FOURTH EDITION

Guidelines for Best Practice in Cross-Cultural Surveys

FULL GUIDELINES

CROSS-CULTURAL SURVEY GUIDELINES

Guidelines for Best Practice in Cross-Cultural Surveys

ISBN 978-0-9828418-2-2 (PDF)

Library of Congress Control Number: 2010937486 2011

Table of Contents

Introduction

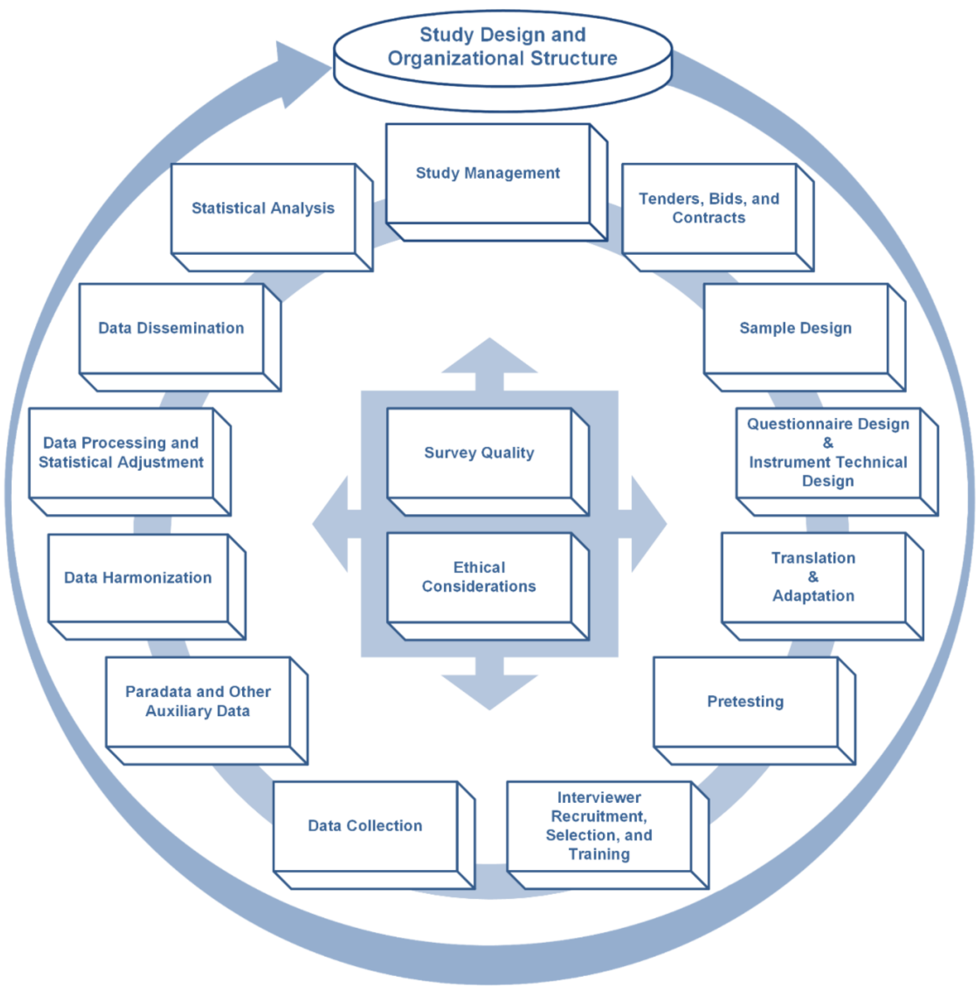
The number and scope of surveys covering many cultures, languages, nations, or regions have increased significantly over the past decade. This has led to a growing need to provide information on best practices across the multiple phases of multinational, multicultural, or multiregional ("3MC") survey design and administration to ensure the collection of high quality comparative data. However, there is very little published information on the details of implementing surveys that is specifically designed for comparative research. For example, little has been published on what aspects of 3MC surveys need to be standardized and when local adaptation is appropriate. The aim of the Comparative Survey Design and Implementation (CSDI) Guidelines Initiative was to develop and promote internationally recognized guidelines that highlight best practice for the conduct of comparative survey research across cultures and countries. The intended audience is researchers and survey practitioners planning or engaged in 3MC research. However, we believe that the Guidelines also could benefit researchers and survey practitioners involved in single country surveys.

The goal of the CSDI Initiative has been to develop Cross-Cultural Survey Guidelines (CCSG) as presented here, which cover all aspects of the survey lifecycle. This currently has resulted 18 chapters and 11 sub-chapters. Three additional chapters on study design and organizational structure, survey quality, and ethical considerations are relevant to all processes throughout the survey production lifecycle. Survey quality can be assessed in terms of fitness for intended use, total survey error, and survey production process quality monitoring. This may be affected by survey infrastructure, costs, interviewer and respondent burden, as well as study design specifications. Figure 1 presents a diagram of the survey lifecycle. The 18 chapters and 11 sub-chapters of the CCSG Guidelines are:

- Study Design and Organizational Structure
- Study Management
- Tenders, Bids, & Contracts
- Sample Design
- Questionnaire Design
- Instrument Technical Design
- Translation
    - Overview
    - Managing and Budgeting
    - Team
    - Scheduling
    - Shared Language Harmonization
    - Assessment
    - Tools
- Adaptation

- Pretesting
- Interviewer Recruitment, Selection, and Training
- Data Collection
    - General Consideration
    - Face-to-Face Surveys
    - Telephone Surveys
    - Self-Administered Surveys
- Paradata and Other Auxiliary Data
- Data Harmonization
- Data Processing & Statistical Adjustment
- Data Dissemination
- Statistical Analysis
- Survey Quality
- Ethical Considerations

Figure 1. The Survey Lifecycle



The CCSG Guidelines draw upon and are based on: (1) general good practice survey methodology, as well as cross-cultural and comparative literature on survey methodology; (2) available study-specific manuals and documentation; and (3) the experiences and lessons learned that authors, reviewers, and editors have added through their work on and with numerous comparative surveys.

Best practices are dynamic and can be expected to evolve over time. At the present time, the Guidelines relate to not just cross-sectional surveys of households and individuals but also computer-assisted personal interviewing modes and the usage of paradata and statistical analyses. At a later point in time, they may be expanded to include establishment and longitudinal surveys.

As more documentation and information about comparative surveys become available, we hope to incorporate the lessons learned from these studies into the CCSG Guidelines. New methodological research will also inform new versions of the CCSG Guidelines. You can greatly help us in these objectives by providing comments and suggestions, or simply alerting us about a topic we need to address. Please contact us at: CCSG-Web-Contact@umich.edu.

Citations:

Please cite these Guidelines as follows: 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. Retrieved Month, dd, yyyy, from http://www.ccsg.isr.umich.edu/.

The CCSG initiative is led by Beth-Ellen Pennell, currently the director of international survey operations at the Survey Research Center, Institute for Social Research at the University of Michigan. Also instrumental in the development and operationalization of the guidelines are Kirsten Alcser and Sue Ellen Hansen of Survey Operations, Survey Research Center, Institute for Social Research. The guidelines were initiated at the 2005 meeting of CSDI and have involved more than 70 individuals from more than 35 organizations worldwide.

We dedicate these guidelines to Dr. Janet A. Harkness. Dr. Harkness passed away in 2012. She initiated the International Workshop on Comparative Survey Design and Implementation where the development of these Guidelines was launched. Dr. Harkness not only contributed to the overall framework and content of the guidelines but she also authored three of the original key chapters: Questionnaire Design, Adaptation and Translation. She inspired this work through her steadfast conviction that resources must be made available to researchers and survey practitioners if we are to improve comparative survey research methods, dissemination and analysis.

Photo images:

Some photo images that appear on this Website are stock photos. Others were provided by the Population and Ecology Laboratory in Nepal (Institute for Social Research, University of Michigan), Professor Jintao Xu at the College of Environmental Sciences and Engineering at Peking University, and Yu-chieh (Jay) Lin and the Institute of Social Science Survey at Peking University.

Contributions to these guidelines have been made by:

Authors

Name	Organization
Kirsten Alcser	Survey Research Center, University of Michigan (UM-SRC)
Christopher Antoun	UM-SRC; now Center for Survey Measurement, U.S. Census Bureau
Dorothée Behr	GESIS – Leibniz Institute for the Social Sciences, Survey Design and Methodology
Grant Benson	UM-SRC
Ipek Bilgen	Survey Research and Methodology Program, University of Nebraska-Lincoln, and UNL Gallup Research Center (UNL-SRAM/GRC); now NORC at the University of Chicago
Emily Blasczyk	UM-SRC
Ashley Bowers	UM-SRC; now Center for Survey Research, Indiana University Bloomington (IN-CSR)
Margaret Brill	UM-SRC; now Bottom Line
Rachel Caspar	Research Triangle Institute (RTI) International
AnaLucía Córdova Cazar	UNL-SRAM/GRC
Judi Clemens	UM-SRC
Julie de Jong	UM-SRC
Brita Dorer	GESIS - Leibniz Institute for the Social Sciences, Survey Design and Methodology
Benjamin Duffey	UM-SRC; now Orbitz Worldwide
Jennifer Durow	UM-SRC
Peter Granda	Inter-University Consortium for Political and Social Research, University of Michigan (UM-ICPSR)
Heidi Guyer	UM-SRC
Sue Ellen Hansen	UM-SRC
Janet Harkness	UNL-SRAM/GRC
Kristen Cibelli Hibben	UM-SRC
Lisa Holland	UM-SRC
Mengyao Hu	UM-SRC
Lei Huang	UNL-SRAM/GRC
Frost Hubbard	UM-SRC; now IMPAQ International
Jennifer Kelley	UM-SRC
Hyun Jung Lee	UM-SRC
Sunghee Lee	UM-SRC
Mingnan Liu	Survey Monkey
Rachel Levenstein	UM-SRC; now American Institutes for Research
Christina Lien	UM-SRC; now Consulting
Yu-chieh (Jay) Lin	UM-SRC
An Liu	UNL-SRAM/GRC; now GfK
Debbie Miller	UNL-SRAM/GRC; now Mixed Methods Consulting
Peter Mohler	University of Mannheim
Rachel A. Orlowski	UM-SRC
Beth-Ellen Pennell	UM-SRC
Emilia Peytcheva	RTI International
Lesli Scott	UM-SRC
Mathew Stange	UNL-SRAM/GRC; now Mathematica Policy Research
Ana Villar	UNL-SRAM/GRC; now City University
Ting Yan	Westat
Dan Zahs	UM-SRC

Editors

Name	Organization	Version
Julie de Jong	UM-SRC	2016
Kristen Cibelli Hibben	UM-SRC	2016
Lars Lyberg	Statistics Sweden; University of Stockholm, now INIZIO	2010
Zeina Mneimneh	UM-SRC	2008, 2010
Peter Mohler	COMPASS and University of Mannheim	2010
Tom W. Smith	NORC, University of Chicago	2016

Reviewers

Name	Organization	Version
Dorothée Behr	GESIS – Leibniz Institute for the Social Sciences, Survey Design and Methodology	2008
Grant Benson	UM-SRC	2008, 2010
Bill Blyth	TNS-Global	2008
Pam Campanelli	Independent consultant	2008
Somnath Chatterji	World Health Organization, Multi-Country Studies Measurement and Health Information	2008
Rory Fitzgerald	City University, Centre for Comparative Social Surveys	2008
Steven Heeringa	UM-SRC	2008
Timothy P. Johnson	Survey Research Laboratory, University of Illinois at Chicago	2016
Achim Koch	GESIS – Leibniz Institute for the Social Sciences, Survey Design and Methodology	2008
Frauke Kreuter	University of Maryland, Joint Program in Survey Methodology	2008
Paul Kussmaul	Retired, formerly Johannes Gutenberg University in Mainz, Germany at Germersheim	2008
Kristin Miller	National Center for Health Statistics	2008
Meinhard Moschner	GESIS – Leibniz Institute for the Social Sciences	2008
José L. Padilla	University of Granada, Social Psychology and Methodology	2008
Alisú Schoua-Glusberg	Research Support Services	2008
Eleanor Singer	UM-SRC	2008
Tom W. Smith	NORC, University of Chicago	2008
Jaré Struwig	Human Sciences Research Council, South Africa	2008
Rick Valliant	University of Maryland, Joint Program in Survey Methodology; UM-SRC	2008
Gordon Willis	National Cancer Institute	2008
Christine Wilson	Heriot-Watt University, School of Management and Languages	2008
Christof Wolf	GESIS – Leibniz Institute for the Social Sciences	2008

Formatting & Copy-Editing

Name	Organization	Version
Gail Arnold	UM-SRC	2008, 2010, 2016
Judi Clemens	UM-SRC	2010, 2011, 2016
Shaw Hubbard	Independent consultant	2008

Graphic Design

Name	Organization	Version
Larry LaFerte	UM-SRC	2008, 2016
Ruth Shamraj	UM-ICPSR	2008

Website Programming

Name	Organization	Version
Jamal Ali	UM-SRC	2016
Tricia Blanchard	UM-SRC	2008, 2010, 2011
Deborah Wilson	UM-SRC	2008, 2010

Study Design and Organizational Structure

Kristen Cibelli Hibben, Julie de Jong, Mengyao Hu, Jennifer Durow, and Heidi Guyer, 2016
(2010 Version: Rachel A. Orlowski and Christopher Antoun)

Introduction

The following guidelines outline a number of study design and organizational considerations that arise when planning multinational, multicultural, or multiregional surveys, which we refer to as “3MC” surveys.

The goal of 3MC surveys is to produce comparable measures across multinational, multicultural, or multiregional populations. To maximize comparability, strict standardization of design is neither always possible nor desired. This is because of the considerable differences in survey context affecting survey design features across cultures and nations. For example, access to up-to-date or good quality sampling frames, the need to accommodate multiple languages (some possibly unwritten), and the available telecommunications, transportation and research infrastructure, are among many other factors that may vary widely (Pennell, Harkness, Levenstein, & Quaglia, 2010; Pennell & Cibelli Hibben, 2016).

3MC study designs that attempt to impose a cookie-cutter or 'one size fits all' approach can actually harm comparability (Harkness, 2008b; Skjåk & Harkness, 2003; Harkness, van de Vijver, & Johnson, 2003; Lynn, Japec, & Lyberg, 2006). For example, an optimal sampling design for one context is rarely optimal, or may be impossible or even detrimental to survey quality if implemented in another context (Heeringa & O’Muircheartaigh, 2010).

Therefore, the challenge in 3MC surveys is to determine the optimal balance between local implementation of a design within each country or culture that will also optimize comparison across countries or cultures (Pennell, Cibelli Hibben, Lyberg, Mohler, & Worku, et al., 2017). The current approach taken by some cross-national surveys is to attempt some level of standardization across country surveys and to monitor and document compliance with the agreed upon standards (for example, see European Social Survey, 2013). Specifications provided to participating countries may require a probability sample but acknowledge that available frames across countries will vary widely. Some frames will require a multi-stage sampling approach where others, such as those in countries with up-to-date registers, may be able to implement a one stage sample design (Heeringa & O’Muircheartaigh, 2010). The European Social Survey, for example, acknowledges these different approaches to sampling in its specifications and in addition to requiring a probability sample design, it also sets a minimum effective sample size, thereby taking into account the design effects (which contribute to sampling error) from the chosen design (European Social

Survey, 2013; Heeringa & O'Muircheartaigh, 2010). See Pennell et al. (2017) for further discussion of challenges in optimizing comparison across countries.

Several factors influence how the overall 3MC study is designed, structured, and implemented, including the source(s) and flow of funding, the research capacity and infrastructure in the participating countries (e.g., availability of sampling frames, field staff, and technical systems). All of these factors will vary from country to country, culture to culture, and from study to study. Yet, before determining other aspects of the study design or the organizational structure, it is critical to clearly define the research questions and the aims and objectives of the study as this should drive subsequent decisions related to other stages in the survey lifecycle. And, it is equally crucial to consider how the ultimate decisions will impact survey quality, assessed in terms of total survey error (TSE), fitness for use, and survey process quality (see Survey Quality for a detailed discussion).

The TSE paradigm is widely accepted as a conceptual framework for evaluating survey data quality (Anderson, Kasper, Frankel, & Associates, 1979; Cochran, 1977) but it can also be used as a blueprint when designing studies (Smith, 2011a). TSE defines quality as the estimation and reduction of the mean square error (MSE) of statistics of interest, which is the sum of random errors (variance) and squared systematic errors (bias). The MSE for each individual statistic in a survey is not typically calculated, due to the following practical problems (see Vehovar, Slavec, and Berzelak (2012) for detailed discussions). First, MSE needs to be calculated differently for different survey parameters (e.g., the survey population mean and variance). It can also differ for each survey item. The fact that a survey usually contains many items and many parameters poses a challenge for the practical application of MSE. Second, the true scores used in bias estimation are often unknown and are usually obtained from a benchmark survey such as Census data or "gold-standard" estimates such as from a face-to-face survey. The accuracy of these estimates, however, is not guaranteed. Third, given that MSE is often a combination of different error sources, it is sometimes difficult to distinguish and separate these error sources. These practical issues become more complicated in 3MC surveys, posing additional challenges to the use of MSE. Despite the challenges, however, the TSE framework helps organize and identify error sources and estimates their relative magnitude, which can assist those planning 3MC surveys to evaluate design and implementation tradeoffs.

TSE takes into consideration both measurement (construct validity, measurement error, and processing error)—i.e., how well survey questions measure the constructs of interest—, as well as representation (coverage error, sampling error, nonresponse error, and adjustment error) (Groves et al., 2009a) —i.e., whether one can generalize to the target population using sample survey data. In the TSE perspective, there may be cost-error tradeoffs, that is, there may be tension between reducing these errors and the cost of reducing them.

Although the TSE paradigm is increasingly used as an organizing framework in the design and evaluation of one-country surveys Pennell et al. (2017) offer a total survey error framework adapted and expanded from Groves et al. (2009a), Tourangeau, Rips, and Rasinski (2000), Smith (2011a), and Smith (2017) for 3MC survey research that integrates error sources with methodological and operational challenges that are unique to or may be more prominent in 3MC surveys (see Figure 1 below).

Figure 9.1 TSE (a) representation in a cross-cultural context

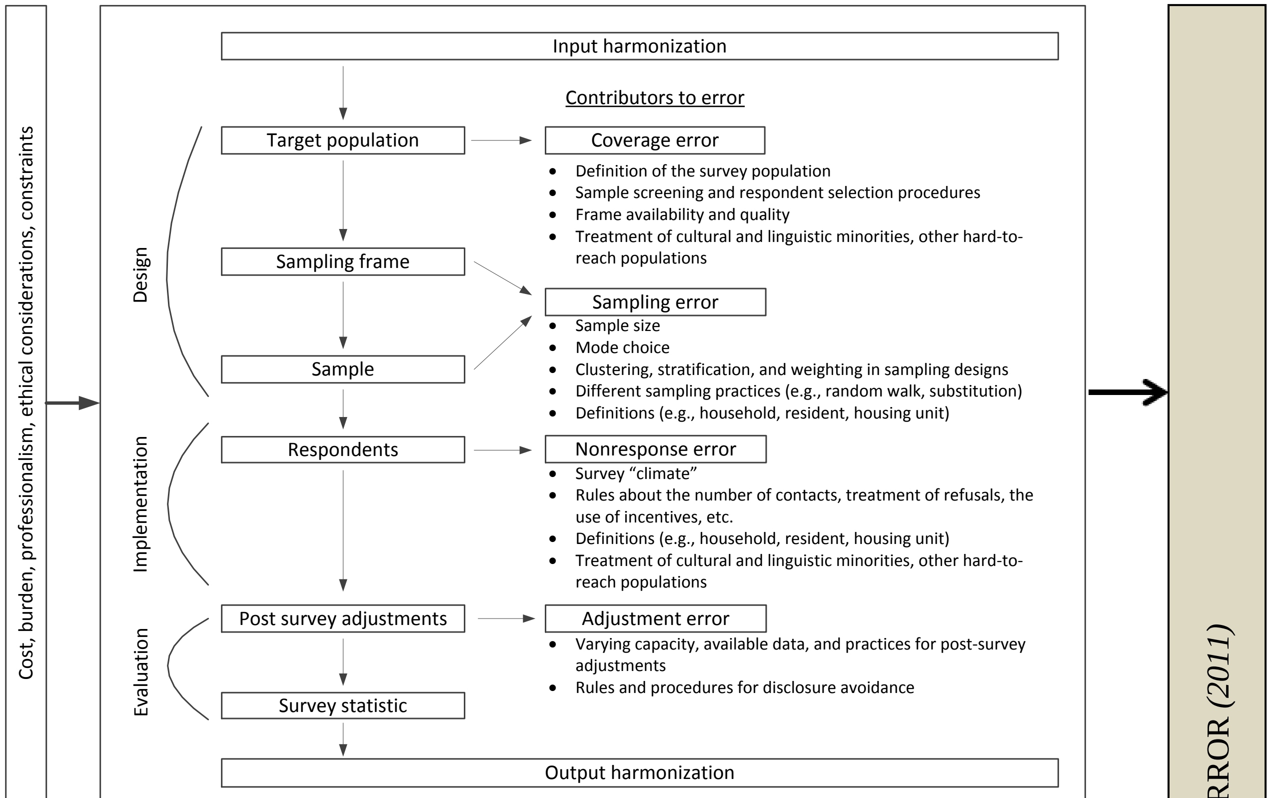
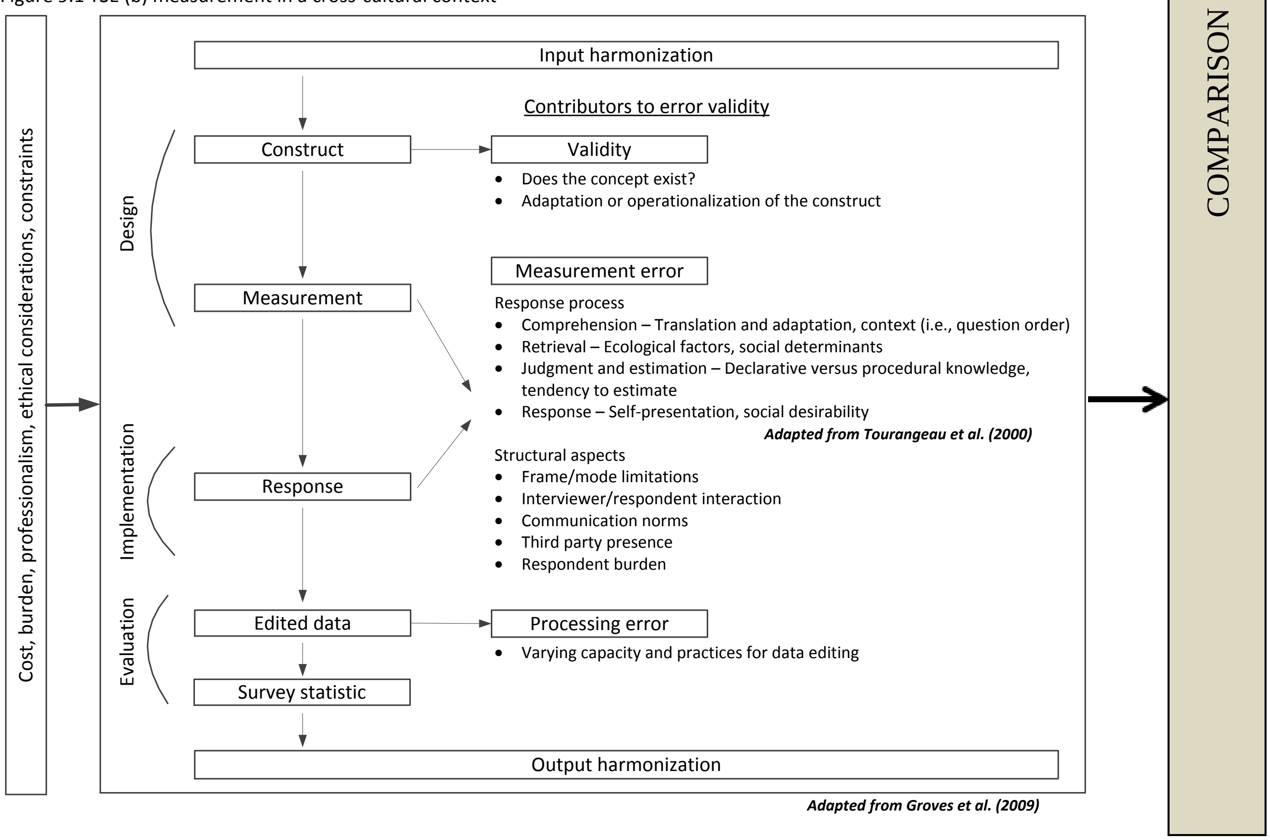


Figure 9.1 TSE (b) measurement in a cross-cultural context



The following describes the main elements of Pennell et al.'s (2017) TSE framework:

- The framework links error sources to the key stages of the survey process: design, implementation, and evaluation.
- Part A of Figure 1 outlines representation error—including coverage error, sampling error, nonresponse error, and adjustment error—which are indicators of how well survey estimates generalize to the target population.
- Part B of Figure 1 encompasses measurement related error—including validity, measurement error, and processing error—which are indicators of how well survey questions measure the constructs of interest.
- As denoted by the resulting "survey statistic" at the end of Part A and Part B, the framework produces statistic-specific error profiles for representation and measurement errors for a single survey statistic. The framework produces statistic specific error profiles because the presence and scale of error may, and frequently does, vary across individual survey statistics.
- The framework incorporates the dimensions of cost, burden, professionalism, ethics, and other design constraints that frequently impose constraints on 3MC survey design and have an important influence on the quality of 3MC surveys.
- The framework includes the role of input harmonization and output harmonization, which are unique to 3MC surveys. Input and output harmonization represent two general approaches to harmonization, which is a term for procedures aimed at achieving, or at least improving, the comparability of different surveys. See Harmonization for further discussion.
- "Comparison error"—a concept introduced by Smith (2011b) —is the conceptual error introduced across each component of a 3MC survey as well as the aggregate of error across all components, which could threaten comparability across surveys.
- For each error component (e.g., coverage error, sampling error, measurement error, etc.), key potential sources of error are identified that may contribute to TSE in individual populations and may present particular challenges to standardizing design and implementation (or establishing suitable localized equivalents) across populations, thereby potentially increasing comparison error. See Pennell et al. (2017) for a detailed discussion of key potential contributions to error and design and implementation challenges across the main stages of the survey lifecycle.

As noted by Smith (2011a), TSE can be used during the design phase for 3MC studies in that each component of error can be considered with the object of minimizing comparison error.

The Cross-cultural Survey Guidelines (CCSG) have been developed to cover all aspects of the lifecycle of 3MC surveys, as shown in the figure on the Chapters page. The lifecycle begins with the guidelines below on establishing aspects of the study design and organizational structure and ends with data dissemination (Data Dissemination). After reviewing the guidelines regarding study design and organizational structure below, we suggest reading Survey Quality followed by Study Management and then the guidelines for each of the elements of the survey lifecycle relevant to your study.

Guidelines

Goal: To consider the key study design decisions that must be addressed in the context of 3MC survey research and how these decisions impact each stage of the survey lifecycle as well as overall survey quality. Additionally, to establish the study's overall structure, the mode of data collection, quality standards from a design perspective, and the elements of the survey lifecycle that are relevant for the study.

1. **Determine key aspects of the overall research design of the study.**

   Rationale

   The first step in designing a 3MC study is to determine key aspects of the overall research design of the study. This includes identification of the research questions and the aims and objectives of the study, assessing the available resources, budget and research capacity of individual study countries and available resources and budget for coordination between study countries, determining the type of study (i.e., cross-sectional or panel), the duration of the study, the populations to be surveyed and the estimated target number of interviews. Subsequent decisions, including those about organizational structure, the mode of data collection, quality standards, and other steps of the survey lifecycle are dependent upon the decisions reached in these key areas.

   Procedural Steps

   1.1 Determine and document the research questions and aims and objectives of the study, ensuring that central and local study goals do not conflict (Biemer & Lyberg, 2003; Federal Committee on Statistical Methodology, 1983). All involved should understand the empirical aims of the research before the organizational and operational work for a study begins. There should be a well-defined direction and purpose of the research, and the aims and objectives should be clearly communicated to all study personnel at the central coordinating center and at study locales. When doing so, consider the following main components:
   1.1.1 Study Aims/Goals: What are the primary research questions or hypotheses the study intends to address?

1.1.2 Representation: What populations are to be studied? See Sample Design and Groves et al. (2009a).

1.1.3 Measurement: What are the measures to be collected? What data are to be collected? See Questionnaire Design and Groves et al. (2009a).

1.1.4 Analysis: What estimates are to be created? (See Data Processing and Statistical Adjustment and Statistical Analysis.

1.2 Investigate how other researchers have addressed similar research questions and consider what data (if any) already exists and what additional data needs to be collected in order to address the research questions.

1.3 Consider whether survey data collection is optimal or whether other methods or mixed methods may be appropriate. Studies involving multiple cultures, countries, regions, or languages may benefit from the use of mixed methods. A mixed methods study "involves the collection or analysis of both quantitative and/or qualitative data in a single study in which the data are collected concurrently or sequentially, are given a priority, and involve an integration of the data at one or more stages in the process of research" (Creswell, Plano Clark, Gutmann, & Hanson, 2003).The different toolkits of qualitative and quantitative data collection methods can be complementary for studies of cross-cultural similarities and differences in attitudes and behaviors that often require different kinds of methods and evidence (van de Vijver & Chasiotis, 2010). van de Vijver and Chasiotis (2010) also provide an in-depth discussion and a conceptual framework for mixed methods studies. Researchers wanting to undertake a mixed methods design or to incorporate mixed methods approaches at different stages of the survey lifecycle may include these considerations when designing the study. Examples and references for mixed methods approaches are provided in Pretesting, Questionnaire Design and Data Collection: General Considerations.

1.4 Assess the available resources and budget for the project, which may affect the scope of the study's aims and objectives that can be realistically undertaken, and will also guide subsequent decisions regarding all steps of the survey lifecycle. In particular, the available resources and budget for the overall coordination of study countries and the resources and research capacity available in individual countries is a key driver of the overall organizational structure for the study. The overall organizational structure of a 3MC survey can be either centralized or decentralized, with a central coordinating center as well as national coordinators in each of the individual study countries. As discussed in further detail in Guideline 2 below, a strong central coordinating center is crucial to effective quality assurance and quality control, but requires significant financial and human resources which may or may not be available depending on the available budget and infrastructure.

1.5 Determine whether to administer a cross-sectional survey or a type of panel survey.

1.5.1 Consider the following attributes of a cross-sectional survey (i.e., a survey where data are collected from selected elements at one point in time) with regard to the aims and objectives of the study.

- Since data are collected at only one point in time in a cross-sectional survey, countries can create an optimal sample design for that specific point in time. If the survey is repeated at a later date, the new cross-sectional study can accommodate changes in the target population which may have occurred, for example, because of migration or other demographic changes
- Since sampling units are only asked to participate once in a cross-sectional survey, the respondent burden over time is less than it would be in a panel survey; this can make it easier to convince the sampling units to participate.
- In a cross-sectional survey, developments or changes on the individual level over time cannot be measured, and it is more difficult to advance a causal argument.

1.5.2 Consider the following attributes of a panel survey (i.e., a survey where the data are collected from selected elements at more than one point in time or data collection waves (Binder, 1998; Kish, 1987; Lynn, 2009) with regard to the aims and objectives of the study. Panel surveys include fixed panel, fixed panel plus births, repeated panel, rotating panel, and split panel studies.

- A panel survey provides the ability to measure changes over time on the statistics of interest at the respondent level.
- In a panel survey, the sampling design, while being optimal at the outset of the panel survey, may be dated and not optimal at a later point in time.
- Changes in the target population are difficult to accommodate (e.g., including new immigrants at a later stage) in a panel survey.
- The initial cost of a panel survey is higher than a cross-sectional survey since both thought and effort need to be expended to plan the best way to capture data over time.
- It can be difficult to convince respondents to participate across multiple waves of data collection, resulting in panel

attrition and reduced sample size in successive waves. With each successive wave of data collection in a panel survey, the cumulative amount of respondent attrition typically increases. Unless the element sample from the original wave of data collection is supplemented with fresh cohorts, the remaining respondents may not accurately reflect the target population.

- For surveys of mobile populations, the attrition rate can be very high. Survey planners should consider how to identify and track panel survey respondents, especially when dealing with a mobile population.
- Question wording and response options need to be comparable across waves in order to allow comparison over time on the statistic of interest.
- Respondents' answers to questions in later waves may be influenced by the interviews conducted in previous waves. This source of error is referred to "panel conditioning" or "time in sample bias" (Sturgis, Allum, & Brunton-Smith, 2009).
- In contrast to a cross-sectional design, a comparative panel survey design implemented across many countries is much more complex. Designers should consider the efforts necessary to achieve comparability simultaneously across each national panel wave and across all countries.

1.6 Determine the timing and duration of the survey.

1.6.1 In some 3MC surveys, particularly those more susceptible to context effects (e.g., a survey of political attitudes), it may be important to complete the data collection in the same timeframe across all study countries.

1.6.2 Other surveys are constrained by a relatively short field period, which may have implications for data collection mode decisions and quality control.

1.6.3 The duration of the study is also dependent on the research goals and type of survey.

1.6.4 When planning the timing of the survey(s), other factors to consider include, seasonal constraints (e.g. rainy seasons), available resources (e.g., longer field period may mean additional cost) and cultural factors (e.g., migration patterns and respondent availability).

1.6.5 The survey duration effects many phases of the survey life cycle, but may have the biggest effect on interviewer recruitment and data collection. See Interviewer Recruitment, Selection, and Training and Data Collection: Face-to-Face Surveys.

1.7 Determine the target population.

- 1.7.1 In a 3MC survey, countries will likely differ in how target populations are defined. From country to country, inclusion criteria may be guided by restricted access to parts of a country's population due to geography, language, instability in the political climate, and other factors. See Heeringa & O'Muircheartaigh (2010) and Pennell & Cibelli Hibben (2016), for examples.
- 1.7.2 The definition of the target population will have implications for the sample design in each country. For example, if the target population is a specific subset (e.g., citizens with a diagnosed health condition), it may be more efficient to develop a sample frame in collaboration with health services rather than launching an area-based probability sample and subsequently screening for this special population.
- 1.7.3 The target population will also impact mode decisions. In the example in Guideline 1.7.2 above, a sample frame developed in collaboration with health services may provide detailed contact information for each person on the sampling frame, which would permit multiple modes of targeting and data collection (e.g., an initial postal mailing informing the respondent of the data collection, a face-to-face contact, and/or ability for a follow-up telephone contact; or a telephone survey rather than a face-to-face survey). In the case of an area probability sample, names and telephone numbers are generally not known ahead of time, limiting mode choices.
- 1.7.4 The target population will also impact most of the other steps in the survey lifecycle, especially in a 3MC study. For example, a country whose target population is multi-lingual or multi-cultural will need to accommodate potential differences in survey items and measurement issues across populations. See especially Questionnaire Design, Translation: Overview, Interviewer Recruitment, Selection and Training, and Data Harmonization.

Lessons learned

1.1. A failure to communicate overall study goals may lead to local decisions that threaten comparability across countries. For example, a country may remove some locally less salient items from the questionnaire in order to reduce the burden of time to both respondents and interviewers without realizing that those items are necessary to measure an important survey construct. Conversely, a country may insert items into the questionnaire in order to study a locally-relevant topic without realizing that those items may affect the quality of the data. When inserting new, or country-specific items, it is necessary to take into account respondent burden, context effects and comparability if the addition of new items is replacing previously existing items.

1.2 The World Fertility Survey (WFS), its successor, the Demographic and Health Survey (DHS), and the International Social Survey Programme (ISSP) are well-known cross-cultural studies which have demonstrated that large-scale probability sample surveys are feasible almost everywhere. For all participating countries in these two studies, sampling frames and resources (including households) were found; local technicians executed complex tasks directed by a centralized international staff; and probability sampling and measurable sampling errors were imposed (Kish, 1994; Scholz & Heller, 2009).

1.3 Survey planners are not always aware of the time and effort required to design and implement quality cross-sectional sampling designs simultaneously across many countries. It might be instructive to consult the extensive documentation of the European Social Survey that includes design, control, and outcomes (European Social Survey, 2010).

1.4 Survey planners are sometimes naïve about the high cost and effort required to maintain a panel survey. When considering the implementation of a panel survey, refer to the literature on longitudinal survey programs such as the Survey of Income and Program Participation (Kasprzyk, 1988), the British Household Panel Survey (Lynn, Häder, Gabler, & Laaksonen, 2007), the European Community Household Panel (Peracchi, 2002), Canada's Survey of Labour and Income Dynamics (Lavallée, Michaud, & Webber, 1993), and additional literature about the methods used in longitudinal surveys (Lynn et al., 2007) and panel surveys (Kasprzyk, Duncan, Kalton, & Singh, 1989). This literature gives a clear sense of the effort and expense necessary to execute a panel survey, and can help survey planners make a more judicious decision regarding the time dimension of the survey design.

1.5 The World Bank's Living Standards Measurement Survey team has developed various household survey design, implementation, and analysis tools such as sample questionnaires and guidelines on questionnaire design, recommendations for maintaining cooperation and avoiding household attrition in longitudinal surveys, example survey manuals and documentation, and guidance for measuring specific topics such as conflict exposure, migration, and fisheries. See
http://econ.worldbank.org/WBSITE/EXTERNAL/EXTDEC/EXTRESE

ARCH/EXTLSMS/0,,contentMDK:23506715~pagePK:64168445~piPK:64168309~theSitePK:3358997,00.html for a list of tools.

2. **Determine the study's organizational structure.**

Rationale

As a rule, the more languages, cultures and countries participating in the survey, the more complex the organizational structure becomes. There are many different ways to organize the structure (see Appendix A for examples), but the key considerations are the locus of control and balancing standardization and localization. The locus of control may be centralized (e.g., all design and operational decisions controlled by a central governing body) or decentralized (e.g., each country makes their own operational decisions while adhering to the study design protocols set by the centralized team). While both centralized and decentralized 3MC surveys are fielded, it is indisputable that a strong centralized infrastructure is needed to maintain quality requirements (Murray, Kirsch, & Jenkins, 1998; Kalton, Lyberg, & Rempp, 1998; Carey, 2000; Pennell et al., 2017; Lyberg, Japec, & Tongur, 2017). For this reason, we outline the advantages of a centralized organization, and only briefly discuss a decentralized organizational structure. As the optimal organizational structure for 3MC surveys, a centralized structure is assumed throughout the guidelines.

Procedural Steps

2.1. Consider maintaining the locus of control as centralized rather than decentralized.
  2.1.1. When the control is centralized, there is a structure with a coordinating center that designs the overall study and assumes the central organizational responsibility to the contracted survey organizations in each country where the study will be carried out. This type of organizing structure is often used in 3MC surveys.
  2.1.2. A coordinating center should include people from different countries, institutions, and affiliations.
  2.1.3. With this organizational structure, the coordinating center will specify the operational structure of the survey for each country to follow. It should determine what elements will be standardized across countries and what elements will be localized; there is a balance between standardization of implementation and adaptation to the cultural context. The coordinating center should inform the survey organizations of the quality standards necessary to execute the study. See Guideline 4 below and Survey Quality.

2.1.4. Although not described here, there are situations where the coordinating center is also responsible for data collection in some or all countries.

2.1.5. When the control is decentralized, each country makes their own operational decisions while adhering to the study design protocols set by the centralized team.

2.1.6 In a decentralized organizational structure, even though all study countries may agree in principle to study design decisions and to protocols for quality assurance and quality control, there is no guarantee that these protocols will be followed. Only strict oversight from a centralized organizational structure can hope to achieve adherence to protocols.

Lessons learned

2.1 Despite knowing the ideal way of executing a study, the available resources often dictate how a study is structured and implemented. For example, multiple sources of funding are typically needed to provide enough support to coordinate a 3MC survey; furthermore, each participating country may be funded separately. Funding sources may have requirements that complicate reporting structures within the study and conflict with the goals of the overall cross-cultural survey. The points at issue may relate to a wide variety of features, from data availability to the content of questionnaires. See Appendix B for examples of how existing 3MC survey programs have been funded.

2.2. As Pennell et al. (2017) note, organizational structure for a 3MC study can be thought of as two extremes. At one end, for a study that is decentralized, a source questionnaire is provided and the details are left up to the participating countries and service providers who deliver the requested data. At the other extreme, "The other extreme can be represented by the ESS that has developed a solid and continuously improving machinery for planning and implementing the survey. One of Sir Roger Jowell's, founder of the ESS, golden rules for comparative surveys was that the number of problems is a function of the number of countries participating in a study (Jowell, 1998; see also Lyberg, et al., 2017). It goes without saying that keeping track of 20 countries is easier than keeping track of 140. In the latter case, the idea that one is in control is very unrealistic without extensive funding for a central infrastructure. We believe that a solid infrastructure is imperative for 3MC surveys to function well and that it is better to limit the number of countries than try to include as many as possible" (Pennell et al., 2017).

3. Determine the mode of data collection to be used and whether it will be standardized across countries and if mixed mode data collection will be permitted within countries.

Rationale

Whether dictated by the coordinating center or left to individual survey organizations to determine, selecting the mode(s) in which the survey will be administered is a major design decision. Surveys can be conducted in numerous ways: face-to-face, by telephone (either conducted by an interviewer or using Interactive Voice Response (IVR)), through the mail, or over the web. The survey instrument format can be paper-and-pencil or computer assisted and either interviewer-administered or self-administered. See Smith and Kim (2015) for a review of surveys modes, their advantages and disadvantages and error structures.

The mode of data collection affects most stages of the survey life-cycle, but arguably the greatest affect is on instrument technical design, data collection, and data processing. Equally affected by mode are issues of comparability, survey cost, and survey error. There is no one "best" mode; rather, the mode(s) of data collection should be selected based on appropriate tradeoffs of time, cost, and error. In a 3MC survey, differences in cultural norms, literacy levels, and logistics may further constrain mode selection.

This guideline focuses on the attributes of different modes vis-a-vis other steps in the survey lifecycle as well the use of a mixed mode design, while also referring the reader to specific chapters for further detail.

Procedural Steps

3.1 Sample design and the mode of data collection are intertwined and the decision about one will affect the decision about the other. When choosing the mode of data collection, consider the following:

3.1.1 The target population for any individual country can influence the decision to collect data via face-to-face, telephone, or self-administered interviews. The following are several examples of the implications of the target population on mode choice.

- If the target population is a nationally representative sample and the geographic region of the country is large (e.g., the United States, Russia, China, etc.), then a face-to-face survey will be significantly more costly than a telephone or self-administered survey.
- If the target population is a population in a climate which is politically unstable, interviewers attempting to complete a survey via telephone may be seen as suspect; and only an

interviewer in a face-to-face setting may be able to obtain cooperation with a respondent.

3.1.2 The availability of the sampling frame and associated infrastructure of the study country can influence the decision to administer a face-to-face, telephone, or self-administered survey.

- For example, many surveys use a sample frame based on an area probability sample and subsequent block listing. Depending on the country's infrastructure, it may or may not be possible to match the household with a telephone number (although this has limitations as well). In such cases where the infrastructure does not permit telephone/address matching, a face-to-face contact or mail survey would be the only way to initially reach the household.

3.2 Consider the length and complexity of the questionnaire when assessing the suitability of different modes.

3.2.1 If the survey is lengthy, a face-to-face interview may be less burdensome to the respondent than a telephone interview (Groves & Kahn, 1979)

3.2.2 If the survey has many skip patterns, then an interviewer administered survey, either by telephone or face-to-face, is preferable to mail survey. A web-based survey may also be suitable if the instrument is programmed so that the respondent does not need to navigate skip patterns,

3.2.3 If the survey is complex and may be difficult for the respondent to understand, then an interviewer administered survey, either by telephone or face-to-face, is advisable so that the interviewer can assist the respondent if necessary.

3.2.4 See Instrument Technical Design and Data Collection: Face-to-Face Surveys for further discussion on questionnaire design vis-à-vis data collection mode.

3.3 Consider the survey topic and potential sensitivity of survey items

3.3.1 If the survey topic is sensitive in an individual study country, a face-to-face interview may serve to put the respondent at ease. Alternately, a survey including sensitive questions may best be, at least partially, self-administered. What is considered as sensitive in one country may not be considered as sensitive in another. See Data Collection: Face-to-Face Surveys and Data Collection: Self-Administered Surveys for a comprehensive discussion of sensitive topics vis-á-vis data collection mode.

3.4 Consider what types of paradata or other auxiliary data might be collected.

3.4.1 Paradata is collected for quality assessment and quality control. An electronic instrument can capture a variety of paradata whereas a paper-and-pencil instrument cannot capture most paradata.

3.4.2 Biomeasures and other auxiliary data can be used for quality assessment and quality control, as well as a complementary data source. Specific auxiliary data may require use of a specific mode of data collection.

- For example, if biomeasures are to be used, face-to-face surveys can facilitate the collection, and indeed may be necessary depending on the type of biomeasures (e.g., blood draw, blood pressure, etc.). However, some biomeasures, such as saliva, can be collected through respondents returning samples through postal mail.

3.4.3 See Paradata and Other Auxiliary Data and Survey Quality for further discussion.

3.5 Consider whether mode will be standardized for a 3MC survey project, or if a mixed mode design will be permitted.

3.5.1 Different modes may produce different survey estimates. These mode-specific differences in measurement might be acceptable to the investigator if nonresponse is sufficiently reduced.

3.5.2 Some studies in the United States employ a mixed mode design in which the least expensive mode is used initially, after which time progressively more expensive modes are implemented in order to reduce nonresponse.

3.5.3 See Data Collection: General Considerations for additional discussion of mixed mode designs and Data Collection: Face-to-Face Surveys for a review of mode effects for sensitive topics.

Lessons Learned

3.1 While a mixed-mode design can reduce the cost of data collection by allowing for increased flexibility to accommodate local contexts, it may also create an additional layer of complexity and, thus, the overall costs for the subsequent harmonization of data by coordinating centers. The Gallup World Poll implements a mixed mode design in which the telephone is used in countries where 80% or more of the target population is covered and face-to-face interviewing is used in countries with lower telephone coverage. The reported costs of telephone surveys are much lower than face-to-face modes (Biemer & Lyberg, 2003), so overall data collection costs are reduced. However, comparability problems due to different modes (phone in one country, face-to-face in another) may exist (Gallup, Inc., 2015). And, this mixed mode approach could lead to non-coverage of up to 20% of a country's population.

3.2 In a cross-national context, the impact of mode can be confounded with cultural differences. For example, when the International Social Survey Programme (ISSP) began, the required mode was a self-administration. However, low literacy levels in some countries necessitated the use of interviewers. Both response rates and reports from substantive measures differed widely, possibly as a result of differences in mode (Skjåk & Harkness, 2003). Therefore, reported variation between countries on survey estimates may indicate substantive differences or may be a result of mode effects and interviewer effects.

3.3 The European Social Survey (ESS) prefers that all data collection be conducted via face-to-face interviews. However, due to local survey infrastructures and costs, some countries want to consider paper-and pencil mode or computer-assisted interviewing or a combination of modes. Extensive research carried out by the ESS to date indicates that the disadvantages would strongly outweigh the advantages of a mixed mode approach in the ESS (Martin & Lynn, 2011). For now, therefore, the ESS has concluded that any move to a mixed-mode data collection would be a threat to comparability.

4. Decide upon quality standards necessary for the implementation of the study from a design perspective.

Rationale

The goal of quality standards is to achieve excellence for all components related to the data (Defeo & Juran, 2010; United Nations, 2005). Setting quality standards is critical to ensuring the same level of methodological rigor across countries (Federal Committee on Statistical Methodology, 1983). Local adaptations will be necessary and appropriate for some aspects of implementation of the study, but any adaptation in the procedure or instrument should be thoroughly discussed, evaluated, and documented beforehand (Mohler, Pennell, & Hubbard, 2008). Frequent measurement and reporting to the coordinating center, along with sufficient methodological support, should allow for timely intervention if problems arise.

Survey quality is a vague concept, which has multiple definitions and has origins in two different developmental paths (Biemer & Lyberg, 2003; Lyberg, 2012). One path is the total survey error paradigm; the other path focuses more on quality management sciences, including fitness for use and survey process quality (Lyberg, 2012). The development of the overall paradigm of survey quality from both the total survey error (TSE) perspective, as well as the quality management sciences perspective, as mentioned by Lyberg (2012), has taken place mainly in official statistics and organizations and has been triggered by the rapid development of technology and other developments. See Survey Quality for a comprehensive discussion of these different survey quality frameworks.

Procedural steps

4.1 Use a Plan-Do-Check-Act cycle (PDCA) by first determining the study's quality standards, then implementing them throughout the research process, while assessing quality indicators at each stage, and finally making appropriate changes to repeat the cycle of PDCA (Biemer & Lyberg, 2003; Deming, 1986).

4.1.1 Consider all potential sources of error in the survey lifecycle, and define quality indicators for key steps in each survey task. See Survey Quality for common sources of error and possible indicators, as well as a thorough discussion of how the TSE, fitness for use, and survey process quality frameworks can guide assessment of error through the steps of the survey lifecycle.

4.2 Acquaint study organizers with important quality control literature that distinguishes between common and special causes of variation, as well as explains how to act on information about these different kinds of variation (Lyberg & Biemer, 2008; Montgomery, 2005; Ryan, 2000).

4.3 Form a team in each country that regularly meets to discuss the quality of the local survey. The team should have or should be provided with methodological expertise needed. The team should document and report any concerns to the coordinating center (Aitken, Hörngren, Jones, Lewis, & Zilhäo, 2003; Biemer & Lyberg, 2003).

4.4 Identify tools that control and maintain operational process quality.

4.5 Implement a certification process or a signing-off procedure for each stage in order to check and document that the study design and specification standards are being followed.

4.5.1 Quickly address and remedy, if possible, any deviations from expectations that may occur (Biemer & Lyberg, 2003).

4.5.2 Invoke sanctions, as specified in the contract, if the certification is not fulfilled.

4.6 Consider site visits to all countries to monitor or support the implementation of quality standards. Make sure these visits are specified in the contract with each survey organization.

4.7 If and where possible, incorporate methodological research. This will inform long-term quality improvement (Jowell, 1998; United Nations, 2005). See also Paradata and Other Auxiliary Data for further discussion on the use of these data for methodological analyses.

Lessons learned

4.1 Variations in country-level research infrastructure, research traditions, and methodological rigor need to be thoroughly investigated and understood when setting quality standards. Some countries will need more assistance in meeting certain standards, and this should be taken into account early in the planning process.

Appendix A

Funding sources

The source and flow of funding impact the structure of a cross-cultural survey. Additionally, the flow of funding or funding structure may change over the course of a study, especially among longstanding studies or programs. Below are examples of how some large-scale, cross-cultural survey programs have been funded. Please see the websites of these programs for further information.

- Some large, cross-cultural studies are European Research Infrastructure Consortiums (ERICs). A ERIC is a specific legal form in Europe between different research groups, established to build and maintain a joint research infrastructure, and is funded by the countries joining the ERIC (http://ec.europa.eu/research/infrastructures/index_en.cfm?pg=what) .
  - The Survey of Health, Ageing and Retirement in Europe (SHARE) (2014) became the first ERIC in March, 2011, giving it legal personality in all EU Member States and other partner countries of the ERIC, as well as some tax exemptions. SHARE-ERIC was initially hosted by the Netherlands; recently its seat was transferred to Munich, Germany. The project investigates health, socio-economic status and social and family networks among adults age 50 and older in over 20 European countries and Israel. Five waves of data collection have taken place beginning in 2004. Austria, Belgium, the Czech Republic, Germany, and the Netherlands are the founding members of SHARE-ERIC, with Switzerland having observer status. Since then, Italy, Greece, Israel, Slovenia, Sweden, and Poland have also become members.
  - Following an application to the European Commission, submitted by the UK on behalf of 14 other countries, the European Social Survey (ESS) (2014) was awarded ERIC status in November, 2013. The ESS is an academically driven cross-national survey that has been conducted every two years across Europe since 2001. The ESS investigates the interaction between Europe's changing institutions and the attitudes, beliefs, and behavior patterns of its diverse populations using face-to-face interviews in over 30 countries throughout four rounds. Before the ESS was awarded ERIC status, it had been funded on a round-by-round basis through the European Commission's Fifth, Sixth and Seventh Framework Programmes, the European Science Foundation (ESF) and national funding councils in the participating countries.

- The International Social Survey Programme (ISSP) (2015) investigates current social science topics in each of 48 participating countries by collecting self-administered questionnaires. Each survey organization has funded all of its own costs; there are no central funds.

- Latinobarómetro (2014) investigates social development, with face-to-face interviews in 18 Latin American countries occurring annually since 1995. Initial funding came from the European Commission. There have been several additional funding sources, including: international organizations (e.g., Inter-American Development Bank, United Nations Development Programme, World Bank), government agencies, and private sector sources.

- The Asian Barometer Survey (ABS) (2014) aims to gauge public opinion on issues such as political values, democracy, and governance across Asia. The survey network includes research teams from 13 East Asian states and 5 South Asian countries. The ABS (formerly the East Asia Barometer) has received financial support from a variety of agencies and organizations. Since 2003, the ABS has received regular funding from the Institute of Political Science at Academia Sinica. The Program for East Asia Democratic Studies has been co-hosting the project since 2005 under the auspice of the Institute for Advanced Studies in Humanities and Social Sciences at National Taiwan University (NTU). The ABS has also received substantial financial support from the Henry Luce Foundation and the World Bank. In addition, many country teams have secured funding from national and international sources to sponsor their own fieldwork.

- The Arab Democracy Barometer (2014) was established in 2005 to produce scientifically reliable data on the politically-relevant attitudes of ordinary citizens, to disseminate and apply survey findings in order to contribute to political reform, and to strengthen institutional capacity for public opinion research. In 2010/11, surveys were conducted in 11 Arab countries with funding provided by the United Nations Development Programme, the International Development Research Council of Canada, and the United States Institute of Peace. The third wave of the Arab Barometer is currently underway and is funded by the Canadian International Research and Development Centre (IDRC).

- Afrobarometer (2014) is an independent, non-partisan research project that measures the social, political, and economic atmosphere in Africa. Afrobarometer surveys are conducted in 35 African countries and are repeated on a regular cycle. Core donors for Afrobarometer Rounds 5 and 6 include the Mo Ibrahim Foundation, the Swedish International Development Cooperation Agency (SIDA), Department for International Development (DFID), UK, and the United States Agency for International Development (USAID) with supplemental funding provided by the World Bank, Institute for Security Studies (South Africa), United States Institute of Peace, Transparency International, and the Bill and Melinda Gates Foundation.

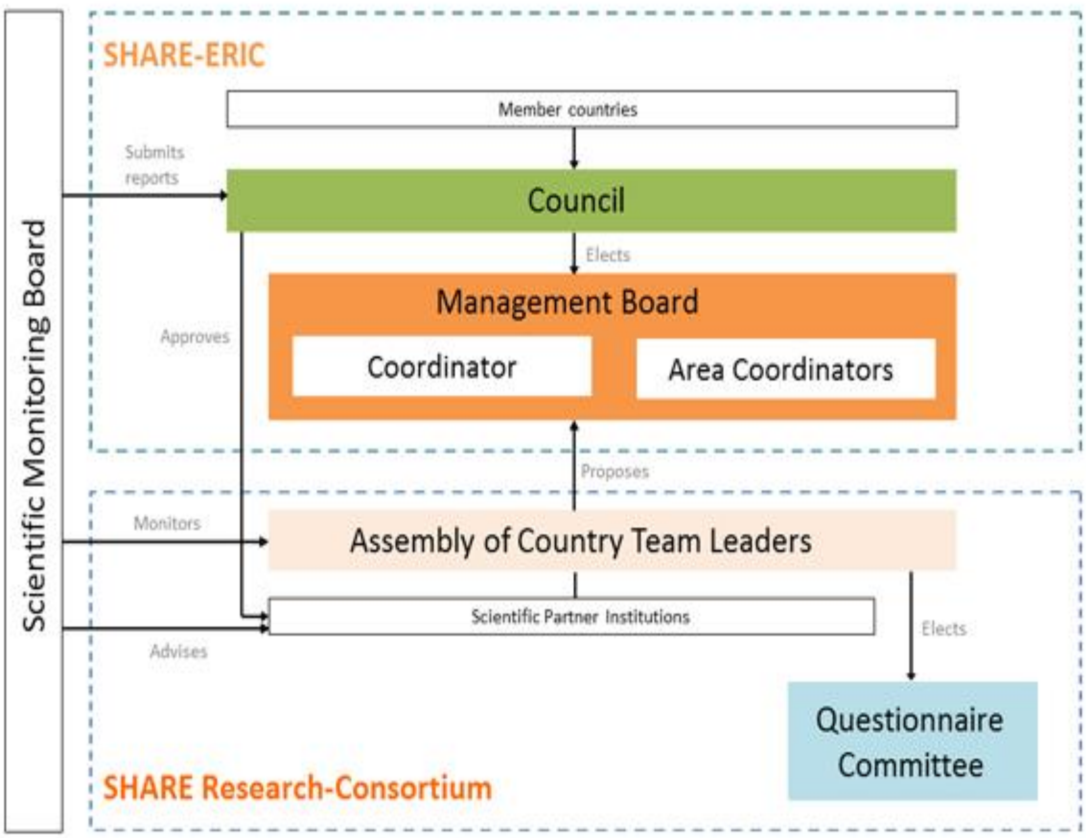
The World Mental Health Surveys (2014) investigate mental disorders with face-to-face interviews in 28 countries since 2000. The World Mental Health Survey Initiative is supported by the National Institute of Mental Health, the John D. and Catherine T. MacArthur Foundation, the Pfizer Foundation, the US Public Health Service, the Fogarty International Center, the Pan American Health Organization, Eli Lilly and Company, Ortho-McNeil Pharmaceutical, GlaxoSmithKline, and Bristol- Myers Squibb. In addition, each participating country has had its own sources of funding.

Appendix B

Organizational structures

Below are descriptions of the organizational structures that have been used on some large-scale, cross-cultural survey programs. These examples are only illustrative. Please visit the survey programs' websites for more information about their structure.

- Survey of Health, Ageing, and Retirement in Europe (SHARE) (Börsch-Supan, Jürges, & Lipps, 2003; Survey of Health, Ageing, and Retirement in Europe, 2014).
  - The governance of the scientific work to build up SHARE involves three separate bodies: a legal entity called SHARE ERIC, a research consortium formed by the scientists who carry out the scientific work in SHARE, and a Scientific Monitoring Board which is independent from the two other bodies and advises both SHARE ERIC and the Research Consortium.



- All members of the SHARE ERIC are represented on the Council, which has full decision-making powers, including the adoption of the budget. The Council appoints the Coordinator, the Vice-Coordinator, and the Coordinator Management as the legal representatives of the SHARE ERIC, and the other members of the Management Board, the executive body of the SHARE ERIC.

- The Management Board proposes all strategic and budgetary decisions to the Council. It is responsible for all financial and governance processes which maintain scientific integrity, cross-national comparability, and an overall balance of the SHARE survey design. Specifically, it is accountable for the SHARE ERIC's finances and deliverables, and for observing legal requirements such as data confidentiality and safety regulations at the European level.
- The Scientific Monitoring Board monitors the scientific quality of SHARE. It gives feedback to the Management Board and the research consortium at least once per year. Every two years, the Scientific Monitoring Board issues a written report to the Council of the SHARE ERIC. This report also assesses the services offered to the users of the SHARE data.

▪ SHARE is organized in various teams, including country teams, area teams, teams providing weights and imputations, programmers, and the central coordination team. SHARE is coordinated in Germany at the Munich Center for the Economics of Aging (MEA), Max Planck Institute for Social Law and Social Policy.
  - Country teams play a crucial role, particularly when knowledge of the language or other country specific issues is needed.
  - Area coordinators are responsible for the central research fields of SHARE: economics, health, health care and social networks.
  - Weights and imputations are managed by expert teams in Italy.
  - The programming of the instrument and data distribution is conducted by CentERdata, located at the University of Tilburg, Netherlands.

- European Social Survey (2015)
  ▪ Each member of the ESS ERIC has a national representative in the General Assembly. The General Assembly appoints the Director, has full decision making powers regarding the operations and management of the ESS ERIC, and has three standing committees: a Scientific Advisory Board (SAB), which provides advice and guidance on the substantive coverage of the ESS ERIC; a Methods Advisory Board (MAB), which provides advice and guidance on methodology; and a Finance Committee (FINCOM), which provides guidance on the financial health of the ESS ERIC.
  ▪ The Central Coordinating Team is responsible for overseeing the entire study and is in contact with the Funders, the Scientific Advisory Board, the Specialist Advisory Groups, and the National Coordinators/Survey Institutes.
    - The Scientific Advisory Board consists of representatives from each participating country, two representatives from the European

Commission, and two representatives from the European Science Foundation.
- The Specialist Advisory Groups have separate teams with expertise in question design, methods, sampling, and translation.
- The National Coordinators/Survey Institutes have one director from each country and one national survey organization from each country. The survey organizations are chosen by their country's respective national academic funding body.

- International Social Survey Programme (ISSP) (2015)
  - The Programme consists of countries which are ISSP members, the ISSP secretariat, the ISSP archive, the ISSP sub-groups drawn up within the ISSP, drafting groups for modules, and methodology research groups.
  - General meetings are held once a year. Each participating nation is entitled to be represented at the General Meetings by not more than three people. If there is no consensus upon a matter, a vote may be taken in which each country has one vote. Decisions are by simple majority of the countries present and eligible to vote at a specific General Meeting. A major function of these meetings is to work on modules. Those members who are not to conduct a particular round of the survey have no vote on the questionnaire for that year. Programme meetings and surveys are conducted according to the ISSP Working Principles, which set out business procedures for meetings, for conducting surveys, and for archiving data.
  - A Standing Committee on organizational matters is elected to assist the Group in making decisions on membership, venues for future meetings, funding of joint activities, etc. The Standing Committee consists of the Secretariat and four other members elected for four-year terms.
  - A Methodology Committee is elected to assist the Group in assessing and enforcing the technical standards of the ISSP. The Methodology Committee has seven members, elected by the General Meeting. Each member is elected for a four-year term. The Methodology Committee may create sub-committees to carry out the various tasks assigned to it. The Methodology Committee may appoint other ISSP members to assist in its tasks and serve on the sub-committees and should consult with experts outside the ISSP as needed.
  - The General Meeting selects a Drafting Group of three to six member nations to prepare a draft questionnaire on behalf of the Group.

- Globalbarometer Surveys (GBS)
  - The Globalbarometer Surveys are a network of regional barometers that have been adapted to world regions undergoing rapid political and economic change. Currently, the Globalbarometer Surveys include Africa (Afrobarometer), East and South Asia (Asian Barometer

Survey), Central and South America (Latinobarómetro), and the Middle East (Arab Democracy Barometer).
  - The organizational structure of the Globalbarometer network is based on the idea of self-governance -- i.e. each regional barometer directs its own roster of research institutes located in the 50 countries where surveys are conducted. For example, the Afrobarometer commissions data collection in Nigeria from Practical Sampling International, and in South Africa from Citizens Surveys. However, to properly coordinate the operation of each regional barometer and maintain high standards of research, the GB network is managed with three bodies:
    - An Executive Board, composed of one representative from each regional barometer. The Executive Board provides leadership and makes decisions for the Globalbarometer, develops proposals for research and funding, plans and coordinates surveys according to a common schedule, and authorizes other actions, including delegating tasks to working groups.
    - A General Meeting, representing the national partners in the network. The purpose of the General Meeting is to discuss GB protocols, to raise new subjects, and to provide inputs into Executive Board decisions. Through region-wide and cross-regional meetings in different cities, the GB network also hammers out questionnaires, develops new methods, and reports results through an iterative process of professional exchange.
    - An Advisory Board, consisting of respected senior analysts and practitioners. The Board provides general advice, technical expertise, and professional contacts on as-needed basis.

- World Mental Health Surveys (Pennell et al., 2009; World Mental Health Study, 2014)
  - The World Health Organization is invested in the objectives of this survey and works closely with two study-level Principal Investigators. These study-level researchers make many of the ultimate decisions for the entire study. The World Health Organization is in contact with the Data Collection Coordination Center and the Analysis Coordination Center.
  - The Data Collection Coordination Center is instrumental in writing and implementing the specifications for pre-production and production activities. The University of Michigan is the Data Collection Coordination Center and its tasks include such activities as selecting survey organizations, training interviewers, and providing assistance during data collection.
  - The Analysis Coordination Center makes decisions regarding post-production activities. Harvard University is the Analysis Coordination Center.
  - The Working Groups are analysis teams that focus on one particular aspect or analytic perspective of mental health. Each Working Group is

led by a Chair. Examples of focal topics include the following: ADHD, drug dependence, gender, social class, suicide, and personality disorders. The Working Groups are in contact with the Analysis Coordination Center and the Principal Investigators from each country.

- The Principal Investigators from each country oversee their respective country's survey.
- The Data Collection Organizations are the survey organizations within each country that carry out the field operations.

References

Afrobarometer. (2014). Retrieved Nov 28, 2014, from http://www.afrobarometer.org

Aitken, A., Hörngren, J., Jones, N., Lewis, D., & Zilhäo, M. J. (2003). *Handbook on improving quality by analysis of process variables.* Luxembourg: Eurostat. Retrieved March 27, 2010, from http://ec.europa.eu/eurostat/documents/64157/4373903/06-Handbook-on-improving-quality-by-analysis-of-process-variables.pdf/b0006e09-1708-4f8e-97e0-6a54d840b92b

Anderson, R., Kasper, J., Frankel, M., & Associates (Eds.). (1979). *Total survey error: Applications to improve health surveys*. San Francisco: Jossey-Bass.

Arab Democracy Barometer. (2014). Retrieved Nov 28, 2014, from http://www.arabbarometer.org

Asian Barometer Survey (ABS). (2014). Retrieved Nov 28, 2014, from http://www.asianbarometer.org

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

Binder, D. (1998). Longitudinal surveys: Why are these surveys different from all other surveys? *Survey Methodology*, *24*(2), 101-108.

Börsch-Supan, A., Jürges, H., & Lipps, O. (2003). *SHARE: Building a panel survey on health, aging and retirement in Europe*. Mannheim, Germany: Mannheim Research Institute for the Economics of Aging. Retrieved March 26, 2010, from http://www.mea.mpisoc.mpg.de/uploads/user_mea_discussionpapers/k3pzhnwk1t1zjymk_dp32.pdf

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

Cochran, W. G. (1977). *Sampling techniques*. New York, NY: John Wiley & Sons.

Creswell, J. W., Plano Clark, V. L., Gutmann, M. L., & Hanson, W. E. (2003). Advanced mixed methods research designs. In A. Tashakkori and C. Teddlie (Eds.), *Handbook on mixed methods in the behavioral and social sciences* (pp. 209-240). Thousand Oaks, CA: Sage Publications.

Defeo, J. A., & Juran, J. M. (2010). *Juran's quality handbook: The complete guide to performance excellence* (6th ed.). New York, NY: McGraw-Hill.

Deming, E. (1986). *Out of the crisis*. Cambridge, MA: Cambridge University Press.

European Social Survey (ESS). (2010). Retrieved May 03, 2010, from http://www.europeansocialsurvey.org

European Social Survey (ESS). (2013). *Round 7 specification for ESS ERIC member and observer countries*. City University, London: Centre for Comparative Social Surveys.

European Social Survey (ESS). (2015). ESS5 – 2010 Documentation Report. Retrieved May 5, 2016, from http://www.europeansocialsurvey.org/docs/round5/survey/ESS5_data_documentation_report_e04_0.pdf

Federal Committee on Statistical Methodology. (1983). *Statistical policy working paper 9: Contracting for surveys*. Washington, DC: Office of Management and Budget.

Gallup, Inc. (2015). How does the Gallup World Poll work? Retrieved Jan 30, 2015 from http://www.gallup.com/178667/gallup-world-poll-work.aspx

Groves, R. M., & Kahn, R. (1979). *Surveys by telephone; A national comparison with personal interviews*, New York Academic Press.

Groves, R. M., Fowler, F. J. Jr., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2009a). *Survey methodology* (2nd ed.). Hoboken, NJ: John Wiley & Sons.

Harkness, J. A. (2008b). 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). New York, NY: Psychology Press Taylor & Francis Group.

Harkness, J. A., van de Vijver, F. J. R., & Johnson, T. P. (2003). Questionnaire design in comparative research. In J.A. Harkness, F.J. van de Vijver, P.P. Mohler, (Eds), *Cross-cultural survey methods* (pp. 19–34). Hoboken, New Jersey: John Wiley & Sons.

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. P. Mohler, … T. W. Smith (Eds.), *Survey methods in multicultural, multinational, and multiregional contexts* (pp. 251–267). Hoboken, New Jersey: John Wiley & Sons.

International Social Survey Programme (ISSP) (2015). Retrieved March 9, 2015, from http://www.issp.org/

Jowell, R. (1998). How comparative is comparative research? *American Behavioral Scientist, 42*(2), 168-177.

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*. Washington D.C.: U.S. National Center for Education Statistics.

Kasprzyk, D. (1988). *The Survey of Income and program participation: An overview and discussion of research issues* (No. 77). Washington, DC: U.S. Bureau of the Census. Retrieved March 26, 2010, from //www2.census.gov/prod2/sipp/wp/SIPP_WP_077.pdf

Kasprzyk, D., Duncan, G.J., Kalton, G., Singh, M. P. (Eds.). (1989). *Panel surveys*. New York, NY: John Wiley & Sons.

Kish, L. (1987). *Statistical design for research*. New York, NY: John Wiley & Sons.

Kish, L. (1994). Multipopulation survey designs: Five types with seven shared aspects. *International Statistical Review, 62*, 167–186.

Latinobarómetro. (2014). Retrieved Nov 28, 2014, from http://www.latinobarometro.org

Lavallée, P., Michaud, S., & Webber, M. (1993). The Survey of Labour and Income Dynamics, design issues for a new longitudinal survey in Canada. *Bulletin of the International Statistical Institute* (49th Session, Book 2) (pp. 99-100).

Lyberg, L. (2012). Survey quality. *Survey Methodology*, *38*(2), 107–130. Retrieved from www.statcan.gc.ca/pub/12-001-x/12-001-x2012002-eng.htm

Lyberg, L. E., & Biemer, P. P. (2008). Quality assurance and quality control in surveys. In E. D. de Leeuw, J. J. Hox, & D. A. Dillman (Eds.), *International handbook of survey methodology*. New York/London: Lawrence Erlbaum Associates.

Lyberg, L., Japec, L., & Tongur, C. (2017). Prevailing issues and the future of comparative surveys. In T. P. Johnson, B-E. Pennell, I. A. L. Stoop, & B. Dorer (Eds.), *Advances in comparative survey methods: Multicultural, multinational and multiregional contexts (3MC).* New York, NY: John Wiley & Sons.

Lynn, P. (Ed.) (2009). *Methodology of longitudinal surveys*. New York, NY: John Wiley & Sons.

Lynn, P., Häder, S., Gabler, S., & Laaksonen, S. (2007). Methods for achieving equivalence of samples in cross-national surveys: The European Social Survey experience. *Journal of Official Statistics, 23*(1), 107–124.

Lynn, P., Japec, L., & Lyberg, L. (2006). What's so special about cross-national surveys? in J. A. Harkness (Ed.), *Conducting cross-national and cross-cultural surveys*, pp. 7-20. ZUMA-Nachrichten Spezial, 12, Mannheim: ZUMA.

Martin, P., & Lynn, P. (2011). *The effects of mixed mode survey designs on simple and complex analyses*. Centre for Comparative Social Surveys, London, U.K., Working Paper Series 4.

Mohler, P. Ph., Pennell, B-E., & Hubbard, F. (2008). Survey documentation: Toward professional knowledge management in sample surveys. In E. D. de Leeuw, J. J. Hox, & D. A. Dillman (Eds.), *International handbook of survey methodology.* New York/London: Lawrence Erlbaum Associates.

Montgomery, D. C. (2005). *Introduction to statistical quality control* (5th ed.). Hoboken, NJ: John Wiley & Sons.

Murray, T. S., Kirsch, I. S., & Jenkins, L. B. (1998). *Adult literacy in OECD Countries: Technical report on the first International Adult Literacy Survey*. Retrieved Jan 03, 2016, from http://eric.ed.gov/?id=ED445117

Pennell, B-E., & Cibelli Hibben, K. L. (2016). Surveying in multicultural and multinational contexts. In C. Wolf, D. Joye, T. Smith, Y-Ch Fu, (Eds.): *The SAGE handbook of survey methodology.* London, UK: 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, … B. West (Eds.), *Total survey error in practice.* New York, NY: John Wiley & Sons.

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. P. Mohler, …T. W. Smith (Eds.), *Survey methods in multicultural, multinational, and multiregional contexts*, (pp. 269–298). Hoboken, New Jersey: John Wiley & Sons.

Pennell, B-E., Mneimneh, Z., Bowers, A., Chardoul, S., Wells, J. E., Viana, M. C., … Vilagut, G. (2009). Implementation of the World Mental Health survey initiative. In R. C. Kessler & T. B. Üstün (Eds.), *Volume 1: Patterns of mental illness in the WMH Surveys.* Cambridge, MA: Cambridge University Press.

Peracchi, F. (2002). The European community household panel: A review. *Empirical Economics*, *27*, 63-90.

Ryan, T. P. (2000). *Statistical methods for quality improvement* (2nd ed.). Hoboken, NJ: John Wiley & Sons.

Scholz, E., & Heller, M. (2009). *ISSP Study monitoring 2007*. GESIS Technical Reports, 2009/5. Mannheim: GESIS. Retrieved March 26, 2010, from http://www.gesis.org/fileadmin/upload/forschung/publikationen/gesis_reihen/gesis_methodenberichte/2009/TechnicalReport_09-5.pdf

Skjåk, K. K., & Harkness, J. A. (2003). Data collection methods. In J. A. Harkness, F. J. R. van de Vijver, & P. P. Mohler (Eds.), *Cross-cultural survey methods* (pp. 179–194). Hoboken, New Jersey: John Wiley & Sons.

Smith, T. W. (2011a). Refining the total survey error perspective. *International Journal of Public Opinion Research*, (23), 464-484.

Smith, T. W. (2011b). The report of the international workshop on using multi-level data from sample frames, auxiliary databases , paradata and related sources to detect and adjust for nonresponse bias in surveys. *International Journal of Public Opinion Research*, *23*(3), 389-402.

Smith, T. W. (2017). Improving cross-national/cultural comparability using the total survey error paradigm. In T. P. Johnson, B-E. Pennell, I. A. L. Stoop, & B. Dorer (Eds.), *Advances in comparative survey methods: Multicultural, multinational and multiregional contexts (3MC).* New York, NY: John Wiley & Sons.

Smith, T. W. & Kim, J. (2015). A review of survey data collection modes: With a focus on computerizations. *Sociological Theory and Methods*, *20*(2), 185-200.

Sturgis, P., Allum, N., & Brunton-Smith, I. (2009). Attitudes over time: The psychology of panel conditioning. In P. Lynn (Ed.), *Methodology of longitudinal surveys*, pp. 113-126. New York, NY: John Wiley & Sons.

Survey of Health, Ageing, and Retirement in Europe. (2014). Retrieved Nov 28, 2014, from http://www.share-project.org

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge University Press.

United Nations. (2005). *Household surveys in developing and transition countries.* New York, NY: United Nations, Department of Economic and Social Affairs. Retrieved March 26, 2010, from http://unstats.un.org/unsd/HHsurveys/pdf/Household_surveys.pdf van de Vijver, F. J. R. & Chasiotis, A. (2010). Making methods meet: Mixed design in cross-national research. In J. Harkness, B. Edwards, M. Braun, T. Johnson, L. E. Lyberg, P. Mohler, … & T. Smith. (Eds.), *Survey methods in multicultural, multinational, and multiregional contexts*. Hoboken, NJ: John Wiley & Sons.

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).

World Mental Health Study. (WMHS). (2014). Retrieved Nov 28, 2014, from http://www.hcp.med.harvard.edu/wmh/

Further Reading

Harkness, J., Braun, M., Edwards, B., Johnson, T., Lyberg, L., Mohler, … Smith, T. (2010) *Survey methods in multinational, multiregional and multicultural contexts.* Wiley & Sons.

Heath, A., Fisher, S., & Smith, S. (2005). The globalization of public opinion research. *Annual Review of Political Science, 8*, 297-333.

Lynn, P. (2003). Developing quality standards for cross-national survey research: Five approaches. *International Journal of Social Research Methodology, 6*(4), 323-336. Retrieved October 8, 2009, from http://www.soc.uoc.gr/socmedia/papageo/developing%20quality%20standards%20for%20cross-national%20surveys.pdf

Study Management

Lesli Scott, Julie de Jong, and Kristen Cibelli Hibben, 2016

Introduction

Conducting a multinational, multiregional, or multicultural survey, which we refer to as a “3MC” survey, involves careful coordination of many elements defined in the survey production lifecycle. Reflecting on the historical development of comparative studies, it is “…a quantum leap in complexity when one moves from the national to the multi-national arena in survey design and implementation” (Mohler, 2007, page 159). Research teams face many challenges in their attempts to manage all the requirements, elements, stakeholders and constraints of 3MC studies. The following guidelines provide a suggested framework for study management activities, incorporating aspects of the Project Management Body of Knowledge, often called PMI PMBOK ® (Project Management Institute, 2013). After reviewing the study management guidelines below, we suggest reading the guidelines for each of the elements of the survey lifecycle relevant to your study (see the Chapters page for an overview and figure of the survey lifecycle).

This introduction begins by discussing why study management is important in the context of 3MC surveys. The next section provides generic project management concepts including definitions for key roles (project manager, team members and stakeholders), followed by a description of core project management topics including scope, schedule, cost and quality. Guidelines and procedural steps are then presented for the main phases of projects including the initiation, planning, execution, and closing phases. Throughout the chapter, there are tools and examples that may be helpful for 3MC study management. The appendices provide templates for some of these tools. Appendix I, presents a table with links to useful project management examples and resources from the European Social Survey (ESS) and the Teaching & Learning International Survey (TALIS).

Study management is critical to successful completion of survey projects. It embodies techniques that can be used to set and attain project goals and to manage activities effectively. Additionally, study management is essential for achieving the comparability and quality standards demanded by 3MC studies. “In comparative survey research, much more than the problems common to all mono-cultural surveys and measures need to be taken into consideration. In addition to depending on the **quality** of each individual national or cultural survey and measurement component, cross-cultural research is also dependent on their **comparability**.” (Johnson & Braun (2016), page 41). In the past, it was often assumed that countries were able to follow instructions or specifications “without much guidance or explanation” but many collaborators in study countries have found it challenging to institute the required protocols because of a lack of experience and infrastructure or may have taken short-cuts in quality assurance and quality control procedures (Lyberg, Japec & Tongur, 2017). Over the past few decades, comparative researchers have attended more rigorously to the planning, execution and evaluation processes that comprise study management. As a result, greater "methodological equivalence" such as that which defines current rounds of the European Social Survey (ESS), are being achieved (Jowell, Kaase, Fitzgerald & Eva, 2007).

The central coordination and local data collection efforts of many 3MC surveys are implemented within the structure of pre-existing programs that carry out ongoing operational activities. The organizations' ongoing program activities may extend through long timeframes and can include many types of endeavors in addition to a specific survey project. The ESS European Research Infrastructure (ESS ERIC), for example, has implemented many surveys as well as conferences since 2001 within a program structure including a headquarters, a general assembly, a core scientific team, a national coordinators' forum and others bodies (European Social Survey, 2016)).

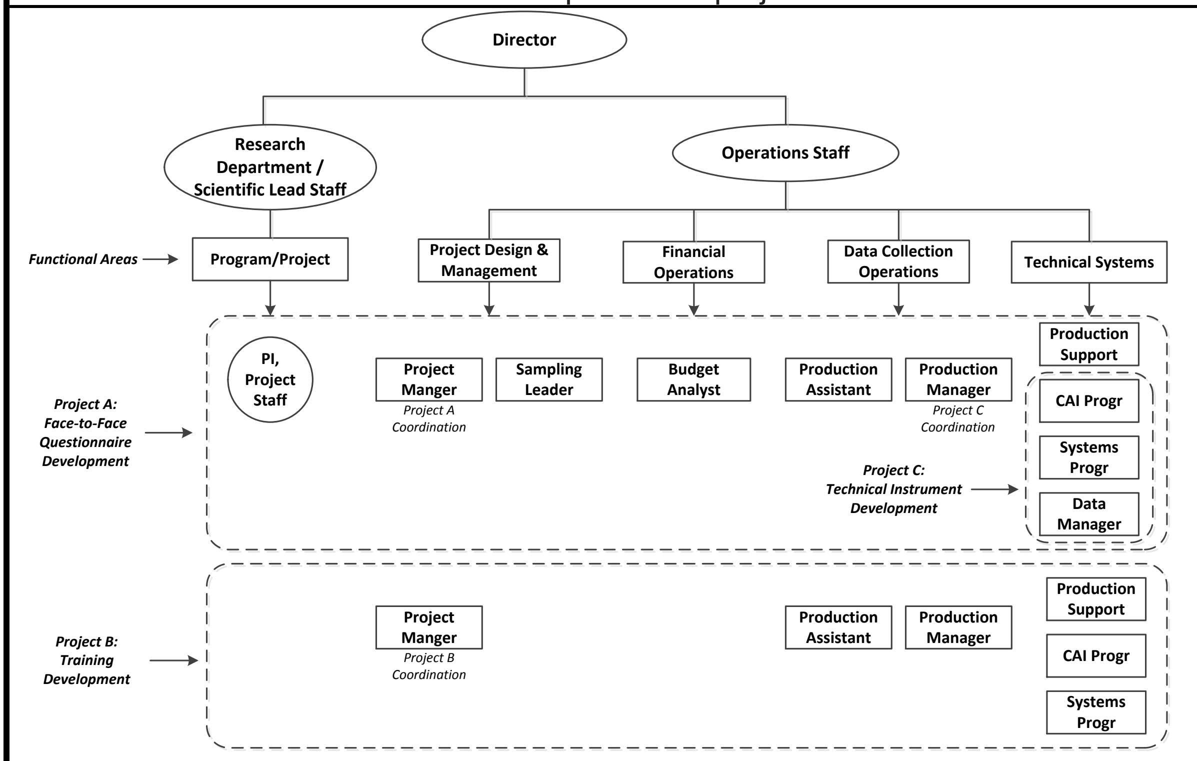
But a specific survey needs to be managed as a distinct **project,** separate from the program(s) it may be associated with. A project (as opposed to a program) has some characteristic features. It is built around a specific and unique goal or set of goals. The project has a beginning and an end. There are limited resources, often revealed through the budget, for implementing the project. The goal or goals are closely tied to the research questions and the methods used to help answer these questions (see Guideline 1 of Study Design and Organizational Structure). The project ends after the goals have been met and all assigned resources have been utilized.

Round 8 of the ESS, for example, is a distinct survey project. There are specific goals which include standard ESS objectives plus new items representing changes from Round 7. The Round 8 survey has a beginning and an end (targeted as May 2015 – October 2017). The resources and budget for implementing Round 8 are pre-determined. The research questions and methods are outlined in the Round 8 Survey Specification (European Social Survey, 2015b). The project ends after the team members finalize and disseminate all Round 8 deliverables.

Note that not all projects are executed as part of the ongoing operations of an organization. Some projects are implemented within new and independent organizational structures that are formed specifically for the project and that dissolve at the end of the project.

The organizational structure chart below shows how a hypothetical organization might utilize employees from several program areas to implement a project. In this case, the organization has a top-level director (sometimes called minister or department head or other). This organization has two program areas including a research department and an operations department (where there are five specific units). The project structure includes individuals from all the program areas. The individuals are assigned to three sub-projects and the sub-projects are coordinated by three managers. Note that the set of individuals assigned to work on “sub-project C” are also separately assigned to work on “sub-project A”.

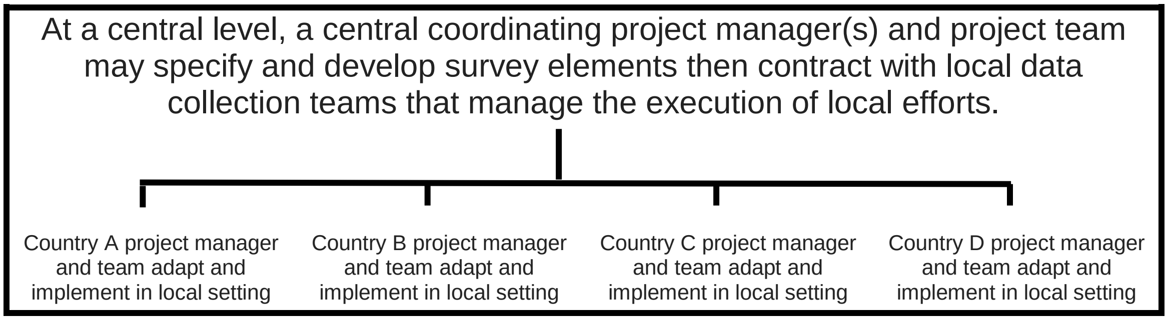
Example of how an organization might utilize employees from several program areas to implement a project



Some survey projects may be unidimensional, where a single project manager and project team handle all the roles and activities to meet the goals of the project. For example, a country’s education ministry might delegate responsibility to a project manager for implementing a survey of paper-based and self-administered questionnaires to regional education superintendents. A small project team might work with the project manager to handle all aspects including sampling, questionnaire development, pretesting, data collection, data processing, and final reporting. All the project phases and processes discussed in this chapter would apply to the tasks carried out by such a unidimensional team.

On the other hand, 3MC surveys typically include multiple dimensions where a central project management team may coordinate the efforts of several country-based or regional-based project teams that each individually implements local components of the survey. For example, a university-based coordinating team might plan and oversee a project that includes contracts with four local data collection teams in separate countries. The central coordinating team (see

Guideline 2 of [Study Design and Organizational Structure](#)) and the local country teams would focus on different elements of the survey production lifecycle. But all the project phases and processes discussed in this chapter would apply to the tasks carried out by both the central coordinating and the local country teams.



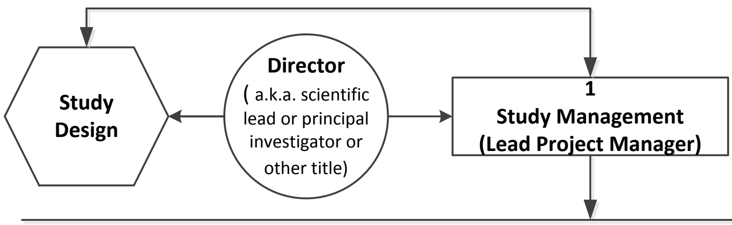
The relationships between many organizations and entities involved in a given 3MC project become increasing complex as multiple countries and cultures join the effort. Distinct study management efforts (for example, discrete management plans) may be warranted at several levels. The coordinating center, each local country team, and each field data collection company may enact the principles that will be discussed in following sections. High levels of communication and clear accountability are required so that multiple study management efforts on a given survey remain synchronized.

Project Management Key Roles

The **project manager** role. The overall director (who might be known as the organization's director, or the scientific lead, or principal investigator, or another title) typically delegates study management authority to a lead project manager. The project manager (who might also be the project director) is the person responsible for accomplishing the project or sub-project objectives. This means completing the project or sub-project on-time and within budget while meeting project specifications and quality. Project managers plan and direct a sequence of activities which involves: identifying requirements, addressing needs, concerns and expectations of stakeholders, maintaining strong communication channels, balancing competing constraints, and completing core processes and phases of the project. An effective project manager has well developed technical skills specific to project management, a strong understanding of the content area for the project, and excellent leadership skills.

The **project team**. The project team is comprised of the people who have assigned roles and responsibilities for completing aspects of the project. Project team members have budgeted effort and cost that will be monitored. The lead project manager may delegate oversight of components of the study to project leads. In the diagram below, twelve project leads have responsibilities for overseeing work and people associated with each of the elements in the survey lifecycle. For larger and more complex projects (as in the diagram), there may be a separate project lead for each element. For smaller projects, any given project lead and project team member may have oversight responsibilities associated with a set of many of the elements.

Project leads may have responsibilities for overseeing work and people associated with the elements in the survey lifecycle.



- Overall project leadership
- Project management plan
- Team meetings and assignments
- Progress monitoring
- Budget and scheduling
- Assure survey quality
- Assure ethical considerations

2 Tenders, Bids & Contracts Lead

- Prepare tenders with detailed requirements
- Conduct a bidding process and select survey organizations
- Negotiate and execute contracts

3 Sample Design Lead

- Define the target population and determine the sample size
- Identify the sampling frame
- Implement a selection procedure

4 Questionnaire Design Lead

- Select a comparative question design approach
- Develop protocols for evaluating questions
- Adopt questions, adapt questions and write new questions

5 Adaptation & Translation Lead

- Identify adaptation needs
- Modify the questionnaire content, format or visual presentation
- Adapt design features
- Find, select and brief translators
- Use existing or develop translation tools
- Complete language harmonization

6 Instrument Technical Design Lead

- Develop design specifications for instruments and a sample management system
- Develop interface design and programming guidelines
- Determine testing specifications
- Determine reporting specifications

7 Interviewer Recruitment, Selection & Training Lead

- Determine required characteristics of interviewers
- Recruit and hire interviewers
- Select interviewer trainers
- Create a training plan and determine the necessary training materials which may involve identifying existing materials or preparing new training materials

8 Pretesting & Data Collection Lead

- Determine the appropriate pretest method and design
- Conduct a pilot study
- Pretest the survey instrument with the target population
- Select the appropriate data collection mode and develop procedures for that mode
- Establish a protocol for managing the survey sample
- Manage data collection and quality control
- Consider potential risks and necessary backup plans if goals are not met

9 Paradata & Other Auxiliary Data Lead

- Investigate the paradata / auxiliary data available and informative to survey errors
- Choose appropriate paradata indicators for survey error and monitor the indicators starting at the initial phases of data collection
- Implement interventions by altering the active features of the survey in subsequent phases or at real-time of the data collection based on cost/error tradeoff decision rules
- Perform analysis using paradata to investigate survey errors

10 Data Harmonization Lead

- Determine a harmonization strategy
- Determine the technical specifications of the system used for data harmonization
- Use a systematic approach to harmonize variables
- Compare and integrate information across data files

11 Data Processing & Statistical Adjustment Lead

- Code survey responses and enter them into electronic form
- Edit and clean data
- Define data quality checks
- Develop survey weights

12 Data Dissemination Lead

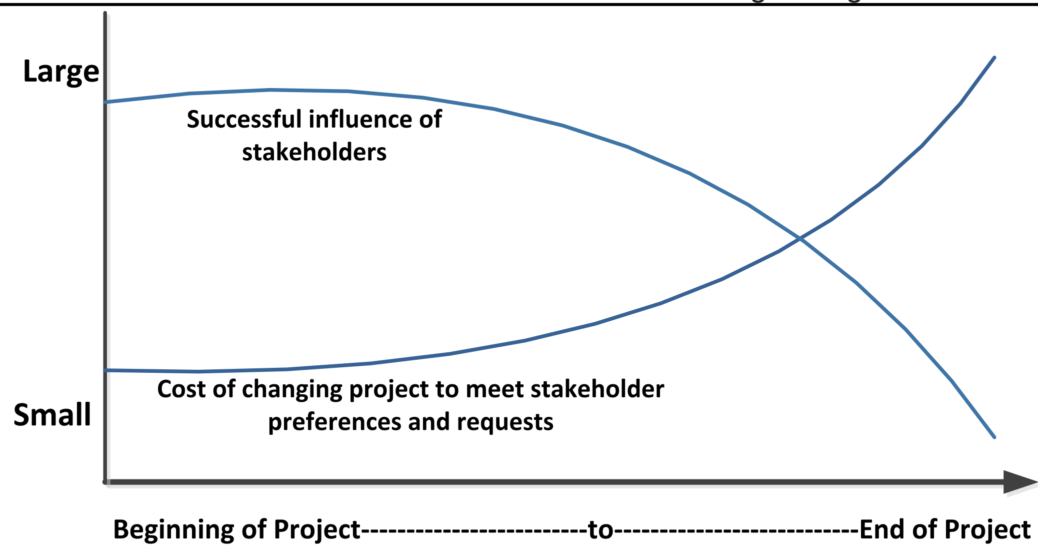
- Preserve key data and documentation files
- Produce public- and restricted-use data files
- Prepare final data deliverables and reports

13 Statistical Analysis Lead

- Apply statistical procedures to data files

The **stakeholders**. The PMI PMBOK defines stakeholders as follows: “A Stakeholder is an individual, group, or organization who may affect, be affected by, or perceive itself to be affected by a decision, activity, or outcome of a project.” Examples of possible stakeholders in 3MC project include the funding sponsor, government representatives in participating countries, a technical advisory board, survey respondents, users of final reports and data systems, and others. Project managers and teams need to influence but do not directly manage stakeholders.

Stakeholder influences diminishes and costs for making change increases.



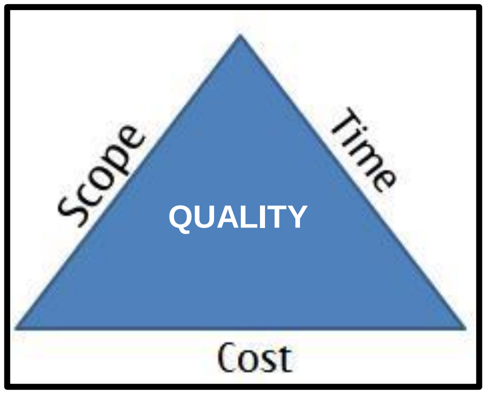
It is most useful to obtain stakeholders’ input at early stages of the project (or lifecycle element) when stakeholders can successfully influence design decisions. If stakeholders try to influence the project (or lifecycle element) at later stages, there will likely be costly revisions and the impact of the stakeholders’ input may be less successful.

Project Management Core Areas

Project management encompasses the application of knowledge, skills, tools and techniques to accomplish the project goals. Some of the broad elements of project management are:

- defining goals, specifying requirements and establishing clear/achievable objectives;
- collecting input from team members and stakeholders then weighing benefits and costs of different approaches;
- balancing competing project constraints of scope, time and cost;
- managing team member and contractor activities; and
- producing and delivering the projects’ services and final products.

There are three dominant constraints (often called 'triple constraints') that interact on all projects: scope, time and costs. Quality influences the triple constraints and itself can be influenced by attempts to balance scope, time and costs.



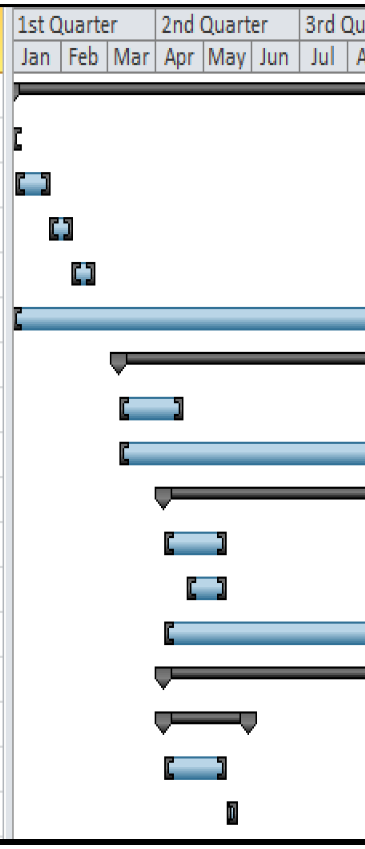
The **scope constraint**. Scope pertains to the work necessary to deliver a product. Scope is about project scope and product scope. Project scope includes the work and activities necessary to complete all the deliverables and requirements. Product scope involves the requirements (e.g., statistical soundness, technical feasibility, ethical integrity, usefulness) that indicate when project deliverables are acceptable.

Scope considers a project's boundaries: what work will be completed during the project lifecycle, and also what work will not be included. Project teams watch out for scope creep, which is a phrase used to describe uncontrolled changes or growth in the scope that must be constrained or may be harmful to costs, schedule and quality.

The **time constraint**. To assure project success, all aspects of the project need to be completed in a timely manner. Project teams typically use schedules to track and adjust time constraints throughout the project lifecycle. The process includes several steps which, depending on the nature of the project might be completed all at once by a single person or might be completed in stages by multiple people. These steps are: creating a detailed list of the activities; putting the activities into a sequence (earliest to latest); estimating the duration of each activity (the work effort and days); and putting them into one or multiple schedules. The following example includes columns for common items on schedule templates. This example also includes a graphical display of scheduling information. When the graphical display is present, this schedule format is called a Gantt chart (see also Tenders, Bids and Contracts, Appendix A).

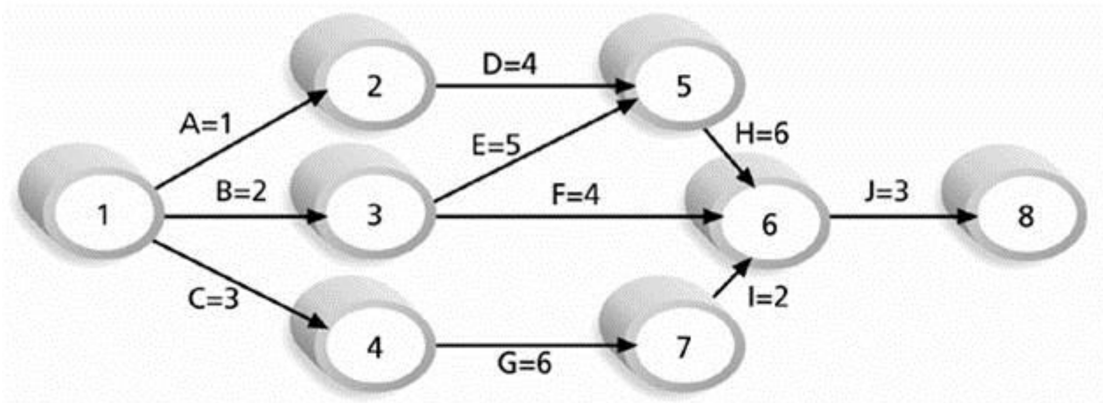
This schedule format is called a Gantt chart.

Task Name	Duration	Start	Finish
1: Study Management	**783 days**	**Wed 1/1/20**	**Sat 12/31/22**
1.1 Hold Kick-off meeting	1 day	Wed 1/1/20	Wed 1/1/20
1.2 Draft project summary	16 days	Fri 1/3/20	Fri 1/24/20
1.3 Draft scope statement	11 days	Fri 1/24/20	Fri 2/7/20
1.4 Draft project charter	11 days	Fri 2/7/20	Fri 2/21/20
....ETC	784 days	Wed 1/1/20	Sat 12/31/22
2: Tenders, Bids, Contracts	**735 days**	**Mon 3/9/20**	**Sat 12/31/22**
2.1 Seek tender proposals	30 days	Mon 3/9/20	Fri 4/17/20
....ETC	736 days	Mon 3/9/20	Sat 12/31/22
3: Sample Design	**715 days**	**Mon 4/6/20**	**Sat 12/31/22**
3.1 Develop sample specs	30 days	Mon 4/6/20	Fri 5/15/20
3.2 Develop weighting plan	20 days	Mon 4/20/20	Fri 5/15/20
....ETC	716 days	Mon 4/6/20	Sat 12/31/22
4: Questionnaire Design	**715 days**	**Mon 4/6/20**	**Sat 12/31/22**
4.1 Prep Respondent Man	**40 days**	**Mon 4/6/20**	**Fri 5/29/20**
4.1.1 Dev draft R manual	30 days	Mon 4/6/20	Fri 5/15/20
4.1.2 Client review	6 days	Fri 5/15/20	Fri 5/22/20



Throughout the life of the project, the project team continuously reviews the schedule(s) and periodically makes adjustments when actual progress occurs differently than originally estimated on the schedule. Some teams use a tool called the "Critical Path" to consider the best ways to adjust a schedule. In a critical path diagram, the tasks are listed in sequence and according to their dependencies. The longest path identifies the timelines for the set of tasks that would need to be adjusted if the overall schedule needs to decrease. The time period for the critical path might be shortened if additional people and their work efforts are added to a task (called crashing) or if some tasks in the sequence have slack and/or they can be rescheduled in parallel with earlier tasks (called fast tracking). The diagram, below, explains how to find the critical path.

Since the critical path is the longest path through the network diagram, Path 2, B-E-H-J, is the critical path for this project.



Assume the durations are shown in months.

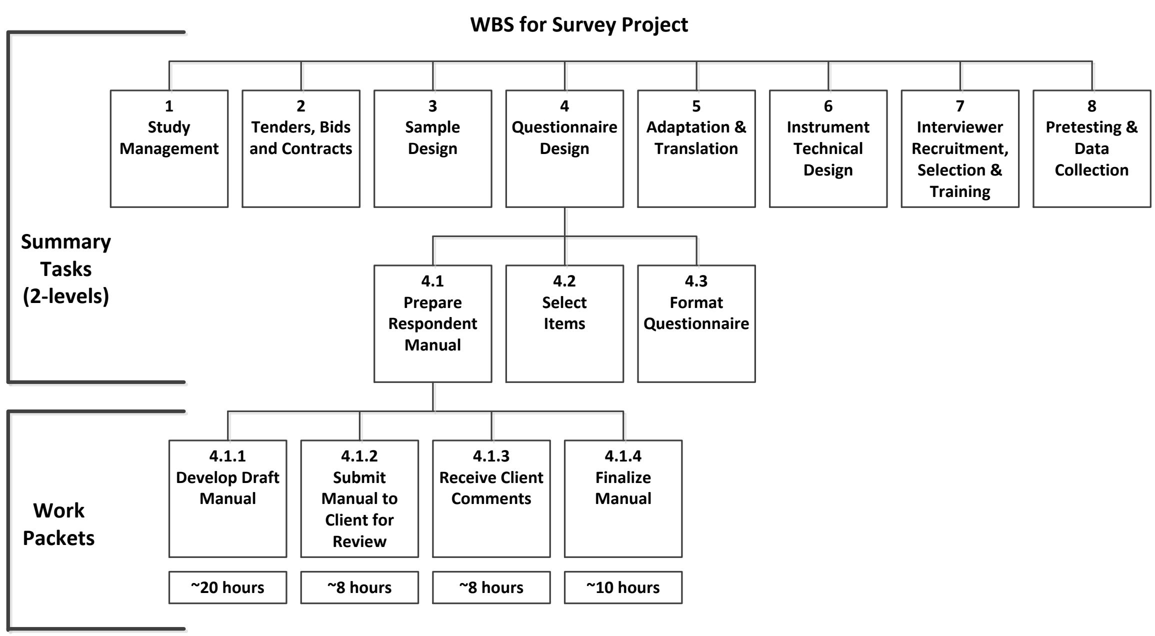
Path 1:	A-D-H-J	Length = 1+4+6+3 = 14 months
Path 2:	**B-E-H-J**	**Length = 2+5+6+3 = 16 months**
Path 3:	B-F-J	Length = 2+4+3 = 9 months
Path 4:	C-G-I-J	Length = 3+6+2+3 = 14 months

The **cost constraint**. Staying within the budget is one of the most important expectations faced during study management. Usually, 3MC projects start with a rough initial budget that was part of the tender or bid packet that led to the project award. The project manager and/or team will then build a more detailed budget by estimating costs to complete work specified in the activity lists and deliverables schedule. The budget includes labor items (salaries and relevant employee benefit costs), and non-salary items (travel, equipment, materials, contractor costs and such).

One tool that can help with labor cost estimation is called work breakdown structure (WBS). A WBS divides the work activities into small pieces that can be assigned to workers and that can be tracked to assure the project stays on track. A WBS has two levels – the summary tasks and the work packets. Summary tasks pertain to elements of the project such as”create the questionnaires’ and ‘train the interviewers”. There may be a second level of summary tasks. For example, “produce paper versions of questionnaires” may have a second level that includes “produce paper versions of the parent questionnaire” and “produce paper versions of the children questionnaire”. For complex projects, there may be many levels of summary tasks.

The work packets are the lower level tasks that provide details of work that will be assigned. As a rule of thumb, work packets might require about eight to eighty work hours to complete – a reasonable numbers of hours that can be specified and that makes sense as an assignment. In the WBS diagram, below, there are four work packets for the summary task called "4.1 Prepare Interview Manual." The work packet 4.1.1 might require about 20 work hours and might be completed by a small team. The work packet 4.1.2 might require about 8 work hours and might be handled by one project manager. That same project manager might complete the work packet 4.1.3 which might require about 8 work hours. And the work packet 4.1.4 might require 10 work hours that the small work team members could split.

A work breakdown structure has two levels–summary tasks and work packets.



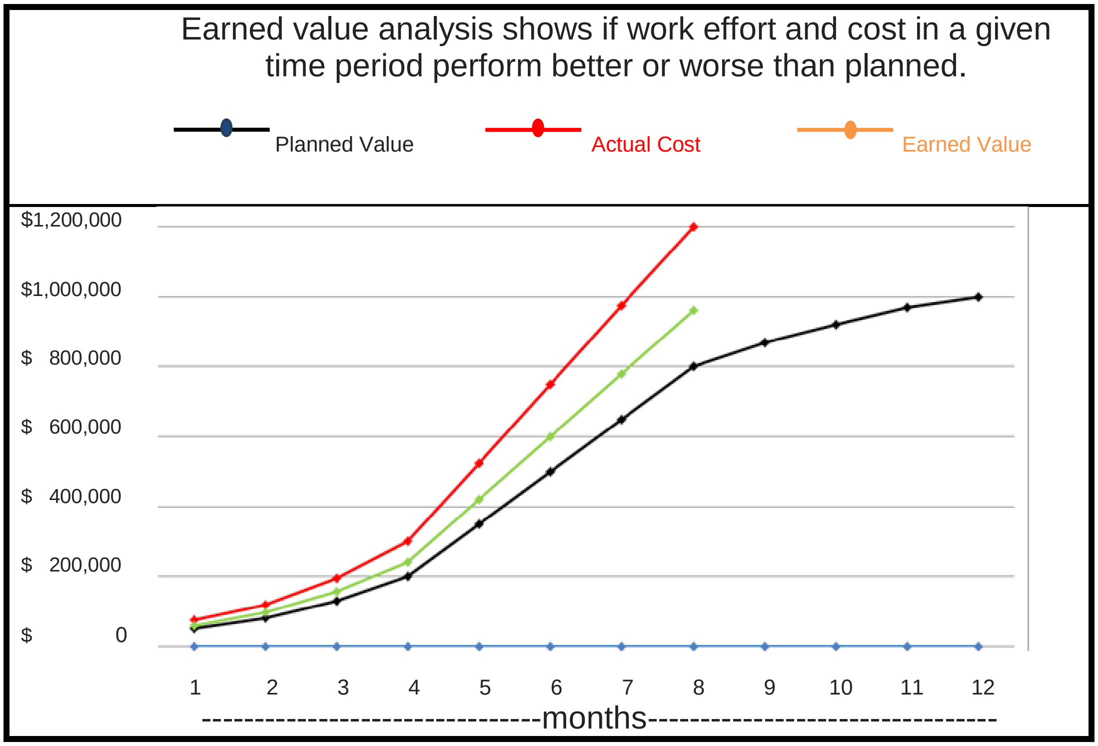
Project managers will consider how to distribute the work packets to specific team members. All the work packet hours for a given team member will be summed and multiplied by that person's hourly rate (cost per unit). Adjustments can be made if the overall WBS budget total does not match the budget amount that the project was awarded.

After estimates are created, the project manager will input all team members' information plus the non-salary cost estimates into a master budget. A basic master budget example is given, below.

Basic master budget example.

	Number	Number of Months	Cost per Unit	Total Cost
Base salaries				
Project Manager				
Data Manager				
Fieldwork Manager				
Accountant				
Assistants				
Supervisors				
Interviewers				
Data entry operators				
Drivers				
Translators				
Computer programmers				
Incentive payments				
Travel				
Researchers				
Interviewers				
Materials				
Computers				
Printers, etc.				
Computer/printer supplies				
Photocopier/Fax machine				
Office supplies				
Communications (phone, fax, postage, etc.)				
Equipment maintenance				
Printing costs				
Questionnaires				
Training manuals				
Reports				
Miscellaneous (maps, listings, manuals, etc.)				
Consultant costs				
International consultants				
International per diem				
Local consultants				
Local per diem				
Local travel				
Contingency (100%)				
TOTAL COST				

As the project is implemented, there are different techniques for monitoring whether project expenses are in line with project progress. Some projects incorporate the cost information into the project schedule so they can project the expected budget balance at several defined time points. Project management software can be useful for this. Some projects use a technique called ”earned value analysis” which determines if the work effort expended and the costs incurred (i.e., actual cost) yielded the expected progress (i.e., planned value).



The **quality** dimension. Project managers are always thinking about quality issues as they balance the ‘triple constraints.’ The guidelines on Survey Quality provide guidance for assessing and managing quality of 3MC survey deliverables and outputs (for example, data sets and response rates and bias). Those guidelines discuss project management processes and tools used to address quality, including: quality planning, quality assurance steps, as well as monitoring and controlling activities.

Beyond these core project management areas, there are other areas, including: human resources, communications, risk, procurement and change-management. These may be discussed as part of the core areas in the base management plan or these may have distinct importance on some 3MC projects and thus warrant their own management plans. More information about all project management areas are available through the PMI website (Project Management Institute, 2016).

Project Management Phases

There are four **project management phases** that all projects pass through: initiation, planning, executing and closing phases (some project teams may break the executing phase into two separate pieces – implementing and controlling/monitoring). During the initiation phase, the project managers gather information about the project and obtain authorization to move forward with project work. During the planning phase, the project managers create a management plan that addresses the core areas of scope, time, costs, and quality as well as other relevant areas (human resources, communications, risk, procurement and change-management). During the executing phase, the project team members complete the work activities and produce survey deliverables using project management monitoring and controlling processes. During the closing phase, the project managers archive project elements, obtain acceptance of deliverables, and document the lessons-learned which may help future survey efforts.

The guidelines below outline activities that are completed in each of these phases. In large and complex projects, there may be sub-project components that independently pass through the phases at different times. The four core project management phases are useful for all projects, regardless of size and complexity.

Guidelines

Goal: To establish a project structure for managing the 3MC survey lifecycle and to use project management processes and tools to effectively complete the study management phases: initiation, planning, execution, and closing.

1. **Implement the Initiation Phase of the project.**

   Rationale

   After there is a trigger indicating the project will occur (e.g., a funding announcement), it is important to establish that project sponsors and decision-makers are committed to moving forward. Early in the initiation phase, project managers educate project leaders in ways to help everyone agree about the goals and approaches. In this phase, project leaders authorize the project managers and team to carry out the project.

   Procedural Steps

   1.1 Develop a clear understanding of the project. This begins by clarifying the research questions, aims and objectives described in Study Design and Organizational Structure. Incorporating and sharing this understanding will help keep project leaders, project stakeholders and team members aligned. Steps that can be taken include:

   1.1.1 Create a project summary that is easy to share and can be included in future documents. The summary can include:
   - A problem statement that describes what needs to be solved and why;
   - Project goals (which integrate research aims and management aims) that are high level targets and that state the end results; TALIS

- Clear, measureable and realistic objectives that provide specific details for the goals and may include: research objectives, financial objectives, business objectives, quality objectives, technical objectives, performance/completion objectives and such.

1.1.2 Begin to create a detailed list of project stakeholders and anticipate their relationships to the work activities.

1.1.3 Determine the approaches that will work best within the culture and organizational structure of the project. For example, if there are multiple project managers, determine whether each will have a distinct budget to manage or if all managers will use the same master budget. There will be different strategy options and it may be helpful to assemble project team members to brainstorm ideas, assess feasibility and consider the desirability of different approaches.

1.1.4 Gather requirements which include details of what the research and management outcomes will look like. Some requirements might be stated in the contract with the sponsor (e.g., the project might need to complete certain deliverables before the second year of funding is released). Some requirements will describe the products (e.g., the sample design, survey questionnaires, etc.) and details might be gathered from stakeholders including the sponsor and the principal investigator. To verify understanding of product requirements, some projects create prototypes before moving into the planning phase. Many of the requirements will pertain to the management aspect: what are the quality standards, what ethical issues need to be met, what milestones belong on the schedule, what level of expertise is required from scientific/technical human resources, what are the budget issues?

1.1.5 List the deliverables and each of their success criteria. For example, the list might include the item "survey questionnaire" and the success criteria might include "short enough so response rate is not compromised" and "sufficiently tested so response bias is low/acceptable".

1.1.6 Identify the assumptions for the project. Stakeholders may have specific expectations. For example, household respondents may expect that husbands may join their wives' interviews. Or, data end users may expect to receive a specific file format. Early interactions with stakeholders can help avoid later misunderstandings.

1.1.7 Identify potential risks.

1.1.8 Gather information from past similar projects.

1.2 Draft a scope statement. A scope statement provides the road map that guides the project team throughout the project and lets everyone know what is expected. As quoted from baseball-legend Yogi Berra, "You've got to be very careful if you don't know where you're going, because you might not get there" (FREEP, 2015).

1.2.1 The components of a typical scope statement are listed below.

- Scope description, based on project summary.
- Deliverables list, which may be elaborated during planning.
- Acceptance criteria, which indicates what the sponsor and/or director (a.k.a. principal investigator or scientific lead) require in order to accept the deliverables, and may include quality elements.
- Exclusions, which include those things that are out-of-scope for the project team.
- Constraints, which are factors that may limit or have an impact on final project results.
- Assumptions, which may be a list of major cost drivers that impact the deliverables (e.g., the period of performance and the target response rates).
- Staffing/scheduling plan indicating the individuals that will work on the project during what time periods.

1.2.2 Use the scope statement to define the boundaries of the project – what team members should expect to work on and especially what is out-of-bounds. For example, separate from the project budget, the sponsor might deliver a sample frame from a “sampling database vendor” to the project team. In this case, it would be out-of-scope (and duplicative) for the project team to design and create or to seek a sample frame for the project.

1.3 Write a project charter and obtain signatures from the project decision-makers. The project charter is a high-level document that provides a synthesis and authorizes the project. It is short and concise and does not change over the life of a project unless there is a dramatic scope revision. Signatures are added to the document to: a) demonstrate the project team commitment; b) provide authorization to start the project from the top-level decision makers; and c) specify the project manager, PI, or scientific lead, additional managers, and pre-committed core staff members.

1.3.1 The components of an example project charter are listed below.

- Basic project information such as name and related projects.
- Project management team members and internal authorities to whom the team reports.

- Sponsor and significant stakeholders (often called the customer).
- Business objective (why the performing organization is interested in conducting the project).
- Project objective (top-level statement of research aims).
- Deliverables (top-level list of primary deliverables).
- Risks, constraints and assumptions (top-level lists that all parties understand and agree can be tolerated).
- Schedule milestones (most significant events on the schedule).
- Overall budget.
- Signature section.

Lessons Learned

1.1 Taking time to complete the steps in the initiation phase yields great benefits. The process helps build team cohesion and sponsor confidence. The outputs are useful for future communications and they facilitate the steps in the other project phases. Especially important, the initiation activities result in full commitment to the project from team members and stakeholders.

1.2 Sponsors and directors (e.g. principal investigators / scientific leads) will vary on how much they desire to participate in the initiation phase. They may not be interested in learning specific project management vocabulary or tools. And, they may not understand the value behind disciplined use of documents like the scope statement and project charter. Whether or not these specific documents are shared, the decision makers will need to be involved in two things: verifying the project scope and budget, and providing authorization for the project team to launch the study.

1.3 Establishing effective communication channels with project stakeholders early in the project can help reduce the chance that barriers will slow down the project progress. During the initiation phase, project managers should be able to identify most of the stakeholders they need to consider during the project life cycle.

1.4 As the project team begins to form, project managers should spend time informing team members about the elaborating details of the project and seeking input from those individuals that have knowledge about specific elements of the project. It takes time to build an effective team. Especially in the early phase of the project, project managers will benefit if they invest time building team understanding and commitment.

1.5. When there are clear sub-projects, especially when they have their own distinct budgets, the relevant project team members may participate in multiple initiation phase activities. For example, a late decision to fund the creation of a data repository at the end of a project might result in a distinctly budgeted sub-project and a few of the project team members will carry out the initiation phase activities when it is time to launch this component.

2. Implement the 'Planning Phase' of the project.

Rationale

Project plans are used to guide how projects will be implemented, including what will be done, who will do what, how stakeholders and team members will receive information, how progress will be tracked, ways the project plan might be corrected when risks are encountered and how quality will be ensured.

Procedural Steps

2.1 Hold an initial planning meeting with all the team members which might include review of the documents developed during the initiation phase, discussion of team members' roles/assignments across the project elements, and consideration of the project milestones.

2.2 Working through the elements in the production lifecycle, create activities lists and work breakdown structures.
- 2.2.1 In addition to the elements in the production lifecycle, use scope and deliverables documents to list top level summary activities.
- 2.2.2 Under each summary activity, list the specific tasks.
- 2.2.3 Create groupings that can be completed in 8-80 hours (i.e., WBSs).
- 2.2.4 Describe the work using a detail level that clarifies 'what to do' but also recognizes that the person or team that will implement it knows more about 'how to do it'.

2.3 Put together the project schedule. There are project management software packages that provide technical tools for developing schedules. Some software integrates schedules with resource allocation and budgets. 3MC projects may have sub-projects and some teams might develop separate schedules for the sub-projects. Steps for developing the schedule include:
- 2.3.1 Put activities in order from earliest to latest.
- 2.3.2 Estimate the duration of activities.

- Most projects estimate activities' durations from the top to the bottom (major phases broken to smaller) but it sometimes works better to estimate from the bottom to the top.
- If available, historical information from past projects and input from experienced colleagues might help with estimating.
- Techniques can be used to analyze whether or not the combined durations can be accomplished within the scheduled project period. The critical path tool (discussed in the Introduction of this chapter) uses activity durations to find the longest sequence of tasks in the project. The 'program evaluation and review technique' known as PERT graphically displays the pathway and durations between milestone activities. Information about these techniques is available at the PMI website.

2.3.3 Specify the dependencies among activities – what is required to be complete before each next task? The most commonly used type of dependency is 'finish-to-start' (finish task-A then start task-B). Other types of task dependencies include 'finish-to-finish,' 'start-to-finish,' and 'start-to-start.'

2.3.4 Place the activities on a calendar.

- Consider if there are pre-determined deadlines for some tasks then work backwards from them.
- Include start and end dates for the activities based on the duration estimates.
- Add milestones which are key project events that don't have durations but mark important things like achievements and due dates.

2.4 Assign activities to individuals.

2.4.1 Based on activity durations, determine the number of hours needed then consider how many individuals are needed. Then, based on their available hours per day, determine how many days the activity will take. Take into account that individuals will have other commitments (for example, department meetings) and some may work on multiple projects in the same week.

2.4.2 Consider how efficient individuals might be when estimating the hours and days needed. Think about all things that influence productivity – for example, multi-tasking tends to decrease productivity.

2.4.3 Integrate the calendar and the results of assigning activities to see if any changes in the calendar are needed. For example, is the overall schedule delayed because some specialists are only available to start activities at later points than anticipated?

If needed, consider techniques for shortening the schedule such as 'crashing' or 'fast-tracking' as discussed in the Introduction of these guidelines.

2.4.4 Obtain agreement from staff supervisors that the project plan can include the individuals for time periods and numbers of hours desired.

2.5 Specify details about the people and resources.

2.5.1 Add the names of individuals to the schedule and using actual pay rates, calculate salary cost of the labor.

2.5.2 Create a responsibility matrix which designates who leads and who works on major areas of the project.

2.5.3 Create an organizational chart for the project.

2.6 Create a project budget.

2.6.1 The project budget will include both labor costs and non-salary costs. A summary budget may include a line (row) for each broad category, for example, managers, programmers, supervisors, interviewers and such. A detailed budget may include line (row) for each individual (by name) and each specific non salary item in the budget.

2.6.2 Typically, there is a rollup budget showing the total budget combining all years of the study as well as individual year-by-year budgets. Often this is accomplished by using tabs in a spreadsheet for yearly budgets that link to a master tab/sheet with the rollup budget.

2.6.3 The budget may be broken in to finer time periods if these will be needed for monitoring and reporting purposes.

2.6.4 Some projects hold a budget line with 'contingency funds,' that is undesignated funds that can help cover unanticipated costs.

2.7 Assess and plan for project risks. "Risk can be defined as the function of three variables, an event that could disrupt the project, the probability that the event could happen, and the impact the event will have on the project if it does happen" (Cook, 2005). Every project faces the chance that anticipated risks, as well as risks that were never imagined, can interrupt the project. By planning ahead, the impact of risks on the project can be reduced.

2.7.1 During early planning stages, the project team should attempt to identify potential project risks. Team members as well as others that have handled projects in the past may be able to help create lists of risks. There are several general areas that may introduce risk. Examples of conditions that may increase risk include: high-levels of project complexity, new technology that may be poorly tested or not work as promised; geographic dispersion of team members that may increase miscommunication; and lower-level of experience by team members.

2.72 There are two key questions that can be asked for each identified risk: what is the potential impact if the risk occurs and what is the likelihood that the risk might threaten the project?

2.7.3 For planning purposes, the team might use a form to list and analyze the risks.

2.7.4 For each risk, the team can indicate a response. Some of the techniques include:

- Acceptance (planning no action and living with the consequences which makes sense if the costs of handling the risk are greater than the cost of the risk).
- Avoidance (changing the project plan to eliminate the risk).
- Transference (hiring a third party to handle the risk).
- Mitigation (making small project updates that won't eliminate but will reduce the probability or impact of the risk).

2.8 Determine the communication needs for the project. It is especially important to establish good communication when a project is complex and geographically dispersed, like many 3MC projects. Steps to help plan for effective communication include:

2.8.1 Determine what stakeholders and team members need to know about the project.

2.8.2 Consider what communication channels work best and under what circumstances.

- When and how often are written/posted status reports most effective?
- When is it beneficial to hold face-to-face sessions?
- Which team members can effectively receive and participate in email exchanges, conference calls and videoconferencing?

2.9 Specify what project changes the team should track and manage. Since all projects experience changes, it is essential to create a plan for change management. A useful strategy is to choose baseline documents, such as the scope statement, the schedule, the management plan, and then handle changes through version control as these documents are updated to reflect change. Some projects require team members to submit written change requests and receive approval before aspects of the project can be amended.

2.10 Compile the written project management plan. Often kept as an electronic document, the project management plan includes sections for the project components and it contains project planning documents and project monitoring materials. The project management plan will address all phases (initiation, planning, execution and closing) and core processes (scope, schedule, cost – all informed by quality). One single project management plan might include all the project components (for example, the survey production lifecycle components) and all the additional processes (communication, risk, change, and procurement). Alternatively, the project team may decide to work with multiple project management plans that break processes and/or components for convenient oversight.

Lessons Learned

2.1. As projects become more complex, it is usually necessary to add extra effort and time specifically to account for the complexity. This is because there will be more interactions with increased communications requirements and extra management needs.

2.2. Many aspects of planning (and project management, in general) require high levels of communication with project team members and stakeholders. Special steps may be required on 3MC projects to account for different languages and communication norms.

2.3. Developing plans that work well in specific local areas may require project managers to consult with local residents or experts. Several examples of ways local stakeholders contribute to the sample plan development are discussed in Guideline 2 of Sample Design. And, Guideline 3 of Questionnaire Design discusses the importance of including local participants when defining the approach for creating questionnaires. Guideline 5 of Instrument Technical Design suggests that plans for usability tests should consider the involvement of interviewers and it discusses issues when the interviewer and participants are from the same or from different cultures.

2.4. Plans that include the use of technology will need to include extra schedule time for development and testing. Guideline 6 of Instrument Technical Design discusses lessoned learned in this regard in Burkina Faso.

2.5. Quality assurance and quality monitoring should be addressed early in the design planning process. An example of how Thornton et al. (2008) handled this is discussed in Guideline 6 of Questionnaire Design.

2.6. The project team should develop and share thorough written documentation of plans and adjustments to plan. Beginning this work in early phases of the project and continuing through the lifecycle will result in high quality final written products.

3. **Implement the 'Execution Phase' of the project.**

Rationale

The "Execution Phase' of the project includes implementing steps, as well as monitoring and controlling activities.

Procedural Steps

3.1 Build and strengthen the project team. The most important assets on a project are the people. Project managers use strong interpersonal and leadership skills to gain high performance from individuals and the overall team. Specific techniques include:

3.1.1 Through the project period, review and clarify the roles and responsibilities each team member holds. Help a given team member understand his or her part and importance on the project. Communicate information to all team members so they understand how their own area of responsibility interacts with their colleagues' areas of responsibility.

3.1.2 Provide specific and achievable goals. Through work packets or other means of assigning work, provide directions that help a given team member determine what will result in 'success' and how to measure progress towards completing goals.

3.1.3 Show respect and be honest in all interactions with the team. Individuals enjoy their work more and perform better when they feel valued and believe they can trust leaders.

3.1.4 Provide feedback to individuals and small teams in a timely manner. When effective, correcting feedback and positive/affirming feedback can help teams understand how well they are performing and help build self-confidence.

3.1.5 Support the individuals and the overall team. If there are obstacles hindering success, provide support and commitment to mitigate the problems.

3.1.6 If the team suffers from 'people problems' (which might include conflicts or mixed commitment to the project or other issues), address these issues immediately.

3.1.7 Devote sufficient time in meeting and communicating about the project with the team members. Encourage team leaders to communicate frequently with their team members. Up to 80% of project management involves communication and this activity will significantly enhance project success.

3.2 Maintain the project planning documents. As the project is executed, many aspects of the original plan need to be updated to fit actual and changing circumstances.

3.2.1 Retain the baseline documents to help with 'lessons learned' and to provide evidence when justifications for changes are required.

3.2.2 Use version control practices to produce and maintain revised planning documents. Team members should have easy access to those planning documents that help them with their work.

3.3 Carry out monitoring activities for all project activities. Do all activities comply with ethics and standards set for the project? Is the specified quality being met? Are all aspects of the scope (product and project) being met? Are there any scheduling problems? Is the work being completed as budgeted? Have new risks been identified?

3.3.1 Gather and analyze data about the survey production elements. The data may cover status, budget, quality and auxiliary areas.

3.3.2 Define what reports will be useful and how they will be produced.

- On a regular basis, gather information from teams (for example, in regular team meetings) and from technical systems.
- Consider using electronic reporting systems and/or web-based 'dashboards' that provide information to team members and stakeholders, on-demand.

3.3.3 Compare data to expectations in the planning resources.

- Status information may focus on progress towards meeting the scheduled milestones. Did tasks begin on time and does the actual duration match the planned duration? A Gantt chart may help identify any variances.
- Cost variances may occur if there are extra hours per task or higher costs for units of work or materials. As discussed in the Introduction of this chapter, earned value analysis can help determine if schedule and cost variances are significant.
- Survey data may be analyzed during the data collection phase to determine if the questionnaires are performing as expected and with acceptable levels of bias (see Paradata and Other Auxiliary Data).

3.4 When projects are 'off-track,' carry out corrective steps in ways that manage and control the changes.

3.4.1 'Responsive design' as discussed in Survey Quality provides guidance for correcting project issues.

3.4.2 Consider whether or not the project manager and project team have authority to implement specific changes. Some changes require higher authority approval.
3.4.3 Consider what type of impact the proposed changes might have on the budget. Might they require approval to utilize contingency funds?

Lessons Learned

3.1 During the executing phase, project team members and their managers may become consumed with work activities related to production of the survey. But more than ever, project managers need to spend time on project management activities in order to keep the project on track. Almost always, this is the most costly portion of the project and it is critical to pay attention to scope, schedule, cost and quality performance.

3.2 The value of communication with team members and stakeholders in this phase cannot be over-stated. As noted in Guideline 4 of Questionnaire Design, not all participating groups in a 3MC project will be confident about providing input. It is important to emphasize that every contribution is valued even when not all suggestions are incorporated into design modifications.

3.3 Implementing quality control protocols from the start of a project permits the survey organization and coordinating center to monitor performance and take corrective action when required. Guideline 5 of Interviewer Recruitment, Selection, and Training, for example, discusses how interviewer certification protocols might be implemented in conjunction with additional interviewer training when interviewer candidates fail to pass on their first try.

3.4 Many organizations have used a quality control technique known as adaptive or responsive design which uses paradata collected during survey implementation to determine if performance such as non-response and response bias indicates the project needs to adapt/correct the original design (Groves & Heeringa, 2006).

3.5 Lessons from other projects can be helpful and the European Social Survey (ESS) provides evidence of success using continuous improvement techniques for planning and implementing the survey (Pennell, Cibelli Hibben, Lyberg, Mohler, & Worku, 2017).

3.6 A growing number of organizations are adopting professional project management frameworks to conduct their project activities. The Project Management Institute (PMI) and the International Project

Management Association (IPMA) provide two of the most commonly followed frameworks. Many organizations follow institute-wide professional project management best practices and encourage staff members to become certified project managers.

4. Implement the 'Closing Phase' of the project.

Rationale

As work is 'wrapped up' for the final elements of the survey production cycle, managers need to take steps to effectively close the project.

Procedural Steps

4.1 Distribute the final deliverables. These deliverables may include data systems, reports, instruments, and other products. See Data Dissemination for additional details.

4.1.1 Data sets for surveys typically need to be transferred with special consideration to assure that privacy and identities of survey participants are protected. Many projects are required to operate under data management plans. These plans may specify:

- Some data may be restricted from distribution beyond the protection of the project's secure storage system and may need to be destroyed during the closing phase.
- There may be specific methods that must be used to transfer data from the project to other parties. Legal agreements such as data transfer and data use agreements may be required.

4.1.2 Final reports and other products may be developed for public dissemination or may be transferred only to the sponsor and specific stakeholders.

- The final report should be nearly complete at the end of the study if the project team keeps the management plan updated throughout the project.
- Final reports and data products may be compiled for hard copy production (for example, the Survey of Health, Ageing and Retirement in Europe (SHARE)) or for on-line access (for example, the ESS).

4.2 Obtain formal notice of acceptance of deliverables.

4.2.1 The project might use basic procedures such as email confirmation that each deliverable is acceptable.

4.2.2 Some projects provide 'acceptance' documents that verify that 'success criteria' are fully met.

4.3 Carry out steps to gather ‘lessons learned’ information.
   4.3.1 Throughout all phases of the project, managers will gather information that can inform a final summation of the ‘lessons learned.’
   4.3.2 At the end of the project, it is useful to hold debriefings with team members and stakeholders to consider ‘what went well’ and ‘what could have gone better.’

4.4 Close all contracts and complete all requirements in legal agreements such as non-disclosure documents, memos of understanding, data use agreements, human subject protections documents, and such.

4.5 Archive project items, including a project close out report.

4.6 Verify that all team members are transitioned off the project.

4.7 Acknowledge successful completion of project. Congratulations!

Lessons Learned

4.1. The last parts of the project ‘execution phase’ may experience very tight timelines which may compromise the timeframe for the ‘closing phase.’ With this in mind, many of the closing activities can be started even as earlier phases are in progress.

4.2. The archiving activities and production of final reports are much easier when project management processes are maintained throughout the project life and when project teams produce on-going survey and project management documentation.

Appendix A - Initial Project Summary Template

Date:
Version:

Project Name:

Project Summary Version Control:

Version	Date	Change Description
1		
2		
3		

Problem Statement (what needs to be solved and why):

Project Goals (high level targets that state the end results):

Project Objectives (clear, measurable, realistic, and specific):

May include:

Research Objectives
Financial Objectives
Business Objectives (what organization gains)
Quality Objectives
Technical Objectives
Performance/Completion Objectives
Other Objectives

Appendix B - Stakeholder Analysis Template

Date:
Version:

Project Name:

Stakeholder Analysis Version Control:

Version	Date	Change Description
1		
2		
3		

Stakeholder Name	Organization	Role on Project	Project Interests (Goals & Motivations)	Importance of Interests (low, medium, high)	Power & Influence (low, medium, high)	Communication Needs

Appendix C - Project Scope Statement Template

Date:
Version:

Project Name:

Scope Statement Version Control:

Version	Date	Change Description
1		
2		
3		

Project Scope Description:

Project Deliverables:

Project Acceptance Criteria:

Project Exclusions:

Project Constraints:

Project Assumptions: (Possible items given, below)

- Production Period:
- Sample Size:
- Total Interviews:
- Interview Length:
- Response Rate:
- HPI:
- Interviewing Hours:

Staffing/Scheduling Plan:

Appendix D - Project Charter Template

Date:

Project Name:
Related Projects:
Project Leaders: [Lead and other managers]
Customer: [Principal Investigator(s), Sponsor]

Other project stakeholders:

Business Objectives:

Project Objectives:

Project Deliverables:

Initial Risks:

Constraints:

Assumptions:

Milestones:

[Period | Phase | Task]	Milestone Outcome	Expected Completion Date

Project Charter (continued)

Top-level Budget: (for example, breakdown by survey lifecycle elements)

Task	Cost
01 – Study Management	
02 – Tenders, Bids, Contracts	
03 – Sample Design	
04 – Questionnaire development	
05 – Adaptation and Translation	
06 – Instrument Technical Design	
07 – Interviewer Selection/Training	
08 – Pretesting & Data collection	
09 – Para & Auxiliary Data	
10 – Data Harmonization	
11 – Data Processing/Stats Adjustment	
12 – Data Dissemination	
13 – Statistical Analysis	
Total:	

Authorization:

The project decision maker has received this charter and authorized the project.

[Signature]
___________________________________ ______________________

Appendix E - Survey Items for Activities Lists Template

Project teams can begin to build detailed activities lists selecting from items in this guideline. By creating a detailed list of survey tasks, the project team can ensure that no aspect of the study structure has been overlooked and can then use this list to assign organizational responsibilities.

Tenders, Bids, and Contracts
- Prepare tenders with detailed requirements.
- Conduct a bidding process and select survey organizations.
- Negotiate and execute contracts.

Sample Design
- Define the target population and determine the sample size.
- Identify the sampling frame.
- Implement a selection procedure.

Questionnaire Design
- Select a comparative question design approach.
- Develop protocols for evaluating questions.
- Adopt questions, adapt questions, and write new questions.

Adaptation
- Identify adaptation needs.
- Modify the questionnaire content, format, or visual presentation.
- Adapt design feature.

Translation: Overview
- Find, select, and brief translators.
- Use existing or develop translation tools.
- Complete language harmonization.

Instrument Technical Design
- Develop design specifications for instruments and a sample management system.
- Develop interface design and programming guidelines.
- Determine testing specifications.
- Determine reporting specifications.

Interviewer Recruitment, Selection, and Training
- Determine required characteristics of interviewers.
- Recruit and hire interviewers.
- Select interviewer trainers.
- Create a training plan and determine the necessary training materials. This may involve identifying existing materials or preparing new training materials.

Survey Items for Activities Lists (continued)

Pretesting
- Determine the appropriate pretest method and design.
- Conduct a pilot study.
- Pretest the survey instrument with the target population.

Data Collection
- Select the appropriate mode and develop procedures for that mode.
- Establish a protocol for managing the survey sample.
- Manage data collection and quality control.
- Consider potential risks and necessary backup plans if goals are not met.

Paradata and Other Auxiliary Data
- Investigate para / auxiliary data available & informative to survey errors.
- Choose appropriate paradata indicators for survey error and monitor the indicators starting at the initial phases of data collection.
- Alter features of the survey based on cost/error tradeoff decision rules.
- Perform analysis using paradata to investigate survey errors.

Data Harmonization
- Determine a harmonization strategy.
- Create technical specifications for systems used for data harmonization.
- Use a systematic approach to harmonize variables.
- Compare and integrate information across data files.

Data Processing and Statistical Adjustment
- Code survey responses and enter them into electronic form.
- Edit and clean data.
- Define data quality checks.
- Develop survey weights.

Data Dissemination
- Preserve key data and documentation files.
- Produce public- and restricted-use data files.
- Prepare final data deliverables and reports.

Survey Quality
- Document the survey process.
- Develop quality standards and a quality assurance plan.
- Monitor and support the implementation of quality standards.

Ethical Considerations
- Create informed consent forms and ensure the rights of respondents.
- Observe professional standards and local laws.

Appendix F - Project Schedule Template

List the beginning and end dates, and key project milestones. Add detail to the schedule as the project moves through the planning phase and tasks are elaborated.

Task/Activity/Milestone	Duration	Start date	End date	Team member

Appendix G - Project Budget Template

	Number	Number of Months	Cost per Unit	Total Cost
Base salaries				
Project Manager				
Data Manager				
Fieldwork Manager				
Accountant				
Assistants				
Supervisors				
Interviewers				
Data entry operators				
Drivers				
Translators				
Computer programmers				
Incentive payments				
Travel				
Researchers				
Interviewers				
Materials				
Computers				
Printers, etc.				
Computer/printer supplies				
Photocopier/Fax machine				
Office supplies				
Communications (phone, fax, postage, etc.)				
Equipment maintenance				
Printing costs				
Questionnaires				
Training manuals				
Reports				
Miscellaneous (maps, listings, manuals, etc.)				
Consultant costs				
International consultants				
International per diem				
Local consultants				
Local per diem				
Local travel				
Contingency (100%)				
TOTAL COST				

Appendix H - Project Management Plan/Manual Template

Date:
Version:

Cover Section:

Project Name:

Version Control:

Version	Date	Change Description
1		
2		
3		

Period(s) of Performance:

	Start Date	End Date
Project		
Data Collection		

Project-Specific Considerations:

Tailored Management Processes:

Financial Accounts:

Project Initiation Documents

Initial Project Summary

Project Stakeholders Analysis

Project Scope Statement

Project Charter

Project Management Planning Documents

Project Organizational Chart

Activities List (possibly built from the Survey Task List Template)

Outline for Project Management Plan/Manual (continued)

Project Management Planning Documents (continued)

Work Breakdown Structure

Responsibility Matrix

Project Schedule

Project Budget

Risk Assessment

Subsidiary Project Management Plan Documents
- **Quality Management Plan**
- **Human Resource Management Plan**
- **Communications Management Plan**
- **Risk Management Plan**
- **Procurement Management Plan**
- **Change Management Plan**

Project Executing Documents

Regular Progress Reports (when element is active)

Note–different projects may use written, or verbal or on-line dashboard reports

Schedule Changes/Updates

Budget Changes/Updates

Quality and Risk Changes/Updates

Client Reports

Project Closing Documents

APPENDIX I - EXAMPLES

The documents referenced in this table contain items that could be compiled into project management plans or other tools discussed throughout the Study Management guidelines. (Other similar examples can be found on the websites of other specific 3MC projects.) Examples of the types of documents listed in the first column (Project Management Items) can be found in the four documents (two from the ESS and two from TALIS) listed in the subsequent columns (under Source Documents). The table provides references to the chapters or sections in the documents where examples can be found. Website links to the source documents are provided, below.

	Source Documents			
Project Management Items	**ESS Blueprint (Numbers refer to chapters)**	**ESS Round 7 Specifications (Numbers refer to chapters / specified Zip Folder)**	**TALIS 2013 NPM Manual (Numbers refer to chapters)**	**TALIS Technical Standards (Numbers refer to chapters or standards)**
Problem statement	2. The Case for a European Social Survey (pages 7-9)		2.1 Background and purpose of TALIS	Chapter 1 INTRODUCTION
Goals/aims (high level)	3. Designing the European Social Survey (pages 10-17)	Main: 1.1. Aims, coordination and funding	2.1 Background and purpose of TALIS	Chapter 1 INTRODUCTION
Objectives (specific/measurable)			2.6 Communication	Chapter 3 COMMUNICATION
Standards / Scope specifications	Appendix 5 Survey Specification of a Proposed ESS (pages 58-60)	3. Specification for survey, ALSO: 4. Time table ESS activities … 5. Preparing the questionnaire 6. Sampling	3 SURVEY PHASES , ALSO: 4 ISCED LEVEL 2 CORE SAMPLE 5 PREPARING INSTRUMENTS 6 SCHOOL COOPERATION	Chapters 2 through 11 provide standards for all aspects of the project scope
Stakeholder analysis	ALSO see the following Chapters: 3. Designing the European Social Survey (pages 17-22) 4. Selection of Themes for European Social Survey 5. Methodological Research 6. Data Management, Archiving and Distribution	7 … MATERIALS TO SCHOOLS 8 ADMINISTER QSTNAIRES 9 QUALITY CONTROL 10 DATA MANAGEMENT	7 … MATERIALS TO SCHOOLS 8 ADMINISTER QSTNAIRES 9 QUALITY CONTROL 10 DATA MANAGEMENT	
Activities lists/ Work breakdown structure		2.2. NC Activities	Specific NPM activities are listed in Scope sections – see above.	
Project / Task schedule		Table 1: ESS7 Project Timetable (April 2013-October 2015)	Table 3.1 Key milestones Field Trial Table 3.2 Key milestones Main Survey – S Table 3.3 Key milestones Main Survey – N	
Budget / Staff hours	8. Costs 9. Funding			Standard 2.6
Organizational chart	7. Organisational Structure		2.2 Project governance	
Roles / responsibilities		1.2. National level appointments	2.4 National Project Managers 2.5 National centres ANNEX F - PROFILE AND ROLE OF NPM	Standard 2.5-2.7 (National Project Manager) Standard 2.8-2.10 (National Centre/3rd party) Standard 6.7-6.8 (School Coordinator)
Risk analysis		10. Quality, comparability and compliance	6.2 Confidentiality and ethics	Chapter 11 CONFIDENTIALITY, SECURITY, INTL DATABASE
Progress Reports		ZIP Folder # 2. ESS7 Fieldwork progress reporting guidance.pdf		Standard 7.15 Standard 8.12

Links to Source Documents	
ESS Blueprint	Original specifications for the European Social Survey with focus on the **central coordinating center functions**. Available at - http://www.europeansocialsurvey.org/docs/about/ESS_blueprint.pdf
ESS Round 7 Specifications	Country-level specifications for the 7th round of the European Social Survey with focus on **national coordinator functions**. Main manual available at - http://www.europeansocialsurvey.org/docs/round7/methods/ESS7_project_specification.pdf Zip folder with various resources available at - http://www.europeansocialsurvey.org/docs/round7/methods/ESS7_project_specification_manual.zip
TALIS 2013 NPM Manual	National Project Manager's Manual for the OECD 2013 Teaching and Learning International Survey Main manual available at - http://www.oecd.org/edu/school/TALIS2013_NPM_Manual.pdf
TALIS Technical Standards	Country-level requirements outlined in standards format for implementation of the OECD 2013 Teaching and Learning International Survey http://www.oecd.org/edu/school/TALIS2013_Technical_Standards.pdf

References

Cook, C. R. (2005). *Just enough project management: The indispensable four-step process for managing any project, better, faster, cheaper*. New York: McGraw-Hill.

Detroit Free Press (FREEP). (2015). http://www.freep.com/story/sports/mlb/2015/09/23/yogi-berra-quotes-isms/72669094/

European Social Survey (ESS). http://www.europeansocialsurvey.org/data/

European Social Survey (ESS). 2015b. *Round 8 survey specification for ESS ERIC member, observer and guest countries*. Retrieved August 17, 2016 from http://www.europeansocialsurvey.org/docs/round8/ESS8_project_specification.pdf

European Social Survey (ESS). (2016). *European Social Survey European Research Infrastructure Consortium ESS ERIC Annual Activity Report 2014-2015*. Retrieved 8/17/2016 from http://www.europeansocialsurvey.org/docs/about/annualreports/ESS-ERIC-Annual-Activity-Report-2014-2015.pdf

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, 439–457.

Johnson, T. P., & Braun, M. (2016). Challenges of comparative survey research. In C. Wolf, D. Joye, T. Smith, Y-Ch Fu, (Eds.): *SAGE handbook of survey methodology*. London, UK: Sage.

Jowell, R., Kaase, M., Fitzgerald, R., & Eva, G. (2007). The European Social Survey as a measurement model. In R. Jowell, C. Roberts, R. Fitzgerald, & G. Eva (Eds.), *Measuring attitudes cross-nationally: Lessons from the European Social Survey* (pp. 79-94). London: Sage

Lyberg, L., Japec, L., & Tongur, C. (2017). Prevailing issues and the future of comparative surveys. In T. Johnson, B. Pennell, I. Stoop, & B. Dorer (Eds.), Advances in comparative survey methods: Multinational, multiregional and multicultural contexts (3MC). New York, NY: John Wiley & Sons.

Mohler, P. Ph. (2007). What is being learned from the ESS. In R. Jowell, C. Roberts, R. Fitzgerald & G. Eva (Eds.), *Measuring attitudes cross-nationally: Lessons from the European Social Survey* (pp. 79-94). London: Sage.

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. West (Eds.), *Total survey error in practice*. New York, NY: John Wiley & Sons.

Project Management Institute (PMI). (2013). *A guide to the Project Management Body of Knowledge (PMBOK guide), fifth edition*. Newtown Square, PA: Project Management Institute.

Project Management Institute (PMI). 2016. https://www.pmi.org/

SHARE. http://www.share-project.org/data-access-documentation.html

Thornton, A., et al. (2008). Creating questions and protocols for an international study of ideas about development and family life. In J. A. Harkness, M. Braun, B. Edwards, T. P. Johnson, L. E. Lyberg, P. Ph. Mohler … T. W. Smith (Eds.), *Survey methods in multinational, multiregional, and multicultural contexts* (pp. 59-74). Hoboken, NJ: John Wiley & Sons.

Further Reading

Biafore, B. (2011). *Successful project management: Applying best practices and real-world techniques with Microsoft Project.* Sebastopol, CA: O'Reilly Media.

Biafore, B. (2016). *Project management fundamental* (on-line video course), https://www.lynda.com/Project-tutorials/Project-Management-Fundamentals/424947-2.html

Frame, J. Davidson. (2003). *Managing projects in organizations: How to make the best use of time, techniques, and people. 3rd ed*. San Francisco, CA: Jossey-Bass.

Lewis, J. P. (2007). *Fundamentals of project management. 3rd ed.* New York, NY: American Management Association.

Prewitt, K. (1983). Management in survey organizations. In P. Rossi, J. Wright, & A. Anderson (Eds.), *Handbook of survey research*. (pp. 123-145). New York, NY: Academic Press.

Tenders, Bids, and Contracts

Rachel A. Orlowski, Christopher Antoun, Rolfe Carlson, and Mengyao Hu, 2016

Introduction

The following describes the process for preparing tenders, soliciting bids, and drawing up and executing contracts. The tenders should be based on the specifications outlined in Study Design and Organizational Structure. Bids should be obtained from as many qualified organizations as possible to ensure a high quality and cost effective survey for the budget available. The goal of entering into a contract is to create a legally-binding agreement between the organization coordinating the study and the organizations collecting the data.

Complications will inevitably arise over the course of the study, so it is important that the mutually agreed upon contract defines, in as much detail as possible, the specifications and expectations regarding procedures, responsible parties, and outcomes for all parts of the study across all participating organizations. Towards this end**,** the tenders, bids, and contracts should all be as specific and detailed as possible (Worcester, Lagos, & Basañez, 2000).

In multinational, multicultural, or multiregional surveys, which we refer to as “3MC” surveys, the tendering, bidding, and contracting process will involve various parties (e.g., the survey organizations, central coordinating center, and funder(s)), and there are many ways for these parties to conduct the process. For example, the coordinating center may prepare tenders and solicits bids from survey organizations in each country where the study will be carried out. The resulting contracts are between the coordinating center and each selected survey organization. There are other situations in which the coordinating center signs a contract with an international organization that is responsible for data collection in several countries. Sometimes, the funder(s) prepares tenders for the central coordinating center and survey organizations separately and the coordinating center submits, rather than solicits, a bid. Thus, there is a contract between the funder(s) and the coordinating center, as well as separate contracts between the funder(s) and local survey organizations. Finally, there are other situations in which the central coordinating center is not involved with any contractual work, and contracts are individually arranged and signed at the country level. In this situation, the central coordinating center may provide specifications and supervise the process, but the contract is an agreement between the local funder(s) and local survey organizations. See Study Design and Organizational Structure for further detail.

The guidelines presented here address the general approach used in the tendering, bidding, and contracting process described in the first case—outlining a competitive bidding process between a central coordinating center and survey organizations (in particular, survey organizations selected at the country level).

Nevertheless, many of the guidelines below also apply to the other contracting arrangements.

Guidelines

Goal: To prepare tenders and conduct a competitive bidding process that will result in detailed contracts within defined budget parameters and an established legal framework.

1. **Prepare a tender based on study specifications while adapting it, when appropriate, for each individual country.**

Rationale

The tender is the first step to soliciting bids and executing contracts. The specifications in the tender will have long-term effects on the final contracts and the implementation of the study. With the tendering process, the coordinating center should consider the amount of risk it is willing to assume and specify the type of contract it will offer. The tender should outline study details and requirements of the bidding survey organizations. Requesting detailed information on technical and business aspects of the survey organization's bid reduces the opportunity for misunderstanding to go unnoticed and helps ensure that the study specifications have been fully understood and adequately accounted for in the plan and budget. In the final preparation of the tender, local adaptations should be considered, and multiple tenders may need to be developed for multiple countries to set reasonable expectations based on the culture and availability of resources.

Procedural steps

1.1 Determine the appropriate tendering process in each participating country.

1.1.1 Decide between open tendering and restricted tendering.

- Open tendering allows any survey organization to provide a bid. It is advantageous because it protects against favoritism. Open tendering is absolutely necessary if the coordinating center is not familiar with the availability of qualified survey organizations in a country.
- Restricted tendering limits the bidding process to a few survey organizations pre-selected by the coordinating center. Restricted tendering is used when the coordinating center has prior knowledge of survey organizations that are capable of implementing their country's portion of a 3MC survey.

1.1.2 Become familiar with the local requirements for tendering (e.g., some countries prohibit restricted tendering if using public funds).

1.2 State in the tender which type of contract will be offered: fixed-price, cost-reimbursable, or time and material. The decision on which type of contract will be offered depends on the level of risk the coordinating center (or funding source) and the survey organizations are willing to take (Project Management Institute, 2004). These guidelines assume cost-reimbursable contracts, but these three types of contracts are defined below.

1.2.1 A fixed-price (or lump-sum) contract requires stating upfront a fixed total price for the clearly-defined scope of work and deliverable(s). Fixed-price contracts may also allow for bonuses if expectations are exceeded. The coordinating center incurs little risk while the survey organizations incur much risk.

1.2.2 A cost-reimbursable contract requires paying the survey organizations for the actual costs necessary to complete the agreed-upon scope of work and production of the deliverable(s); it may include paying them a fee — typically received as profit. Cost-reimbursable contracts also allow for bonuses if expectations are exceeded. This type of agreement is riskier for the coordinating center than for the survey organizations. Thus, it is important for the coordinating center to carefully evaluate survey organizations during the bidding process and to monitor progress during survey design and implementation.

1.2.3 A time and material (T&M) contract has elements of both the fixed-price and the cost-reimbursable contract. Time and material contracts may require a fixed level of effort by a specific class(es) of resources (staff) at the survey organizations or may have a variable level of effort by a specific class(es) of resources at an agreed-upon rate of pay for the specific class(es). These contracts may be open-ended, such that the exact price for the scope of work and/or deliverable(s) may not be determined when signing the contract. This type of contract is rarely used for the implementation of an entire survey project; it is sometimes used when contracting work for a particular task in the survey lifecycle (e.g., contracting with an organization to perform the post-collection data analysis).

1.3 Ask bidders to provide specific technical information about their survey organization and their plan to execute the survey within the study specifications, generally referred to as the Technical Proposal.

As suggested by the International Organization for Standardization (ISO), the proposals submitted by bidders should reference ethical codes in social, market, and public opinion research (ISO, 2012).

1.3.1 Request the following from the survey organization:

- Examples of similar studies the bidder has conducted (describing the size, complexity, topic, etc.).
- Examples of the bidder's training and supervisory materials, details of procedures used, and reports from studies previously conducted.
- References or contact names for previously completed work.
- Number and relevant qualifications of all levels of staff assigned to the study, as well as an organizational chart and outline of responsibilities for this survey.
- Organizational capacity (e.g., size of field interviewing staff).
- Financial capacity (e.g., adequate financial resources to pay staff and expenses until the reimbursement payment is received).
- Technical system capability (e.g., any computer-assisted interviewing, sample management capabilities, and/or data entry software).
- Facilities and equipment (e.g., computers, internet access, and e-mail).

1.3.2 Request the following regarding their plan to execute the survey:

- Timeline with survey tasks, milestones, and deliverables. See Study Design and Organizational Structure for details about creating a timeline and see Appendix A for an example of a timeline of effort.
- Staff responsibilities for each survey task. See Appendix B for an example of a person loading chart describing how responsibilities are assigned.
- Consent, confidentiality, and data protection procedures. See Ethical Considerations.
- Sampling methods (e.g., sample size, type of frame, etc.) See Sample Design and The European Society for Opinion and Market Research (ESOMAR) (2001).
- Questionnaire development and translation methods, if applicable. See Questionnaire Design and Translation.
- Pretesting methods, including pretesting the questionnaire, technical instrument and field procedures. See Pretesting.
- Design of survey instrument. See Instrument Technical Design.

- Interviewer recruitment, selection and training protocol (i.e., number of hours of training, topics covered, etc.) See Interviewer Recruitment, Selection, and Training.
- Interviewer characteristics (e.g., age, education, gender, and experience). See Interviewer Recruitment, Selection, and Training.
- Unique identification of the interviewers
- Interviewer payment plan (typically by the hour or by completed interview). See Interviewer Recruitment, Selection, and Training.
- Interviewer employment structure (i.e., employees of the survey organization or contract workers).
- Ratio of interviewers to supervisors (see Data Collection: General Considerations).
- Mode of data collection proposed, and, if using a mixed mode design, whether multiple modes will occur concurrently or sequentially. See Study Design and Organizational Structure).
- How information about the contact attempts will be collected and reported (e.g., time, day, interim disposition codes). See Paradata and Other Auxiliary Data.
- Production requirements (e.g., minimum number of contacts to attempt to obtain a complete interview, minimum response rate, etc.).
- Local quality monitoring (e.g., evaluating recorded interviews, re-interviews on key survey items). See Paradata and Other Auxiliary Data and Survey Quality.
- Plans in place to address nonresponse bias (see Data Collection: General Considerations).
- Procedures for data transfer during the production period from the field to the study country's office (e.g., mail or electronic transfer of completed survey questionnaires and other materials).
- Procedures for processing, managing, and storing data (see Data Processing and Statistical Adjustment).
- Procedures and methods for providing data to the coordinating center.
- Procedures developed to handle unexpected problems (i.e., risk management) (ESOMAR, 2001).

1.4 Ask bidders to complete a separate Business or Cost Proposal (allowing for concurrent independent reviews of the technical and business/cost proposals). Have the bidders organize the business proposal by each major survey task—tailoring the budget to the specific country's implementation of the study (Federal Committee on Statistical Methodology, 1983). All of the Cross-Cultural Survey

Guidelines could be considered viable survey tasks (see Appendix A in Study Design and Organizational Structure for a brief description of each survey task).

1.4.1 Within each organized survey task, ask bidders to prepare a detailed budget by the two general categories: direct costs and indirect costs (Project Management Institute, 2004). Direct costs typically consist of salary and non-salary costs, and indirect costs are typically calculated as a percentage of some or all of the total direct costs (both salary and non-salary costs).

- Salary costs include: labor (both regular and temporary staff), fringe (calculated as a percentage of the regular staff labor costs), and overhead (calculated as a percentage of the total labor and fringe costs) (Glewwe, 2002). For each staff position, budget the number of hours or percent of effort needed for each staff member for each survey task in which he or she will contribute. See Appendix C for a salary budget example template that specifies labor hours for the pretesting task.
- Nonsalary costs include general sample purchase; supplies (e.g., pencils, folders, binders, etc.); printing (e.g., letterhead, training materials, respondent booklets, maps, reports, etc.); postage; communications (e.g., local and long distance telephone calls and service, high-speed internet connection, etc.); computing (e.g., laptop computers, printers, equipment maintenance, software licensing, security protection, etc.); interviewer recruitment (e.g., advertisements, community meetings, etc.); interviewer training (e.g., hotel arrangements, meals, travel, etc.); interviewer travel during the production period (e.g., lodging, mileage, vehicle rental, vehicle maintenance, fuel, etc.); respondent incentives; and consultant fees (e.g., stipend, per diem, travel, etc.). See Appendix D for a non-salary budget example template that specifies costs for the pretesting task.

1.4.2 Require bidders to provide written justifications for all direct and indirect costs, as well as to be explicit with the budgeting assumptions taken (e.g., the duration/dates of each survey task, the questionnaire length, the number of hours needed to receive a complete interview, the average distance interviews will travel, the expected response rate, the expected interviewer attrition rate, the cost of each supply item, etc.) (Glewwe, 2002).

1.4.3 For study designs with a lot of uncertainty, advise bidders to include contingency (possibly 10%) into the budget to account for this risk (Harik, 1987).

1.4.4 For studies lasting longer than one year, suggest the inclusion of a cost-of-living increase (Glewwe, 2002).
1.4.5 For areas with rampant inflation, require frequent updates to the projected budget.

Lessons learned

1.1 Interacting with survey organizations upfront to discuss project details can help avoid possible complications during the bidding process, especially if a culture is unfamiliar with a formal bidding process. However, any information shared with one potential bidder needs to be shared with all potential bidder to prevent an appearance of impropriety or collusion in the bidding process (see Guideline 2 below).

1.2 Gathering information about constraints on survey organizations before issuing tenders will improve the bidding process. These constraints include legal requirements, cultural norms, lack of organizational capacity (e.g., does not have computer-assisted interviewing capability), and standard organizational practice (e.g., organization usually only provides interviewers two days of training but the tender requires a week), etc. This information should be used to adapt specifications in tenders to each country as appropriate.

1.3 Survey organizations may hesitate to mention any obstacles to conducting the study as outlined in the tender specifications for various reasons. Organizations should be encouraged in a culturally appropriate fashion to be open and explicit about anything that would conflict with the study specifications. Some obstacles may be quickly remedied if identified in advance. For example, it may be necessary to appoint male interviewers to some locations (such as lumber camps or mines) or to notify gatekeepers of the study and explain the need to contact given respondents. Strategies and schedules should be developed to accommodate this.

2. Ensure a fair and competitive bidding process.

Rationale

If the research capacity of a country is unknown to the central coordinating center, the bidding process is one way to illuminate this and to determine if any methodological or substantive expertise may be needed to supplement local resources. A competitive bidding process is not always possible; sometimes, there are only one or two competent survey organizations within each location being studied.

As suggested by the International Organization for Standardization (ISO), a series of standards are needed to provide a generic and standard set of processes, procedures and methods so that a fair, equitable, transparent, competitive and cost-effective system can be created (ISO, 2010).

Procedural steps

2.1 Request bids in a language understood by the reviewers from the central coordinating center, or arrange for language resources for the reviewing team to enable them to evaluate the bids.

2.2 Provide bidders with the evaluation criteria (ISO, 2011), such that they will then know what is expected at each phase of the survey lifecycle as well as what deliverables are required at each phase (Federal Committee on Statistical Methodology, 1983).

2.3 Encourage consortium bids as seems relevant because, in contexts with sparse resources, partnerships may enable survey organizations to make stronger bids if together they have a broader set of proficiencies (United Nations, 2005).

2.4 Set a timeline for the bidding process that includes time for clarification of any questions which may be raised and discussion between the contracting parties and for the bidder(s) to develop complete and comprehensive bids.

2.5 Encourage bidding organizations to identify any elements required in the tender specifications that they are unable or unwilling to meet (ESOMAR, 2001). Doing so helps avoid bids which the bidding organizations will not or cannot fulfill.

2.6 Check bids for potential problems, such as the following:

- 2.6.1 Can a proper sampling frame be obtained? See Study Design and Organizational Structure and Sample Design.
- 2.6.2 Does the bidding survey organization have access to the sample elements on the frame (e.g., will political conflicts or travel restrictions limit the areas in which the survey organization can contact individuals)? See Sample Design and Data Collection: Face-to-Face Surveys.
- 2.6.3 Is the concept of probability sampling understood and its implementation assured? See Sample Design.
- 2.6.4 Are suitable protocols and trainers available for interviewer training and interviewer motivation? See Interviewer Recruitment, Selection, and Training.

2.6.5 Are essential nonresponse bias reduction techniques realized? See Data Collection: General Considerations and Paradata and Other Auxiliary Data.
2.6.6 Are adequate quality control procedures in place? See Survey Quality.
2.6.7 Are necessary facilities, such as hardware, software, and internet access, available? See Data Collection: Face-to-Face Surveys
2.6.8 Is the specification of budget details adequate?
2.6.9 Are there local research “traditions,” such as quota sampling or undocumented substitution, that may conflict with study specifications?

2.7 Keep the bidding process transparent, open, and fair.
2.7.1 Provide the same level of help or assistance to every survey organization (Fink, 1995).
2.7.2 If new information becomes available that would be useful in preparing a bid, distribute this information to all bidders.

Lessons learned

2.1 Following up with the survey organizations to make sure they know what is expected is one way to maintain a fair bidding process. By clarifying aspects of the survey organization’s bid, the coordinating center can avoid possible complications later in the implementation of the survey. For example, in many countries the research tradition is to pay interviewers by the completed interview and not by hours worked. The coordinating center may want to explain that this practice might work well if all interviewer assignments are of the same difficulty and if the length of the interview administration is within well-defined limits. However, if assignments vary in difficulty (longer travel times, for example) or the length of the interview can vary widely (dependent upon the respondent’s answers), this will not work as well. It is important for the coordinating center to emphasize the risk of paying interviewers by the completed interview. Interviewers might be tempted to use strategies to keep interviews as short as possible in order to complete more cases. In the worst scenario, interviewers might be tempted to falsify the interview (i.e., interviewer falsification) (see Interviewer Recruitment, Selection, and Training, Ethical Considerations, and Data Collection: Face-to-Face Surveys).

3. Select the survey research organization or firm best suited to carry out the survey in each country within the constraints.

Rationale

The decision to select a survey organization or collaboration of organizations that will carry out the study, based on pre-specified and agreed-upon evaluation criteria, is a critical one. A poor choice of an organization will divert attention and resources away from other aspects of the study and may have a lasting impact on the entire endeavor.

Procedural steps

3.1 Form a bid evaluation team within the coordinating center that is comprised of a substantive expert, a statistical advisor, a methodological advisor, a financial reviewer / advisor and, as relevant, legal and local expertise.
- 3.1.1 When necessary, involve additional consultants throughout the contracting process, from preparing the tender to signing the contract (Fink, 1995).
- 3.1.2 Ensure there are no pre-existing relationships between the bid evaluation team members and the bidding survey organizations, which could violate the fairness of the process.
- 3.1.3 Determine in advance the process for final decisions on survey organization selection, in case disagreements among the review team should arise.
- 3.1.4 Have the Technical Proposals and Cost Proposals evaluated separately on their own merits.
- 3.1.5 Have each evaluation team member evaluate the survey organizations individually and make written notes.
- 3.1.6 Organize among the team a group discussion of the strengths and weaknesses of various bids.
- 3.1.7 Even if there is only one bid for a given country, conduct evaluation as described above with notes and a group discussion.
- 3.1.8 If the final required work scope and budget cannot be met by the bidding organization(s), decide whether a new round of bids is necessary or if some other alternative is available.

3.2 Use the following indicators as the basis of evaluation criteria for choosing an organization:
- 3.2.1 Local knowledge of the population of interest (United Nations, 2005).
- 3.2.2 Organizational and staff expertise in the subject area and survey methods envisioned (Fink, 1995).

- 3.2.3 Knowledge of and experience with conducting similar types of surveys (both the organization as a whole and the management/personnel assigned to the project) (ESOMAR, 2001; United Nations, 2005).
- 3.2.4 Ability to estimate the costs to complete the entire work scope.
- 3.2.5 Transparency of procedures.
- 3.2.6 Organization of field staff, including the planned supervisory structure and implementation strategy (e.g., whether interviewers are stationed throughout study areas or travel extensively in teams to different sampled locations).
- 3.2.7 Demonstrated or projected ability to meet the timeline and various specified outcomes (United Nations, 2005).
- 3.2.8 Demonstrated or projected availability of management staff and statistical support.
- 3.2.9 Affiliations with professional organizations.
- 3.2.10 Cost.
- 3.2.11 Methodological rigor and quality of the technical proposal.
- 3.2.12 Adequate proficiency of the language used by the coordinating center among, at the minimum, those key personnel in the survey organization who will be working on the project.

3.3 Find out as much about the culture as possible before negotiating strategies with survey organizations. In particular:

- 3.3.1 Make use of local or regional feedback about the survey organizations. It can be very useful to ask local contacts (these may not be directly local but at least in the region) to provide information about the organizations.
- 3.3.2 Try to become aware of any local tendencies in terms of management and likelihood of acknowledging obstacles. Encourage people to point out difficulties in terms of the knowledge of local tendencies. If you lack knowledge of what could be involved and do not have someone suitable to act as an informant, then introduce the topics you need to know about (for example, “We have sometimes found organizations fear their bid will not be considered if they admit they have trouble meeting requirements. We have learned to recognize information about local constraints as very important. Is there anything you would like to raise with us?”).
- 3.3.3 Learn to wait longer than you may be accustomed for a response and listen attentively for indirect mention of a constraint.
- 3.3.4 Try to become aware of local survey traditions or their absence. If through preparation for local negotiations it becomes clear that the study specifications run counter to local traditions, ask for information about how the organization intends to address this difference (ISO, 2006).

3.3.5 Try to become adept at recognizing and addressing hesitancy, as people or organizations may be reluctant to engage in unfamiliar procedures.

3.3.6 If something is known or found to be unusual in a given context, ask for a demonstration of its usefulness.

3.3.7 Aim to persuade those involved to try out suggested techniques or help adapt them to local conditions before deciding on their use. In other words, work with survey organizations to try out techniques before determining them as not feasible.

3.4 Negotiate work scope and costs with the most promising organization.

3.4.1 If the specifications change significantly, then reopen the bidding process to all competitive organizations (Fink, 1995).

3.4.2 Agree upon alternative designs prior to signing the contract, since change is more difficult once a study has started (Worcester et al., 2000).

3.5 Throughout this selection process, do not rely on the same person to act as both translator and negotiator with the survey organizations.

3.6 Notify unsuccessful bidders of your selection once the contract has been awarded. Supply them with your reasoning for selection, and provide feedback as to how they could be more successful in future bidding processes (ISO, 2011; Fink, 1995).

Lessons learned

3.1 When evaluating survey organizations, one of the most difficult decisions made is determining whether a survey organization is truly capable of implementing what has been promised in its bid. If two competing survey organizations propose similar technical bids, it is not always prudent to select the organization with the less expensive business / cost bid, even though not doing so might conflict with pre-determined bidding evaluation criteria. It is important to balance the proposed technical aspects, timeline, and budget with the survey organization's and staff's experience and references. Prior work is often very foretelling of future work.

3.2 When evaluating the proposed data collection timeline of each survey organization, seasonal effects must also be taken into account. One country's harvest time may be another's winter months; access to areas may be restricted or facilitated by the season. In certain times of year, large parts of the population may be on vacation or working away and difficult to reach at their usual residence (Worcester et al., 2000). See Data Collection: General Considerations for further discussion of scheduling and timelines in 3MC surveys.

4. **Execute a contract that addresses the rights and obligations of all parties involved and references local legal requirements, if applicable.**

Rationale

The final contract that the coordinating center drafts is legally binding and thus must fall under the auspices of a recognized legal authority with the power to sanction contract breaches. The sanctions should be explicit, up to and including nullifying the contract. The contract needs to be properly signed and dated by authorized representatives. Local, independent legal advice is critical to this process.

Procedural steps

4.1 Write the contract based upon the study design and specifications as described in the Tender.

4.2 Tailor contracts to the funding source, contracting organizations, and countries, as necessary. Each may carry additional requirements, such as stipulated delivery of reports, an ethics board review, and so forth.

4.3 Require official pre-approval of any subcontracting. Any known need for subcontracting in any form should be disclosed in advance by the survey organization(s) (Fink, 1995).

4.4 Incorporate bonus schemes in the contract and cost estimates as appropriate. Examples may include:
  4.4.1 Interviewer bonuses, based on performance.
  4.4.2 Organizational bonuses, such as a payment for completing interviews beyond the expected total.

4.5 Identify and specify the coordinating center's right to observe aspects of data collection (e.g., live interviews, call-backs to selected households for verification, spot checks of original questionnaires, and electronic control files) (Jowell, Roberts, Fitzgerald, & Eva, 2007).

4.6 Set reasonable production benchmarks, where possible (Fink, 1995).
  4.6.1 Define targeted response rates as one of the production benchmarks (see Data Collection: General Considerations).

- For the purpose of response rate calculation, provide the survey organizations with a defined list of the disposition codes to be used uniformly (see Data Collection: General Considerations Appendices D - G for a description of disposition codes and templates for calculating response rates).
- Go through the list of disposition codes, checking applicability of each for the local situation and define the need for additional codes to account for local conditions.

4.6.2 Require field monitoring progress reports (possibly at the individual interviewer, interviewing team, or region level) to ensure benchmarks are met. See Data Collection: General Considerations, Paradata and Other Auxiliary Data, and Survey Quality.

4.7 Establish and specify in writing ownership of the data and respondents' sample and contact information within the limits of any confidentiality restrictions.

4.8 Specify requirements for how the local survey organization will execute the data delivery and the frequency of updates on data collection progress to the coordinating center (see Data Collection: General Considerations, Data Collection: Face-to-Face Surveys, Data Harmonization, and Data Dissemination).

4.9 Specify any deliverables (such as sample specifications, instrument specifications, and source questionnaires), expected delivery dates and commitments from other parties involved, including any central organization to local organizations (e.g., advisory boards and help lines)

4.9.1 Identify and specify all required documents.

- Agree on format for these as well as who has the responsibility to develop the format.
- Include provisions for training for those required to provide documentation.
- Consider requiring copies of the consent form, translated questionnaire, training materials, and methods report (see Ethical Considerations and Interviewer Recruitment, Selection, and Training).

4.10 Specify copyrights for data and documents, including stipulations for data release (by when and by whom) and plans for data access rights (taking into account any legal restrictions and/or legal requirements).

4.11 Define the necessary security level of respondent data (e.g., contact information and survey responses) for both physical and electronic storage and transfer.

4.12 Define any restrictions on the survey organization's ability to present and publish any of the substantive or methodological results, with or without review.

4.13 For longitudinal studies, indicate, as appropriate, decisions about the protocol for possible respondent recontact. If potential for future follow-up exists, consider introducing this possibility at the time of initial contact with the respondents and ask the survey organization to budget for this activity.

4.14 Instruct the survey organization to notify the coordinating center of any potential need to change or modify the contract (Worcester et al., 2000).

Lessons learned

4.1 Although it is important to enforce adherence to specifications, a measure of flexibility is also needed. Natural disasters, unexpected political events, and outbreaks of disease can interrupt data collection and make agreed-upon deadlines impossible.

4.2 Approving the use of subcontractors may impact the coordinating center’s level of control. For example, monitoring data collection will be problematic if subcontractors restrict the right of the coordinating center to observe aspects of the survey process. Certain study specifications, such as the required security level of respondent data, can be difficult to ensure while working with subcontractors.

5. Define upfront the quality standards that lay the quality framework for the rest of the survey lifecycle.

Rationale

The bidding process may be the first interaction the survey organizations have with the coordinating center. Hence, it is essential for the coordinating center, from the conception of the survey, to demonstrate and emphasize the importance of quality.

Procedural steps

5.1 Develop a quality management plan (see Survey Quality). Use this plan as the outline for expectations of the survey organizations throughout the entire study.

5.2 Ask bidding survey organizations to detail their quality control and quality assurance procedures, and include minimum quality requirements in the criteria used for evaluating the bidders.

5.3 Consider re-releasing the tender if no bidding survey organization can meet the requested quality standards.

5.4 Define progress approval points throughout the research process (e.g., sample selection, questionnaire design, interviewer training, and data collection milestones) to ensure each party involved achieves the study's objectives.
   5.4.1 Require certification from the coordinating center at these formal points before a survey organization can proceed with the study.
   5.4.2 Sanctions for unnecessary delays or specification deviations should be specified, in the contract, before the study begins.

Lessons learned

5.1 Since budgets are often underestimated, it is critical to monitor the overall budget throughout the survey lifecycle to avoid a potential overrun at the end of the study. In addition, individually monitoring the budget of each survey task is an important quality assurance procedure. If the budget for each survey task is more detailed (i.e., specified budgets for each direct cost component), it is useful to systematically assess the status of the budget and weigh the quality trade-off by monitoring costs at the lower levels (see Survey Quality).

6. Document the steps taken while preparing tenders, soliciting bids, and drawing up and executing contracts.

Rationale

The coordinating center can use the contract resulting from the bidding process to enforce its expectations of the survey organizations. Thus, it is very important that steps taken throughout the process be clearly noted and transparent to those involved. No one involved should be surprised at how the study is to be structured, what production actions are required, and when the final deliverables are to be completed.

Procedural steps

6.1 Clearly state specifications in tenders.

6.2 In advance of releasing tenders, document the evaluation criteria to be used when assessing bids.

6.3 Keep a record of the information exchanged with each survey organization to make sure no one organization receives differential treatment during the bidding process.

6.4 Document bid evaluation team scores for each survey organization's bid.

6.5 Collect notes from each member of the bid evaluation team as to how they arrived at their selection decision.

6.6 Make sure each survey organization formally details all aspects of their anticipated scope of work in their bid. Information beyond what is written in the bid (e.g., from other forms of correspondence) should not be considered when evaluating the survey organization—so as not to give differential treatment.

6.7 Keep records of all notifications to unsuccessful bidders of your selection.

6.8 Write contracts that are tailored to the involved parties (e.g., funding source, coordinating center, survey organization, etc.). When writing the contract, include all specifications of the scope of work, budget, and timeline for which each survey organization should commit.

6.9 In the contract, establish responsibility for documenting all aspects of the study.

6.10 Request documentation of any subcontracts from the survey organizations.

6.11 Have a signed agreement regarding the ownership of the data and respondent information, within the limits of confidentiality regulations (see Ethical Considerations).

6.12 Keep a copy of the tenders, all bid materials provided by any survey organization submitting a bid, and a copy of the contracts as well as any modifications.

6.13 Documents must be subject to a version control process to identify changes (ISO, 2012).

Appendix A

Timeline of effort by survey task example

When bidding for a 3MC survey, it is important for a survey organization to outline how it plans to meet the specified deliverables' deadlines. This can be achieved by creating a timeline that demonstrates when the survey organization will work on each task of the survey lifecycle and how much effort (i.e., how many hours) is necessary to perform that task.

Below is an example of a timeline with an expected 24-month duration (specified in actual calendar months and years) and survey tasks corresponding with each of the Cross-Cultural Survey Guidelines. The 'X's are placeholders for the number of hours assigned to each survey task per month (for the entire staff's effort). It is critical that the total number of hours for all tasks for all months equal the total number of hours for all assigned staff (see Appendix B).

TASK	**MONTH OF SURVEY**																								
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	Total
Study, Organizational and Operational																									
Survey Quality																									
Ethical Considerations																									
Tenders, Bids and Contracts																									
Sample Design																									
Questionnaire Design																									
Adaptation																									
Translation																									
Instrument Technical Design																									
Interviewer Recruitment, Selection and Training																									
Pretesting																									
Paradata and Other Auxiliary Data																									
Data Collection																									
Data Harmonization																									
Data Processing and Statistical Adjustment																									
Data Dissemination																									
Total																									

Appendix B

Person loading chart example

When bidding for a 3MC survey, it is important for each survey organization to outline how it plans to assign responsibilities for each task of the survey lifecycle to which staff members and how much effort (i.e., how many hours, days or percent of effort) is necessary for each staff member to accomplish each given task. This can be achieved by creating a person loading chart.

Below is an example of a person loading chart with example study roles and survey tasks corresponding with each of the Cross-Cultural Survey Guidelines. If the name of the staff member fulfilling the role is known, include the name. If the name of the staff member is not known, include the job title. Indicate if multiple people will be necessary for a given role. The 'X's are placeholders for the number of hours budgeted to staff for each task. It is critical that the total number of hours for all staff, for all tasks, equal the total number of hours for all months of the survey (see Appendix A.)

	SURVEY TASKS															
STAFF	Study, Organizational, and Operational Structure	Survey Quality	Ethical Considerations in Surveys	Tenders, Bids, and Contracts	Sample Design	Questionnaire Design	Adaptation	Translation	Instrument Technical Design	Pretesting	Interviewer Recruitment, Selection, and Training	Data Collection	Data Harmonization	Data Processing and Statistical Adjustment	Data Dissemination	Total
Project Manager	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	
Quality Coordinator	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	
Budget Analyst	X	X		X												
Office Assistant	X	X	X	X						X	X	X				
Statistician		X			X					X		X	X	X	X	
Questionnaire Designer		X				X	X	X	X	X						
Translator		X				X	X	X	X							
Data Manager		X			X				X	X		X	X	X	X	
Information Technologist		X							X	X		X	X	X	X	
Programmer		X							X	X		X		X		
Field Manager		X							X	X	X	X		X		
Field Support Staff		X							X	X		X				
Data Processing Manager		X				X								X		
Interviewer Recruiter		X									X					
Interviewer Supervisor		X								X	X	X				
Interviewer		X								X	X	X				
Interviewer Aide/Driver		X								X		X				

Appendix C

Salary budget template example

Specifying the salary costs for each survey task is an important component of a bid. For each staff member, estimate the number of hours, days or percent of effort that he or she will contribute. In this example, the staff members expected to work on the pretesting task are listed by job title, with only one person needed for each role.

If the name of the staff member completing the role is known, include the individual's name and actual hourly rate. If the name of the staff member is not known, include the job title and average hourly rate for the staff members with that title. If several people have the same job title, include separate entries for each (however, "interviewers" may be listed as a single line). When a survey task is completed across multiple years, the budget estimate should account for the expected changes in salary rates.

STAFF	HOURS	HOURLY RATE	TOTAL COST
Project Manager			
Quality Coordinator			
Office Assistant			
Statistician			
Questionnaire Designer			
Data Manager			
Information Technologist			
Programmer			
Field Manager			
Field Support Staff			
Interviewer Supervisor/Trainer			
Interviewers*			
Interviewer Aide/Driver			
TOTAL HOURS:		TOTAL COST:	

* Note that in some countries, interviewers may be paid by the number of interviews completed.

Appendix D

Non-salary budget template example

Specifying the non-salary costs for each survey task is an important component of a bid. For each survey task, estimate the number of items and cost per unit. In this example, the items expected to be used for the pretesting task are listed. When a task is completed across multiple years, the budget estimate should account for the increases in per unit costs.

ITEMS	NUMBER	COST PER UNIT	TOTAL COST
Facilities			
Meals			
Laptop Computer			
Software Licensing			
General Supplies			
Communications			
Postage			
Printing			
Respondent Incentive Payments			
Respondent Recruitment Expenses			
Travel			
Other			
		TOTAL COST:	

References

American Association for Public Opinion Research (AAPOR). (2015). *Standard definitions: Final dispositions of case codes and outcome rates for surveys* (8th ed.). Retrieved March 2, 2016, from https://www.esomar.org/uploads/public/knowledge-and-standards/codes-and-guidelines/AAPOR_Standard-Definitions2015_8theditionwithchanges_April2015_logo.pdf

The European Society for Opinion and Market Research (ESOMAR). (2001). *ESOMAR world research codes & guidelines: How to commission research* [brochure]. Amsterdam: The Netherlands.

Federal Committee on Statistical Methodology. (1983). *Statistical policy working paper 9: Contracting for surveys*. Washington, DC: Office of Management and Budget.

Fink, A. (Ed.). (1995). *The survey kit: The survey handbook* (Vol. 1). Thousand Oaks, CA: Sage Publications.

Glewwe, P. (2002). Overview of the implementation of household surveys in developing countries. In United Nations Statistical Division (Eds.), *Household sample surveys in developing and transition countries* (pp. 54-66). New York, NY: United Nations.

Harik, I. (1987). Some political and cultural considerations bearing on survey research in the Arab world. In M. A. Tessler, M. Palmer, T. E. Farah, & B. L. Ibrahim (Eds.), *The evaluation and application of survey research in the Arab world* (pp. 66-74). Boulder, CO: Westview Press.

International Organization for Standardization (ISO). (2006). *Market, opinion and social research—terms, definitions and service requirements* (ISO 20252). Geneva. Retrieved April 30, 2010, from https://www.astandis.at/shopV5/Preview.action%3Bjsessionid=0D8F503B3BD48375403D96257760FABB?preview=&dokkey=342276&selectedLocale=en/

International Organization for Standardization (ISO). (2010). *Construction procurement — Part 1: Processes, methods and procedures* (ISO 10845). Retrieved March 18, 2016, from https://www.iso.org/obp/ui/#iso:std:iso:10845:-1:ed-1:v1:en

International Organization for Standardization (ISO). (2011). *Part 3: Standard conditions of tender* (ISO 10845). Retrieved March 18, 2016, from https://www.iso.org/obp/ui/#iso:std:iso:10845:-3:ed-1:v1:en

International Organization for Standardization (ISO). (2012). *Market, opinion and social research — Vocabulary and service requirements* (ISO 20252). Retrieved March 18, 2016, from https://www.iso.org/obp/ui/#iso:std:iso:20252:ed-2:v1:en

Jowell, R., Roberts, C., Fitzgerald, R., & Eva, G. (Eds.). (2007). *Measuring attitudes cross-nationally: Lessons from the European Social Survey*. London: Sage Publications.

Project Management Institute. (2004). *A guide to the project management body of knowledge (PMBOK® guide)* (3rd ed.). Newtown Square, PA: Project Management Institute, Inc.

United Nations. (2005). *Household surveys in developing and transition countries.* New York, NY: United Nations, Department of Economic and Social Affairs. Retrieved March 26, 2010, from http://unstats.un.org/unsd/HHsurveys/pdf/Household_surveys.pdf

Worcester, R., Lagos, M., & Basañez, M. (2000). *Cross-national studies: Lessons learned the hard way*. Paper presented at the 2000 AAPOR/WAPOR Conference, Portland, OR.

Sample Design

Frost Hubbard, Yu-chieh (Jay) Lin, Dan Zahs, and Mengyao Hu, 2016

Introduction

Although one could employ a census to measure the entire population, it is more common to take a sample of the population. A properly designed probability sample can be used to make estimates for not only the sample itself, but also for the underlying population from which it was selected. A probability sample is one in which each element of the (underlying) population has a known and non-zero chance of being selected. That is, every person has a chance to be included in the study and have his or her characteristics, opinions, etc., become part of the data. It should be noted that everyone does not have to have an equal chance of being selected – just a known non-zero chance of being selected.

Probability samples have several desirable characteristics. They enable us to put a margin of error or confidence interval on our estimates – essentially a measure of how accurate the estimate is compared to the same estimate calculated on the full population. Probability samples make it possible not only to compare the sample to the population, but also to compare a sample from one population to a sample from another population, as occurs in multinational, multicultural, or multiregional surveys, which we refer to as "3MC" surveys.

An optimal sample design is one that maximizes the amount of information obtained per monetary unit spent within the allotted time and meets the specified level of precision (Heeringa & O'Muircheartaigh, 2010). One important prerequisite for 3MC surveys is that all samples are full probability samples from comparable target populations (Kish, 1994).

Different nations have different sampling resources and conditions. For a multinational survey, this means that the optimal sample design for one country may not be the optimal design for another. (Please note this chapter uses the term "participating country" to encompass any participating country, culture, region or organization in a 3MC study.) Therefore, allowing each participating country some flexibility in its choice of sample design is highly recommended, so long as all sample designs use probability methods at each stage of selection (Häder & Gabler, 2003; Kish, 1994).

This chapter outlines the decisions that need to be made when designing a 3MC probability survey sample. It encourages survey organizers to allow sample designs to differ among participating countries while, at the same time, ensuring standardization on the principles of probability sampling.

Although flexibility will usually be necessary in sample design and implementation, the guidelines and recommendations discussed in this chapter are only directly applicable if a probability framework is followed. Probability designs can be expensive. To reduce costs, some survey organizations select nonprobability samples, including convenience samples and quota samples. While not all nonprobability samples are biased, the risk of bias is extremely high and, most importantly, cannot be measured—a survey that uses nonprobability sampling cannot estimate the error in the sample estimates (Heeringa & O'Muircheartaigh, 2010).

Please note that this chapter assumes that the reader has a basic understanding of statistics and terms such as "variance" and "standard deviation." Please refer to Further Reading or an introductory statistics textbook if a statistics refresher is needed.

Guidelines

Goal: To select an optimal, cost-efficient probability sample in each participating country that is representative of the target population and allows researchers to make inferences to the target population, and to standardize sample designs without hampering optimal designs in each participating country.

1. **Define the target population for the study across all countries and the survey population within each participating country.**

   Rationale

   The survey planners of any 3MC survey need to develop a detailed, concise definition of the target population in order to ensure that each participating country collects data from the same population. Without a precise definition, one country may collect data that include a certain subgroup, such as noncitizens, while another country excludes this subgroup. This difference in sample composition may influence the estimates of key statistics across countries. In addition, a precise definition will let future users of the survey data know to which exact population the survey data refer. The data users can then make a more informed decision about whether to include the survey data in their analyses.

   Procedural steps

   1.1 Define the target population across all participating countries as clearly as possible, including the elements of the population and the time frame of the group (Groves et al., 2009). For example, a target population might be defined as, "All persons above the age of eighteen, who usually slept most nights in housing units in South Africa, Zimbabwe, Lesotho, and Swaziland during April, 2007." Note that this definition would, in turn, require definitions of the terms "usually," "most," and "housing unit."

1.2 To ensure a clear description of the target population, think about all potential inclusion/exclusion criteria. For example, the target population might exclude:
   1.2.1 Persons outside a defined age range.
   1.2.2 Persons in institutions, such as hospitals, nursing homes, prisons, group quarters, colleges, monasteries, or military bases.
   1.2.3 Persons living in certain sparsely populated or remote geographic regions.
   1.2.4 Non-citizens, ethnic minorities, homeless or nomadic populations, language groups.

1.3 Define the survey population within each participating country by refining the target population based on cost, security, or access restrictions to all target population elements (Groves et al., 2009).
   1.3.1 In some cases the target and final survey populations are the same. However, there can be differences. The target population can be thought of as “Who we want to include in our survey” while the survey population is “Who we can include in our survey”. Differences can occur due to frame availability, screening ability, cost, etc. For example, the survey population may exclude those residing in war-torn areas, or the data collection period may be narrowed in areas with civil disturbances that are threatening to escalate.
   1.3.2 Ideally, the resulting survey populations are comparable across all countries. To help analysts determine comparability, the decisions involving the definition and implementation of the target and survey populations should be well documented.

Lessons learned

1.1 Large established 3MC surveys have defined their target and survey populations differently, depending upon the goals and topics of the study.
   1.1.1 The Afrobarometer is an independent, nonpartisan research project that measures the social, political, and economic atmosphere in Africa. Afrobarometer surveys are conducted in more than a dozen African countries and are repeated on a regular cycle. Participants in Round 6 of the Afrobarometer Survey had to be citizens of their country and of voting age on the day of the survey. They had to complete the interview in their country’s national language or in an official local language for which a translation was available. People living in areas of armed conflict or natural disasters, national parks and game reserves, and in institutionalized settings were

excluded. Special cases, like areas of political unrest, were reviewed on a case-by-case basis (AfroBarometer, 2014).

1.1.2 The Asian Barometer is an applied research program studying public opinion on political values, democracy, and governance in thirteen East Asian political systems (Japan, Mongolia, South Koreas, Taiwan, Hong Kong, China, the Philippines, Thailand, Vietnam, Cambodia, Singapore, Indonesia, and Malaysia) and five South Asian countries (India, Pakistan, Bangladesh, Sri Lanka, and Nepal). The target population of the Asian Barometer was defined as citizens who were at least 20 years of age and were eligible to vote (i.e., were not disenfranchised due to mental illness or incarceration) (see http://www.asianbarometer.org/survey/survey-methods).

1.1.3 The European Social Survey (ESS) is an academically-driven social survey designed to chart and explain the interaction between Europe's changing institutions and the attitudes, beliefs and behavior patterns of its diverse populations. Round 7 of the ESS covers more than 22 nations and includes persons 15 years or older who are resident within private households, regardless of nationality, citizenship, or language; homeless and institutional populations are excluded from the sample (ESS, 2014).

1.1.4 The Living Standards Measurement Study (LSMS) was established by the World Bank in 1980 to explore ways of improving the type and quality of household data collected by statistical offices in developing countries. Its goal is to foster increased use of household data as a basis for policy decision making. Respondent requirements and exclusions vary across participating countries (LSMS, 1996).

1.1.5 The Survey of Health, Ageing and Retirement in Europe (SHARE) studies the health, socio-economic status and social and family networks of individuals, aged 50 or over, in nearly 20 European countries and Israel. In addition to the age requirement, respondents had to be residents of the country and speak (one of) the official language(s) of the country. The study excludes seasonal or vacationing residents, persons physically or mentally unable to participate, those who died before the start of the field period, or who are unable to speak the specific language of the national questionnaire. It also excludes residents of institutions, except facilities for the elderly (De Luca, Rossetti, & Malter, 2015).

1.1.6 The World Value Survey is conducted by a non-profit association seated in Stockholm, Sweden, to help social scientists and policy makers better understand worldviews and changes that are taking place in the beliefs, values, and motivations of people throughout the world. Respondents are adults, 18 years and older; some countries also place upper limits on age (http://www.worldvaluessurvey.com/).

1.1.7 The World Mental Health (WMH) Survey studies mental illness in selected countries in Europe, Asia, Africa, and North and South America. One of the major goals of the WMH Study was to compare the age of onset of disease across countries. Best practice might suggest strictly defining the age of majority (e.g., 18 years old). However, the WMH study organizers recognized that strictly defining this inclusion criterion would be difficult, given that age of majority varies by country (and even within a country). Also, a strict definition would affect study protocols such as ethics reviews and informed consent (seeking permission to interview minors). Therefore, the WMH Study had to make a difficult decision to allow the age eligibility criterion to vary across countries. In the end, the WMH Study allowed the age range to vary, with 16 years of age being the youngest lower age limit; some countries also set upper age limits. This was taken into consideration in the analysis stage (http://www.hcp.med.harvard.edu/wmh/). Participating countries were also allowed to vary in whether or not respondents must be citizens or be fluent in specific languages (Kessler, Ustun, & World Health Organization, 2008).

1.2 An increasingly common form of housing seen in international studies is workers' quarters. Survey designers may want to explicitly state in the definition of the target population whether workers' quarters should be included or excluded.

2. Identify and evaluate potential sampling frames. Select or create the sampling frame that best covers the target population given the country's survey budget.

Rationale

An ideal sampling frame contains all of the elements of the target population. However, very few sampling frames exist that allow access to every element in the target population. The goal, then, is to choose a sampling frame or a set of sampling frames that allows access to the largest number of elements in the target population and contains the fewest number of ineligible elements, given the constraints of the survey budget.

Procedural steps

Note: Although this chapter focuses heavily on the method and practice of in-person interviews, telephone interviews using Random-Digit-Dialing (RDD) frames (Tucker, Lepkowski, & Piekarski, 2002) or other lists are also widely used. In cross-national surveys, one country may conduct interviews over the telephone while another country conducts face-to-face interviews (see, for example, the Gallup World Poll: http://www.gallup.com/178667/gallup-world-poll-work.aspx). This difference in the mode of data collection, driven by the available sampling frames, might lead to differences in the results. (See Data Collection: General Considerations for more information about mixed-mode design and mode effects.)

2.1 Have each participating country identify a pre-existing list (or lists) of desired elements or clusters of elements of the target population to create a sampling frame. Examples include:
   - 2.1.1 Official population registries.
   - 2.1.2 Health registries.
   - 2.1.3 Lists of schools.
   - 2.1.4 Postal registries.
   - 2.1.5 Electoral rolls.
   - 2.1.6 Utility customer lists.
   - 2.1.7 Pre-existing sampling frames used by other surveys.
   - 2.1.8 Telephone directories.
   - 2.1.9 Random-Digit-Dialing (RDD) telephone frames.
     - The phone system in many nations makes it possible to generate telephone numbers for sampling purposes without first generating all possible telephone number combinations.
     - In some places there might be separate lists or frames for landline telephone numbers and cellular telephone numbers.
   - 2.1.10 Lists of email addresses.
     - These are only appropriate in special circumstances.
     - In some populations, many people may not have an email address; at the same time, some people may have multiple email addresses.
   - 2.1.11 Other list(s) of addresses, phone numbers or names.

2.2 Create a sampling frame via area probability sampling methods if there are no appropriate pre-existing lists of elements (or if the lists are not made available for use by the survey project) of the target population. Even if such lists do exist, it is wise to assess the cost and coverage errors associated with creating an area probability sampling frame and how this might compare to the use of a pre-existing list. Many texts and documents provide detailed guidance regarding the development of area probability samples (Kish, 1965; Üstun, Chatterji, Mechbal, & Murray, 2005). Below, we outline a simple two-stage area probability sample of households, including the following steps used in many 3MC surveys. (Additional information can be found in Appendices A and C).

2.2.1 Create a list of primary sampling units (PSUs) based on geographic clusters. In the United States, for example, these clusters are typically census enumeration areas.

2.2.2 Using a probability sampling method, select a sample of PSUs.

2.2.3 Determine the appropriate method for listing the housing units (secondary sampling units (SSUs)) within selected PSUs.

2.2.4 Send staff to list the housing units in selected PSUs, maintaining a uniform definition of what constitutes a "housing unit."

2.2.5 Once the housing units in a PSU have been enumerated, select a random sample of housing units from the list.

2.2.6 During data collection, ask the selected housing units within the PSUs to participate. Once the housing unit has agreed to participate, complete a list of all eligible members within the housing unit. (See Appendix B for more detailed instructions on enumerating eligible members of the housing unit.)

2.2.7 Using a probability method, select one or more eligible members within the housing unit (Koch, 2017).

- Train the interviewer or, where possible, program the computer, to select an eligible respondent based on the selection method specified.
- While some "quasi-probability" and "non-probability" or "quota" within-household selection methods can be used, be aware that such procedures produce a non-probability sample.
- Some studies may want to survey the most knowledgeable adult, the one with primary child care responsibilities, or with some other specific characteristics, rather than randomly select from among the household members. Note that this would be part of the definition of the target population and, thus, does not violate probability sampling.

2.3 Evaluate how well each potential sampling frame covers the target population (Groves, 1989).

2.3.1 Examine the sampling frame(s) for a one-to-one mapping between the elements on the sampling frame and the target population. There are four potential problems:

- Undercoverage (missing elements): elements in the target population do not appear on sampling frame.

- Ineligible elements: elements on the sampling frame do not exist in the target population.
- Duplication: several sampling frame elements match one target population element.
- Other mismatches: for example, one sampling frame element matches many target population elements (i.e., only a street address is listed but there might be several apartment units at one address).

2.3.2 Area frames generally have better coverage properties than pre-existing lists of addresses, names, or telephone numbers because area frames have fewer missing eligible elements, fewer duplications, and fewer ineligible elements. For more information on the creation of area probability frames, see Appendices A and C.

2.4 Consider combining multiple sampling frames which cover the same population if the union of the different frames would cover the target population better than any one of the frames on its own (Hartley, 1962). When combining multiple lists to create a sampling frame, the following steps should be considered (Groves & Lepkowski, 1985):

2.4.1 First, determine for each element on the combined frame whether it is a member of Frame A only, Frame B only, or both Frame A and B, and calculate probabilities (Lohr & Rao, 2006).

- If the membership of each element can be determined before sampling, duplicates can be removed from the sampling frame.
- A variation on this is to use a rule that can be applied to just the sample, rather than to the entire frame. Frame A might be designated as the controlling frame, in the sense that a unit that is in both frames is allowed to be sampled only from A. After the sample is selected, determine whether each unit from B is on the A frame, and retain the unit only if it is not on frame A. This method extends to more than two frames by assigning a priority order to the frames.
- If the membership cannot be determined prior to sampling, then elements belonging to both frames can be weighted for unequal probabilities of selection after data collection (see Data Processing and Statistical Adjustment for best practices for weighting and nonresponse adjustments).

2.5 Assess the cost of obtaining or creating each potential sampling frame.

2.5.1 In most circumstances, it is less expensive to purchase pre-existing lists than to create area probability frames.

2.5.2 While three stage area probability samples are more costly to develop than pre-existing lists, they facilitate cost-effective clustering for interviews.

2.5.3 If pre-existing lists are not up-to-date, potential respondents may no longer live at the address on the list or may have changed phone numbers; tracking these individuals can be very expensive.

2.5.4 Pre-existing lists for household surveys often contain more ineligible elements than area probability frames, increasing survey costs.

2.6 Update an already existing frame, if necessary. For example, World Health Survey (WHS) administrators have suggested that frames that are two years old or more require updating (Üstun et al., 2005). However, that is only a rough rule of thumb. In mobile societies, a frame that is one year old may need to be updated, while in other societies, even older frames might still be accurate.

2.6.1 If the frame is a pre-existing list, contact the provider of the list for the newest version and its quality documentation.

2.6.2 If the frame is an area probability sample and the target population has undergone extensive movement or substantial housing growth since the creation of the frame, then updating the PSUs and SSUs will be required. However, what is most important is the quality of the enumerative listing.

2.6.3 Select the sampling frame based on the undercoverage error vs. cost tradeoff.

Lessons learned

2.1 Most countries do not have complete registers of the resident population and, therefore, construct area frames for sample selection.

2.1.1 Some surveys in countries with limited resources have found that it can be difficult to enumerate the rural, impoverished areas (Bergsten, 1980; Dunzhu et al., 2003; Kalton, 1983) and, consequently, surveys in these countries may under-represent poorer or more rural residents. However, not all survey methodologists agree with the opinions expressed by these authors regarding enumeration in rural, poor areas. Those who disagree argue that the poor enumerations are mainly due to low expectations and insufficient training and supervision. Regardless, if the statistic of interest is correlated with income and/or urbanicity, the sample estimate will be biased. For example, the Tibet Eye Care Assessment, a study on blindness and eye diseases in the Tibet Autonomous Region of China, used an area sampling frame (Dunzhu et al., 2003). One of the PSUs was the township of Nakchu, an area of high elevation that is primarily populated by nomadic herders. Because of the elevation and rough terrain, Nakchu proved difficult to enumerate accurately. As a result, the survey sample underrepresented the residents of the roughest terrain of Nakchu. This was potentially important, as ophthalmologists believe that Tibetans who live in the most inaccessible regions and the highest elevation have the highest prevalence of eye disease and visual impairment.

2.1.2 Even when available, the quality and recency of census data and/or administrative lists can vary across countries. Data from other sources might be available. For example IPUMS-International is an "effort to inventory, preserve, harmonize, and disseminate census microdata from around the world. The project has collected the world's largest archive of publicly available census samples. The data are coded and documented consistently across countries and over time to facilitate comparative research. IPUMS-International makes these data available to qualified researchers free of charge through a web dissemination system" (Minnesota Population Center, 2015**).**

2.2 Local residents can help produce maps for an area probability sample. When measuring the size of the rural population in Malawi, researchers used statistical methods to determine the sample size and selection of villages. Then they asked members of the selected communities to help draw maps, including exact village boundaries, key landmarks, and each individual household (Barahona & Levy, 2006).

2.3 Technology can also permit the associated use of GPS to facilitate sample selection. Vanden-Eng et al., (2007) used PDAs equipped with GPS units in household surveys to rapidly map all households in selected areas, choose a random sample, and navigate back to the sampled households to conduct an interview in Togo and Niger. Incorporating GPS into the sampling method allows researchers to select a random sample of households for interviewing from a complete and up-to-date listing of the households and, as a result, the probability of selection is known. The GPS data collected also provides geospatial information for reports and analyses. Vanden-Eng et al., (2007) were able to generate a preliminary report of survey findings from the aggregated data (including maps of the districts sampled showing all of the households and their sample inclusion status) within a few days of completion of data collection. This can be particularly useful in situations where data are needed quickly to guide public health action, such as in routine monitoring and evaluation and rapid needs assessments.

2.4 Eckman, Himelein, and Dever (2017) discuss several sampling approaches using geographic information system (GIS) tools for face-to-face household surveys. Examples include the use of Google Maps, Google Earth, satellite photos, handheld global positioning system (GPS) devices and location-enabled applications on mobile phones. These technologies can be used to form and select clusters in early stages of selection and to select households in a later stage of selection.

2.5 Koch (2017) provides an overview on within-household selection methods for face-to-face household surveys. Various methods exist, and the two commonly used ones are Kish (see Appendix B) and birthday methods. The Kish method is believed to be the gold standard for within-household selection. The technique requires all eligible persons in a household to be listed. A person is sampled with equal probability from all eligible persons. Birthday methods use household members' birthday information to select the respondent. Both within-household selection methods are used in European Social Survey (ESS) countries.

3. Choose a selection procedure that will randomly select elements from the sampling frame and ensure that important subgroups in the population will be represented.

Rationale

Sample selection is a crucial part of the survey lifecycle. Since we cannot survey every possible element from the target population, we must rely on probability theory to make inferences from the sample back to the target population.

Procedural steps

3.1 Consider only selection methods that will provide a probability sample.

3.1.1 Statisticians have developed procedures for estimating sampling errors in probability samples which apply to any type of population.

3.1.2 Random sample selection protects the researcher against accusations that his or her bias, whether conscious or unconscious, affected the selection.

3.2 Identify the optimal sampling method available in each country. Below are summaries of each selection method. See Appendix C for additional information about each selection method.

3.3 Consider Simple Random Sampling (SRS) without replacement. In SRS, each element on the frame has an equal probability of selection, and each combination of *n* elements has the same probability of being selected. Due to the benefits of stratification, this technique is seldom used in practice.

3.3.1 Advantages of SRS:
- The procedure is easy to understand and implement.

3.3.2 Disadvantages of SRS:
- The costs in attempting to interview a simple random sample of persons can be quite high.
- SRS provides no assurance that important subpopulations will be included in the sample.

3.4 Consider Stratified Sampling (see Appendix C for a detailed description). Stratified sampling uses auxiliary information on the sampling frame to ensure that specified subgroups are represented in the sample and to improve survey precision. Virtually all sample implementations use some form of stratification.

3.4.1 Some examples of commonly used stratification variables are:
- Age.
- Region of the country.
- State/province.
- County.
- City/town, community, municipality.
- Postal code.
- Metropolitan status/urbanicity.
- Size of sampling unit (e.g., population of city).
- Race/ethnicity.
- National origin.

3.4.2 Advantages of stratified sampling:
- Stratified samples can be ether proportionate (which is general used to improve the precision of estimates of the total population) or disproportionate (to help improve sub-group estimates)
- Depending on the allocation of elements to the strata, the method can produce gains in precision (i.e., decrease in sampling variance) for the same efforts by making certain that essential subpopulations are included in the sample.

3.4.3 Disadvantages of stratified sampling:
- For any given frame stratification variables may be limited.

- No gains in precision will be seen if the stratification variables are not correlated with the statistic(s) of interest. In some cases, the precision may even decrease.

3.5 Consider Systematic Sampling to reduce the operational effort needed to select the sample. In systematic sampling, every *k*th element on the sampling frame is selected after a random start.

3.5.1 Advantages of systematic sampling:

- The operational time necessary to select the sample can be reduced substantially.
- If the sampling frame is sorted into groups or ordered in some other way prior to selection, the systematic sampling method will select a proportionately allocated sample (see description below of stratified sampling). This is often referred to as "implicitly stratified sampling."

3.5.2 Disadvantages of systematic sampling:

- If the key selection variables on the sampling frame are sorted in a periodic pattern (e.g., 2, 4, 6, 2, 4, 6…) and the selection interval coincides with periodic pattern, the systematic sampling method will not perform well (Kalton, 1983). If periodicity is a problem, several systematic samples can be selected and concatenated to form the total survey sample.
- If the list is sorted in a specific order before selection, the repeated sampling variance of estimates cannot be computed exactly.

3.6 Consider Cluster Sampling (see Appendix C for a detailed description). With cluster sampling, clusters of frame elements are selected, rather than selecting individual elements one at a time. Within the selected clusters, we can interview all or a sample of households. We can even do multi-stage cluster sampling where we select an additional sample of clusters inside the larger selected clusters.

3.6.1 Advantages of cluster sampling:

- When survey populations are spread over a wide geographic area and interviews are to be done face-to-face, it can be very costly to create an element frame and visit *n* elements randomly selected over the entire area.
- A full frame of all elements in the entire population is not required only the elements within selected clusters are needed. This will reduce listing costs.

3.6.2 Disadvantages of cluster sampling:

- Estimates are not as precise as with SRS, necessitating a larger sample size in order to get the same level of precision. See Appendix D for more information about effective sample size.

3.7 Consider Two-Phase (or Double) Sampling (see Appendix C for a further description). The concept of two-phase sampling is to sample elements, measure one or more variables on these 1st-phase elements, and use that information to select a 2nd-phase subsample.

3.7.1 A common application is to collect 1st-phase data that is used to stratify elements for the 2nd-phase subsample.

3.7.2 Survey samplers use two-phase sampling to help reduce nonresponse, with the stratifying variable from phase one being whether the person responded to the initial survey request. For example, samplers might select a subsample of nonrespondents and try to entice the nonrespondents to participate by offering incentives.

3.8 Consider Replicated (or interpenetrated) Sampling. Replicated sampling is a method in which "the total sample is made up of a set of replicate subsamples, each of the identical sample design (Kalton, 1983)."

3.8.1 Advantages of replicated sampling:

- It allows the study of variable nonsampling errors, such as interviewer variance.
- It allows for simple and general sampling variance estimation (see Data Processing and Statistical Adjustment for further explanation, especially regarding the methods Balanced Repeated Replication and Jackknife Repeated Replication).

3.8.2 Disadvantages of replicated sampling:

- There is a loss in the precision of sampling variance estimators; a small number of replicates leads to a decrease in the number of degrees of freedom when calculating confidence intervals.

3.9 Consider using a combination of techniques such as a stratified multistage cluster design.

3.9.1 Most surveys in countries with limited resources are based on stratified multistage cluster designs (Yansaneh, 2005). The combination of these techniques reduces data collection costs by clustering while striving to increase or maintain precision through stratification.

Lessons learned

3.1 Probability sampling at every stage generally requires more labor and funding than other methods. Therefore, some 3MC studies have used probability sampling in the first stage of selection and then allowed quota sampling or substitution to occur at later stages (Chikwanha, 2005; Heeringa & O'Muircheartaigh, 2010). However, a survey that uses a nonprobability sampling method at any stage of selection cannot estimate the sampling error of the estimates (Heeringa & O'Muircheartaigh, 2010). Therefore, the coordinating center should make every effort to promote the use of a full probability sample and remove any obstacles that would prevent participating countries from using probability methods at each stage of selection. For the first few waves of data collection, the International Social Survey Programme (ISSP) allowed countries to use nonprobability methods at the household level. After recognizing the problem this caused in variance estimation, the ISSP has required countries to use full probability samples since 2000 (Häder & Gabler, 2003).

3.2 Existing 3MC surveys have employed various strategies for selecting a probability sample.

3.2.1 Round 6 of the Afrobarometer Survey uses a clustered, stratified, multi-stage, area probability sample. The sampling design has four stages in urban areas: (1) stratify and randomly select primary sampling units, (2) randomly select sampling start-points, (3) randomly choose households, and (4) randomly select individual respondents within households ; and five stages in rural areas: (1) randomly select secondary sampling units (SSU), (2) randomly select two primary sampling units (PSU) from each SSU, (3) randomly select sampling start-points from each PSU, (4) randomly choose households, and (5) randomly select individual respondents within households (AfroBarometer, 2014).

3.2.2 Sample designs for Round 5 of the European Social Survey (ESS) must use random probability sampling at every stage. Samples are designed by a sampling expert or panel and may include clustering and stratification. Quota sampling and substitutions are not allowed although subgroups may be over-sampled. Sample designs and frames must be documented in full and be pre-approved by a sampling expert or panel. The target minimum response rate is 70% (ESS, 2014).

3.2.3 Sampling frames and designs for the Living Standard Measurement Study Survey (LSMS) vary across participating countries but generally consist of two stages. In the first stage, the sample frame is developed from census files and Primary Sampling Units are randomly selected with probability proportionate to size; in the second stage, households (usually 16) are randomly selected from each of the designated

Primary Sampling Units. Clustering and stratifying are permitted, but all sampling procedures must be documented and made available to data analysts (LSMS, 1996).

3.2.4 Survey of Health, Aging and Retirement in Europe (SHARE) sampling designs vary by country but all are required to be probability samples. Three sampling designs may be used: (1) stratified simple random sampling from national population registers, (2) multi-stage sampling using regional or local population registers, or (3) single or multi-stage sampling using telephone directories followed by screening in the field (De Luca et al., 2015).

3.2.5 Sampling frames for the World Mental Health Survey vary across participating countries, but generally consist of three types of sampling frames: (1) individual contact information databases such as national population registries, voter registration lists, or household telephone directories, (2) multistage area probability sample frames, or; (3) hybrid multistage frames that combine area probability methods and a individual contact database in the final stages. Sampling designs vary across participating countries, including stratification and clustering, but probability sampling is required at all stages. The target minimum response rate is 65% (Kessler et al., 2008).

3.2.6 Probability sampling is strongly recommended, but not required, in the World Value Survey; any deviations from probability sampling are to be reported in the Methodology Questionnaire report (http://www.worldvaluessurvey.org/wvs).

3.3 When there is very little information on the population, surveys (including Europe-wide surveys, surveys in developing countries and worldwide international surveys) sometimes use random route (random walk) as part of the multistage face-to-face interview sampling method.

3.3.1 For each randomly-chosen sampling points (e.g., urban units, small cities, or voting districts), interviewers are assigned with a starting location and provided with instructions on the random walking rules – e.g., which direction to start, on which side of the streets to walk and which crossroads to take. Households are selected by interviewers following the instructions. The routes end when the predefined number of respondents (or households) is achieved (Bauer, 2016). Since the probability of the selected household is unknown, this method is categorized as non-probability sampling methods (Bauer, 2016). See Table 1 in Bauer (2016) fora list of surveys which used this method in data collection.

3.3.2 The advantage of random route sampling is that it is usually cheaper and easy to conduct. Most importantly, this sampling method makes it possible when there is no list of respondents available (Bauer, 2016).

3.3.3 This method assumes that each household in this sampling point has equal probability to be selected. However, a recent study evaluated the bias in random route sampling using registration office data (to verify the impact of selection errors on survey results), and found that "all tested routes strongly violate the equal probability assumption and lead to biased expected values in multiple variables" (Bauer, 2016). Additional error source also includes incorrect interviewer behaviors, which can lead to coverage and sampling error (also see: Paradata and Other Auxiliary Data).

3.3.4 In a review of a survey of blindness in war-torn southern Sudan, Kuper and Gilbert (2006) note that the estimate of trachoma prevalence is higher than previously reported for Africa, including other studies in conflict-affected places. Kuper and Gilbert suggest that methodological problems, including an element of subjectivity resulting from the random route method employed for household selection, may have resulted in an over estimate of the disease.

4. Determine the sample size necessary to meet the desired level of precision for the statistics of interest at population or subgroup levels for the different potential sample selection procedures.

Rationale

After choosing a sample design, and before selecting the sample from the sampling frame, the sample size must be determined. The sample size takes into account the desired level of precision for the statistic(s) of interest, estimates of the statistic of interest from previous surveys, the design effect, and estimated outcome rates of the survey. (See Lynn, Häder, Gabler, & Laaksonen (2007) for a detailed treatment of the approach used in the European Social Survey. For a more extensive example of sample size calculation, see Appendix D.)

Procedural steps

4.1 Specify the desired level of precision, both overall and within key subgroups. Practical experience has determined that often it is easiest for sponsors to conceptualize desired levels of precision in terms of 95% confidence intervals.

4.2 Convert these 95% confidence intervals into a sampling variance of the mean or proportion.

4.3 Obtain an estimate of S^2 (population element variance).
- 4.3.1 If the statistic of interest is not a proportion, find an estimate of S^2 from a previous survey on the same target population or from a small pilot study.
- 4.3.2 If the statistic of interest is a proportion, the sampler can use the expected value of the proportion (p), even if it is a guess, to estimate S^2 by using the formula $s^2 = p(1-p)$.
- 4.3.3 If no information about p is available, a researcher can assume that it is 50%, which will yield a conservative estimate of the sample size.

4.4 Estimate the required number of completed interviews for an SRS by dividing the estimate of S^2 by the desired sampling variance of the mean. See Cochran (1977) for more on sample size computation for SRS.

4.5 Multiply the number of completed interviews by the design effect to account for a non-SRS design.

4.6 Calculate the necessary sample size by dividing the number of completed interviews by the expected response rate, eligibility rate, and coverage rate.
- 4.6.1 The sampler can estimate these three rates by looking at the rates obtained in previous surveys with the same or similar survey population and survey design.

Lessons learned

4.1 Prior to the first implementation of the European Social Survey (ESS), many of the participating survey organizations had never encountered the concepts of sample size determination and calculating design effects (Lynn et al., 2007). Therefore, the ESS expert sampling panel spent considerable time explaining these. In return, the organizations that were new to these methods were very enthusiastic to learn about them, and eager to meet the standards of the coordinating center. In fact, after completing Round 1 of the study, many nations commented that designing the sample was one of the most educational aspects of the entire survey process, and had significantly improved the survey methods within their country.

4.2 Sample size frequently varies among countries participating in cross-cultural surveys. In Round 6 of the Afrobarometer Survey, sample size ranges from a minimum of 1,200 respondents to 2,400 or more in extremely heterogeneous areas (Afrobarometer, 2014); sample size ranges from 800 to 3,200 respondents in the Asian Barometer study (Asian Barometer, 2016); Round 5 of the European Social Survey (ESS) requires a minimum of 800 respondents for participating countries that have a population of less than two million, 1,500 from larger countries (ESS, 2015); the International Social Survey Programme (ISSP) requires a minimum of 1,000 respondents, with a goal of 1,400 respondents (ISSP, 2015); sample size ranges from 1,600 to 5000 households in the Living Standard Measurement Study Survey (LSMS) (LSMS, 1996); the Survey of Health, Ageing and Retirement in Europe (SHARE) requires 1,500 respondents from each participating country (De Luca et al., 2015); samples in the World Mental Health Survey range from 2,357 (Romania) to 12,992 (New Zealand) (Kessler et al., 2008); and the World Value Survey requires a minimum of 1,000 respondents (World Values Survey, 2016).

5. Institute and follow appropriate quality control procedures at each step of the sample design process.

Rationale

Development and implementation of quality control procedures for the sample design are necessary to ensure the highest level of coverage possible and to maintain a probability sample that meets the desired level of precision for key survey statistics. If a failure to meet those standards is detected, protocols should be in place to remedy the failure. In addition, monitoring of procedures related to the sample design of the study should inform efforts to improve the quality and cost-effectiveness of the study over time.

Procedural steps

5.1 Define the target population for the study across all participating countries/cultures as well as the target population within each country/culture. If the study design does not change over time, strive to keep each target population, both overall and within each participating country, consistent over time.

5.2 Prior to selecting sample elements or sampling units, provide the data collection staff with a list of all of the variables on the sampling frame and ask which variables they would like and the format in which they would like these variables delivered once sampling is complete for data collection purposes.

5.2.1 After sample selection, check that each selected sampling unit or element contains this information and is in the specified format.

5.3 If possible, use a responsive survey design (Groves & Heeringa, 2006; Groves et al., 2008) to help achieve an optimal sampling design (see Survey Quality, Paradata and Other Auxiliary Data, and Data Collection: General Considerations for more information about responsive survey designs). A responsive survey design uses prespecified paradata (quantitative indicators of the data collection process such as "contact attempts" or "interviewer success rate") for intervention during data collection.

5.3.1 Advantages of responsive survey designs are the prespecification of interventions instead of ad hoc decisions and the possibility to target efforts on hard to interview groups.

5.3.2 A disadvantage is that the survey designers walk a thin line between full probability and quota samples if they deviate from carefully predefined paradata-driven interventions.

5.4 After each stage of selection, generate frequency tables for key variables from the frame of sampling units to check for the following:

5.4.1 Overall sample size and within stratum sample size.

5.4.2 Distribution of the sample units by other specific groups such as census enumeration areas.

5.4.3 Extreme values.

5.4.4 Nonsensical values.

5.4.5 Missing data.

5.5 Create a unique, sample identification code for each selected sampling unit. This code will allow identifying information to be easily removed after completing data collection. The code value can then be used in place of other values to preserve confidentiality when the original values may contain information that could lead to identifying respondents.

5.6 Whether the participating country or the coordinating center is selecting the sample, assign a second sampling statistician within that organization to check the sample design methodology and the statistical software syntax of the survey's primary sampling statistician.

5.7 Save all data files and computer syntax from the statistical software package used during the sample design process in safe and well-labeled folders for future reference and use.

Lessons learned

5.1 The construction and maintenance of sampling frames constitute an expensive and time-consuming exercise. If a participating country determines that no sampling frame meeting the specified coverage level of the target population exists, they can create a frame from sources such as census data collected by national statistics offices. However, one should be aware that official statistics differ greatly in accuracy, as well as availability, from country to country.

5.2 As discussed in the Lessons Learned section of Guideline 2, the decision to stray from full probability sampling reflects the conflict between standardization and flexibility in 3MC surveys. However, it bears repeating that without probability sampling, one cannot make justifiable inferences about the target population from the sample estimates.

6. Document each step of the sample selection procedure.

Rationale

Over the course of many years, various researchers will analyze the same survey data set. In order to provide these different users with a clear sense of how and why the data were collected, it is critical that all properties of the data set be documented. In terms of the sample design and selection, the best time to document is generally shortly after sample selection, when the information regarding sample selection is fresh in one's mind.

Procedural steps

6.1 Have participating countries document the sample selection procedure while selection is occurring or shortly thereafter. Ideally, set a deadline that specifies the number of days after sample selection by which each participating country must send sampling selection documentation to the host survey organization. Be sure to allow for appropriate time to review and revise documentation when setting the deadline (see Tenders, Bids and Contracts and Study Management.)

6.2 Include the following:
   6.2.1 A clear definition of the survey population, as well as the differences between the target population and survey population.
   6.2.2 The sampling frame:

- Both the sampling frame used and the date the frame was last updated.
- A description of the development of the sampling frame and the frame elements.
- A description of how well the sampling frame is thought to cover to target population and the potential for coverage error.

6.2.3 The data file of selected elements:

- A descriptive and distinct variable name and label.
- Unique variables that contain the selection probabilities at each stage of selection as well as the overall selection probabilities. If a participating country used a nonprobability method in at least one stage of selection and therefore the selection probabilities are unknown, ensure that this is clearly documented.
- A clear description of all variables in the selected element data file, with all variable names, an accompanying description and a codebook, which provides question-level metadata that are matched to variables in a dataset.
- The statistical software syntax used for checking the dataset of selected sampling units or elements.

6.3 For each sample, indicate how many stages were involved in selecting the sample (include the final stage in which the respondent was selected within the household, if applicable), and a description of each stage, including how many sampling units were selected at each stage.

6.3.1 Examples of different stages include:

- State/province.
- County or group of counties.
- City/town, community, municipality.
- Census/election district.
- Area segment/group of neighborhood blocks.
- Housing unit/physical address (not necessarily the postal address).
- Postal delivery point/address.
- Block of telephone numbers (e.g., by regional prefix).
- Telephone number.
- Household.
- Person selected from a household listing or a non household listing (a list, registry or other source)

6.3.2 Examples of how sampling units were selected:

- All selected with equal probability.
- All selected with probability proportional to size; specify the measure of size used. (See Appendix C for more on probability proportional to size sampling methods.)

- Some units selected with certainty, others selected with probability proportional to size; describe the criteria used for certainty selection.
- Census/enumeration (all units selected with certainty).
- Units selected using a nonprobability method (e.g., convenience sample, quota sample).

6.3.3 At each stage of selection, describe the stratification variables and reasons for choosing these variables

6.3.4 At each stage of selection, explain the allocation method used and the sample size for each stratum at each stage of selection. (See Appendix C for more on allocation methods in stratified sampling.)

6.4 If systematic sampling was used at any stage of selection, indicate whether the frame was sorted by any variables prior to systematic selection in order to achieve implicit stratification. If this is the case, describe the variable(s).

6.5 Describe the time dimension of the design (i.e., one-time cross-sectional, fixed panel, rotating panel design).

6.5.1 If a panel study:

- State how many previous waves or rounds of data collection there have been for this panel study.
- Describe the initial sample design for the panel study and any subsequent modifications to the design that are important in documenting this study.

6.5.2 If a rotating panel design:

- Fully describe the rotating panel design for the study (e.g., fresh cross-section is drawn each month and respondents are interviewed once that month, and then reinterviewed once six months later).
- State the anticipated precision of the estimates.
- Explain any problems encountered during the sampling process and any deviations from the sampling plan during implementation.

6.5.3 Additional sampling documentation:

- Report any (additional) subsampling of eligible respondents, carried out in order to control the number of interviews completed by respondents with particular characteristics (e.g., one in two eligible males was interviewed, one in four eligible persons with no previous history of depression was interviewed (describe protocol)).
- Describe any use of replicates (see Data Processing and Statistical Adjustment).

- Explain if releases (nonrandom subsets of total sample) were used or the entire sample was released to data collection staff at the start of the study.
- Recount in detail any substitution or replacement of sample during data collection.

Lessons learned

6.1 As the procedural steps outlined above show, selecting a sample can involve many detailed steps that may be hard to recall after the fact. For example, the coordinating center for the World Mental Health Survey began gathering sampling documentation for weighting and other purposes after many of the participating countries had finished data collection. They found that some countries had a difficult time recalling all the necessary details, such as the sample size for each stratum at each stage of selection. It is wise to document sampling procedures in detail shortly after sample selection (see Data Processing and Statistical Adjustment for further explanation of the weighting process).

Appendix A

Additional information on creating area probability sampling frames

Most of the surveys conducted in countries with limited resources are based on multistage, stratified area probability sample designs. Multistage, stratified area probability sample designs are quite common across countries. The example here is for a two-stage area probability design of HUs where the PSUs are groups of linked United States Census blocks and the SSUs are HUs (both occupied and unoccupied) within the selected blocks.

Creating and selecting primary sampling units (PSUs)

- Create PSUs. PSUs are geographic clusters. In the United States, they are often census enumeration areas, postal codes, or election districts. The size of the geographic clusters should be large enough to contain a population that is heterogeneous with respect to the survey variables of interest, but small enough to realize the travel-related cost efficiencies of clustered sample observations. Good PSUs generally have the following characteristics:
    - They possess clearly identifiable boundaries that are stable over a certain time. (Note that all administrative boundaries such as census enumeration areas, election districts, etc., are regularly updated and changed.)
    - They cover the target population completely.
    - They have measures of size for sampling purposes.
    - They contain auxiliary data for stratification purposes (see Guideline 3).
    - They are large in number.

Defining and enumerating secondary sampling units (SSUs)

- Decide on a comprehensive definition of a housing unit (HU).
    - What defines a HU and who should be counted as a household member can vary greatly across countries. For comparative surveys, often only a general definition is feasible (e.g., all persons living in private households born before xx/xx/xx in country y). Be aware that the size of a typical "private household" also varies among countries. Hoffmeyer-Zlotnik and Warner (2008) provide many household definitions used in the European Union.
    - A commonly used definition in the United States is "a physical structure intended as a dwelling that has its own entrance separate from other units in the structure and an area where meals may be prepared and served (Groves, 1989)."
    - In 1998, the United Nations defined a HU as "a separate and independent place of abode intended for habitation by a single

household, or one not intended for habitation but occupied as living quarters by a household at the time of the census. Thus it may be an occupied or vacant dwelling, an occupied mobile home or improvised HU, or any other place occupied as living quarters by a household at the time of the census. This category thus includes housing of various levels of permanency and acceptability" (United Nations, 1998).

- Determine the appropriate method for listing the HUs (SSUs) within selected PSUs.
  - One option is to use a preexisting list of HUs.
    - Some 3MC surveys have used satellite technology to help identify and list households, settlements, and habitations, especially those in hard to find areas such as mountainous, riverine, and creek regions (Okafor, Adeleke, & Oparac, 2007).
  - Another option is to send staff to list the HUs in selected PSUs.
    - Create maps to help staff efficiently travel to and correctly list all of the HUs within the selected PSUs. (See the section below on maps for creating two stage area probability frames of HUs.)
    - Use standardized protocol to consistently enumerate the HUs in selected PSUs in the field.
    - If a preexisting list of HUs for the specified PSU is available but the list is believed to be incomplete or if the coverage properties of the list are unknown, the participating country can send staff to the PSU with the pre-existing list and instructions to update and revise the list so that all HUs are included.
    - If no pre-existing list is available or the participating country knows from previous experience that the available list greatly undercovers the HUs in the PSU, have staff enumerate all the HUs in the PSU without the aid of a list.
    - If some selected PSUs have lists of HUs that, at least marginally, cover all its HUs and other PSUs do not, a combination of these listing methods can be used.

Creating maps to help staff locate PSUs and enumerating SSUs

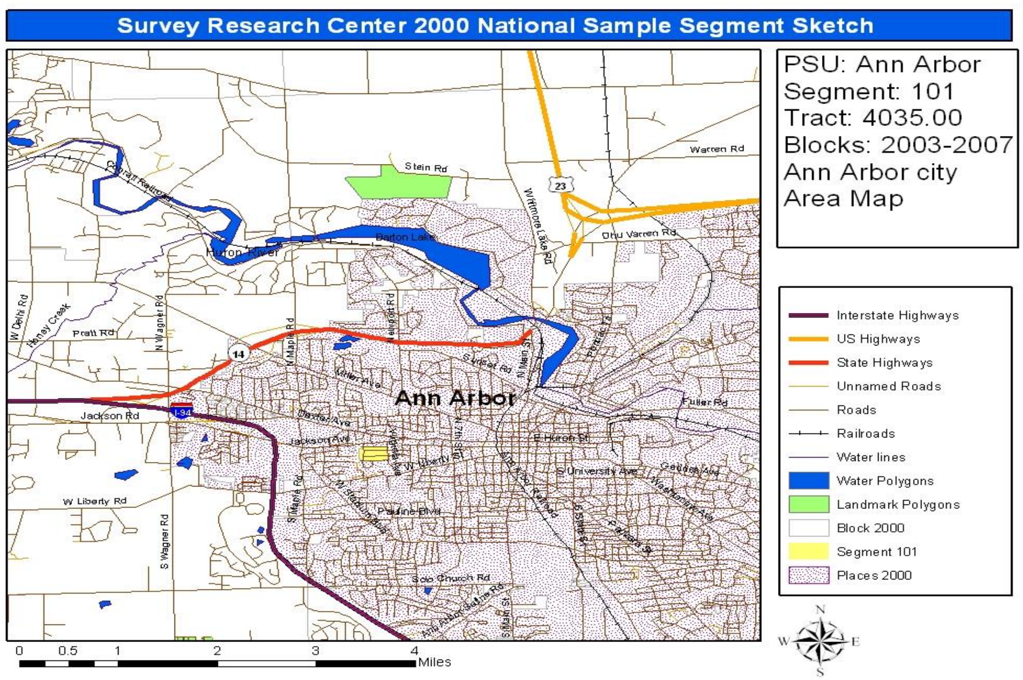
- Maps can be created on paper by hand or electronically with a mapping program like ArcGIS (Mankoff, 2009) that uses geographic data. Likewise, maps may be distributed on paper or electronically.

Area Maps

- The purpose of the area map is to show a geographic area large enough to provide context for locating the selected PSUs. Useful area maps typically contain the following features:

- Map Layout: Create area maps so that the top of the map indicates north and the top right corner of the map page displays the name of survey areas, and their associated area numbers.
- Map Legend: Located under the area information, the legend identifies roads, streets, and highways. Water boundaries for creeks, streams, rivers and lakes can be coded blue. Railroads can be indicated with a cross-hatched line.
- Distance Scale: At the bottom of the map, a scale indicates the range of miles/kilometers the map encompasses.

Example of an area map (Survey Research Center, 2006)



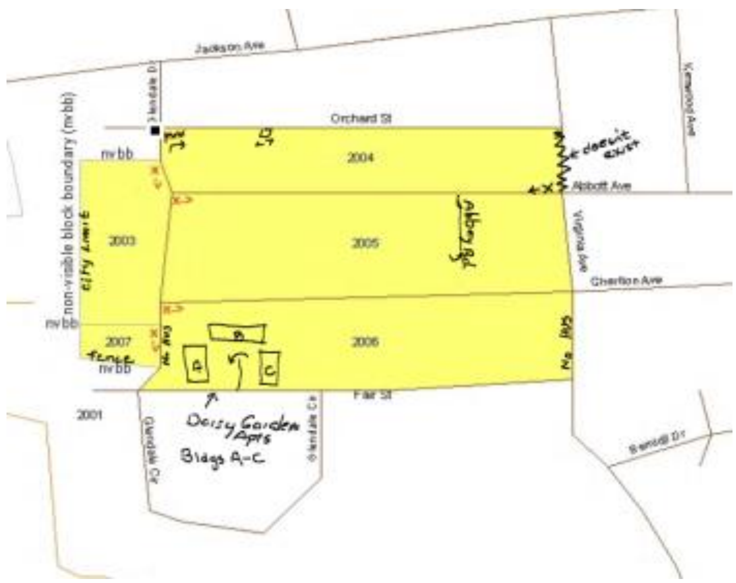
PSU Maps

- The purpose of a PSU map is to update or correct street names, note the line of travel used when listing, and draw landmarks or physical boundaries that will help future interviewers find all the listed HUs in the PSU. Below are detailed instructions for creating PSU Maps:

- Starting X and Directional Arrows: Draw a starting X and directional arrows to assist with the listing assignment. Make an effort to determine a logical starting place for listing each block, like the Northeast corner.

- Once listers have visited the PSU in person, best practice suggests that they update the map to accurately reflect the defined geographical area including:
  - Obtaining information about new streets or housing construction. A visit to the city or county planning office, or the engineering or highway department in the area can usually provide the information to accurately record current conditions in the segment area.
  - Recording “no household units” along any block face that is clearly devoid of HUs, such as those with parks, vacant fields, parking lots, woods, farm land, or only commercial or industrial structures.
  - Recording street names missing from the map(s), drawing in streets, alleys or cul-de-sacs not shown on the map, and correcting misspelled or incorrect street names. Verify that street names are complete.

Example of PSU map (Survey Research Center, 2006)



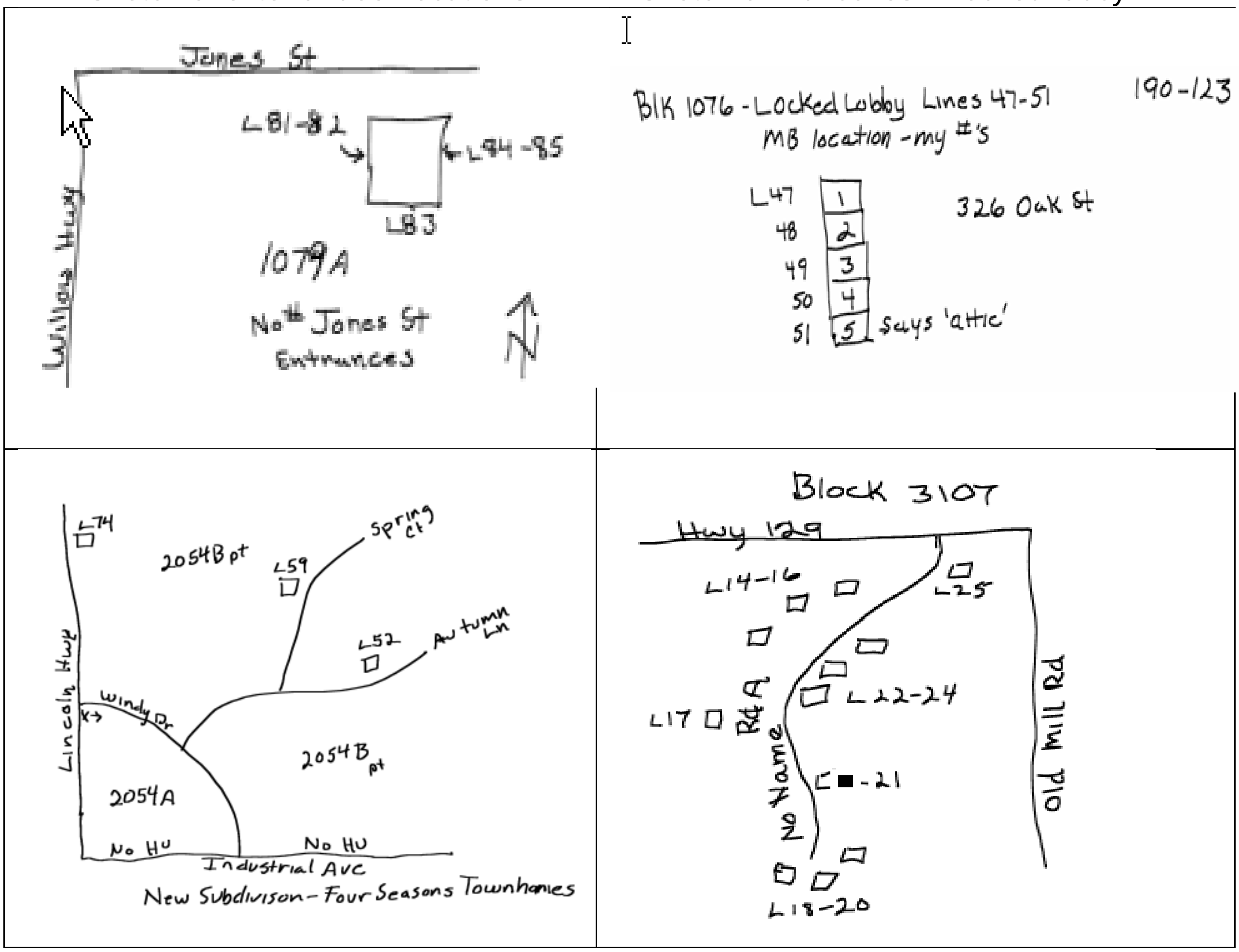
Sketch Maps

- The purpose of a sketch map is to allow listers to supplement the PSU map with their own hand-drawn map when the area and PSU maps provided seem inadequate.
- A few examples of sketch maps are provided below but the list is not all-inclusive (Survey Research Center, 2006). Sketch Map 2 is an example of sketch map used in rural area in China in the China Health and Retirement Longitudinal Study (CHRLS). For more information on the methods for listing in areas without street addresses, rural, or unmapped areas, refer to Survey Research Center (1990) and Survey Research Center (1985).

Sketch Map 1: An example of a sketch map used in a survey in the United States (Survey Research Center, 2006)

Sketch of exterior door locations

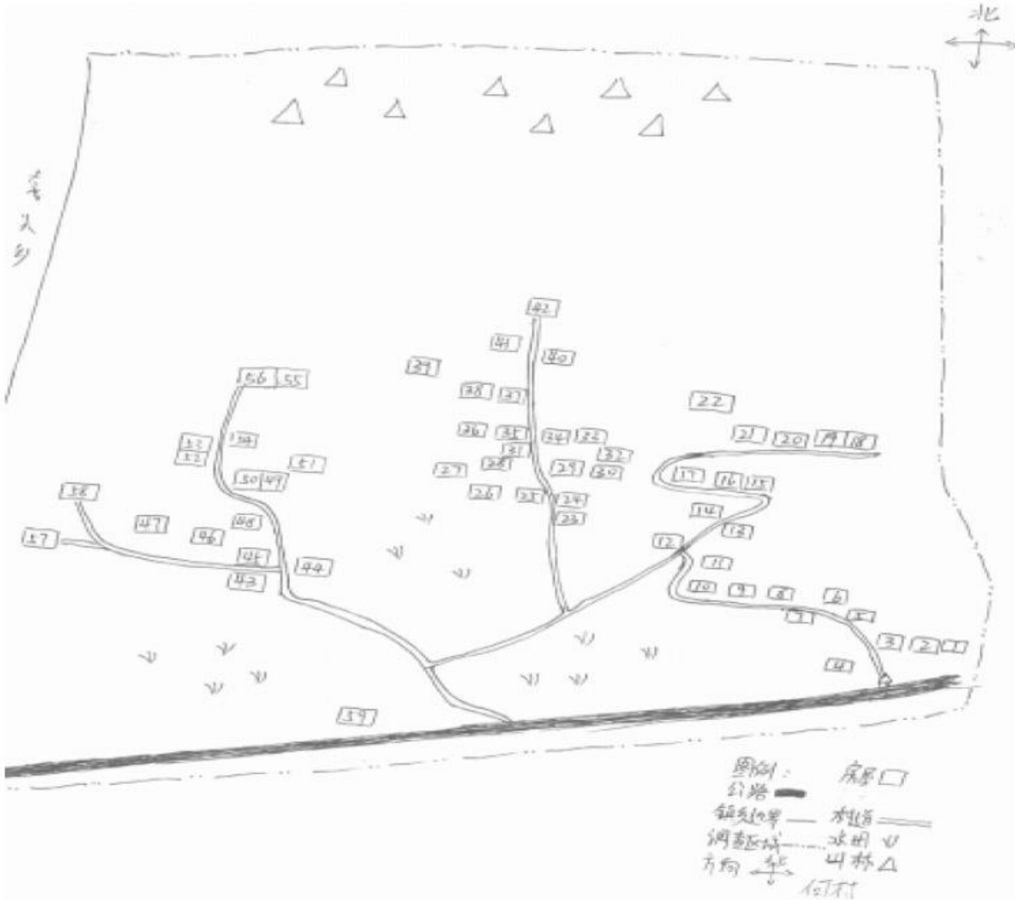
Sketch of mailboxes in locked lobby



Sketch of street configuration for a new subdivision

Sketch of "No Name" Road

Sketch Map 2: An example of a sketch map used in the China Health and Retirement Longitudinal Study (CHRLS) (China Health and Retirement Longitudinal Study, 2009)



公路	镇边界	调查区域	方向 北	房屋	村道	水田	山林
Road	Village Boundary	Survey Area	Direction	House Unit	Village Street	Paddy Field	Forest Area

Tasks for listing staff to complete prior to enumeration of HUs

- Train listers to complete the following tasks prior to beginning the listing procedure:
  - Contact local authorities. A survey organization can provide listers a letter and a form to deliver to the local police station or some other local authority, alerting them to the survey's presence in their area. The letter to the local authorities might define the purpose for listing efforts and also give staff a chance to gather information about the local situation. See Data Collection: Face-to-Face Surveys for further discussion.
  - Scout the selected areas. Most experienced listers make a complete circuit of selected areas once before beginning listing to

get a "feel" for the area. The purpose is to help find the assigned areas and to confirm or correct boundaries if maps obtained are hard to read.
  - It is helpful if listing staff estimate the number of HUs and look for indicators that may explain the discrepancy if it appears that there are twice as many HUs or fewer than half as many HUs as expected by the census count (e.g., new apartment complex or subdivision, HUs which have been demolished in the recent past, or older homes which have been converted to other uses).

<u>Recording the listed HUs</u>

- Listing is an exact record of all HUs, both occupied and unoccupied, that are located in predefined census geographical area boundaries.

- Elements of Listing Format for United States HUs include:
  - Block: borough, planned residential area or village number.
  - Non mailable (NM) indicator which is used to identify addresses that cannot receive postal mail because the address is not complete or not unique.
  - Line Number (Line_No): the first HU recorded in every listing begins at Line Number 1 with the subsequent HUs encountered being numbered consecutively 2, 3, 4 and so on through all HUs found in the PSU.
  - Street Number (Street_No): the street address number should be complete.
  - Street Name: check the spelling of street names on sign posts against the street names given on the PSU maps.
  - Apartment Number (Apt/Lot): this field should be used for apartment or trailer lot numbers only.

Example of listing format

Block	NM	Line_No	Street_No	Street Name	Apt/Lot	Additional Information
3003	X	1	NO#	FIRST AVE		
3003		2	654	FIRST AVE		
3003	X	3	1233	WILSON ST		
3003	X	4	1233	WILSON ST		Same Street_No as Line_No 3
3003		5	1241	WILSON ST		
3004	X	6	NO#	WILLOW HWY		

<u>Additional protocols to help create consistent listings</u>

- Make HUs listings consistent across all selected areas. The suggested listing protocols also include:
  - Begin by listing the lowest numbered block first. Work systematically around the PSU, listing each block in numerical order from lowest to highest block number.
  - Start listing HUs for each block beginning at the red starting "X" and following the directional arrow indicated on the Block Map.
  - Look "over your right shoulder" and record each HU address as it is approached. In other words all listed HUs should be on your right.
  - Walk around the block in a clockwise direction making a complete circuit until reaching the original starting point.
  - List only HUs inside the selected (shaded) PSU boundary.
  - List empty, boarded up, burned or abandoned HUs unless the HU is posted for demolition.
  - List on foot whenever possible when you are working in urban and suburban areas.

- Create general rules to deal with the following situations:
  - Abandoned, boarded up, burnt out, and vacant HUs.
  - Apartment complexes.
  - Locked buildings and gated communities.
  - New construction.
  - Under construction or unfinished construction.

- Check the completed listing.
  - Review the listed addresses against the block map for each block in the PSU. Beginning at the starting "X" and proceeding clockwise around each block, confirm that there are HUs listed for every street in the block or that HUs without a street number have been noted along the proper block face on the PSU map.
  - Confirm that there is only one HU per listed line and that listing lines are used only for HUs. Commercial or public buildings such as churches, schools, or businesses should be recorded only in the PSU observations or noted on the map(s).

    - Make certain that all HUs without a street number are uniquely described in the additional information column and that their locations are noted on the map by line number.

- Review the PSU observations and make sure they are complete. Confirm that information about locked building, seasonal accessibility, and safety issues are noted in detail.

Appendix B

Administration of the within housing unit (HU) listing of eligible persons

- Identify the eligibility of the selected HUs by listing the eligible persons within each selected HU (list of household members).

- Choose a HU residency rule to identify eligible respondents within each HU. Similar to defining a target population, once the rule is defined, it should be consistent across all participating countries. Choose between:
  - *De facto* residence rule – persons who slept in the HU the previous night.
    - Advantage: Easy to remember.
  - *De jure* residence rule – persons who "usually" sleep in HU.
    - Advantage: A better representation of the typical residents of a HU.
  - Design a household enumeration table based on study-specific residence rules and goals.
  - There are at least two sources of within-household undercoverage (Martin, 1999; Tourangeau, Shapiro, Kearney, & Ernst, 1997):
    - Motivated misreporting (deliberate concealment): household reporters deliberately conceal members for a multitude of reasons, including fear that they or another member may be evicted or deported.
    - Poor fit between living situation and definition: membership is complex and shows that household members may have confusion or disagree about who is a member.

Example 1 of a within-household listing table

	HOUSEHOLD LISTING					RESPONDENT SELECTION		
	11 a. Household Member's First Name	11 b. HH Member's Relationship to Informant	11 c. Sex	11 d. Year of Birth	11 e. Language Spoken	11 f. Eligible	11 g. Person Number	11 h. Selected R
MALES			M					
			M					
			M					
			M					
			M					
			M					
			M					
FEMALES			F					
			F					
			F					
			F					
			F					
			F					
			F					

Instructions for using the household listing table

Column 11a (Household Member's First Name): List all members of the household, beginning with the informant.

Column 11b (Household Member's Relationship to Informant): Record each household member's relationship to the informant (e.g., husband or wife, son or daughter, mother or father, brother or sister, friend, etc.).

Column 11d (Age): Record each household member's age.

Column 11e (Language Spoken): This column may or may not be included, depending upon the study requirements.

Column 11f (Eligible): Place a check mark in this column if, based upon the information in columns 11a-11e, the household member meets the eligibility criteria for the study.

Column 11g (Person Number): Assign a sequential number to each eligible household member.

Column 11h (Selected R): Count the number of eligible persons in the household. Find that number in the Kish table in the "If the Number of Eligible Persons is:" column. The selected respondent will be the household member with the "Person Number" corresponding to the "Interview the Person Numbered:" column in the Kish table (For more information about Kish tables, see Data Collection: Face-to-Face Surveys).

- Example 2 below is the 2010 Chinese Family Panel Study enumeration table.

This study found the main challenge of listing to be situations where urban and rural villages were adjacent to one another. These situations contained complicated building structures and mixed populations (part-time and nonresident population). Therefore, the table specifically documents when more than one HU was located within a single dwelling, the reason the registered person had moved out, the time when the registered person moved out, and where the registered person had moved.

Example 2 of household enumeration table (Chinese Family Panel Studies, 2010)

Tile: ____________
City: ____________Village Number: ___________ Lister: ___________ : Supervisor: ___________

Order	Name of HU Head	Registered Address	Registered order of Resident (Person)	Actual order of Resident (Person)	The person has lived here LESS than six months (please check)	If the person has moved out, please fill out this column. Reason (choose one) 1. Marriage. 2. Living with other relatives. 3. Moving to another place. 4. Having business at other place. 5. Not actually living but having registered record here. 6. Others.	Where this person is living? 1. The same city 2. The same county 3. the same province 4. Out f this province	Eligibility (Check if the person is eligible)	HU Number (Assigned by supervisor)	Additional note
	(1)	(2)	(3)	(4)	(5)	(6)	(8)	(9)	(10)	(11)
1										
2										
3										
4										
5										
6										
7										
8										
9										
10										
NO. of HUs in this address										

Appendix C

Additional information on different sampling techniques and terminologies

Simple Random Sampling (SRS)

- SRS uses a sampling frame numbered 1 to *N* (the total number of elements on the frame). Random numbers from 1 to *N* are selected from a table of random numbers or a random number generator.

- Formula for estimating the sampling variance of a simple random sample:

  var ($\bar{y}$) = $\frac{(1-f)s^2}{n}$, where

  f is the finite population correction and is equal to n (the sample size) divided by N (the number of elements on the sampling frame);
  s^2 is the sample element variance of the statistic of interest

  $$s^2 = \frac{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}{n-1}$$

- The finite population correction indicates that, unlike the assumption made in standard statistical theory that the population is infinite, the survey population is finite in size and the sample is selected without replacement (Heeringa & O'Muircheartaigh, 2010).

Systematic Sampling

- Steps of Systematic Sampling.
  - Compute the selection interval (*k*) as the ratio of the population size, *N*, to the sample size, *n*. In a formula, $k = \frac{N}{n}$.
  - Choose a random number from 1 to *k*.
  - Select the element of that random number from the frame and every *k*th element thereafter.

- Example 1.
  - Imagine the size of the sampling frame is 10,000 and the sample size is 1,000, making the sampling interval, *k*, $\frac{10000}{1000} = 10$. The sampler then selects a random number between 1 and 10, for

instance, 6. The sampler will then make selections in this order – 6, 16, 26, 36…9996.
- Additional steps if the selection interval is a fraction:
  - Compute the selection numbers by adding the fractional sampling interval each time.
  - Drop the decimal portion of the selection numbers.

- Example 2.
  - The size of the sampling frame is 10,400 and the sample size is 1,000, making the sampling interval, *k*, $\frac{10400}{1000} = 10.4$. The sampler selects a random number between 1 and 10.4, for instance, 6. The selection numbers would then be – 6, 16.4, 26.8, 37.2…10395.6. After rounding down, the selection numbers become – 6, 16, 26, 37…10395.

Stratified Sampling

- Stratified sampling steps:
  - Find information for every element on the frame that can be used to partition the elements into strata. Use information that is correlated to the measure(s) of interest. Each element on the frame can be placed in one and only one group.
  - Sort the frame by strata.
  - Compute a sample size (see Guideline 5).
  - Determine the number of sample selections in each respective stratum (allocation).

- There are 3 main types of allocation:
  - Proportionate allocation.
    - Selecting the sample so that elements within each stratum have the same probabilities of selection. Another way to conceive of proportionate allocations is that the sampler selects a sample of size n_h from each stratum *h* such that the proportion of elements in the sample from stratum *h,* $\frac{n_h}{n}$, is the same as the proportion of elements on the frame from stratum N_h, $\frac{N_h}{N}$.
  - Equal allocation.
    - An allocation where the same number of elements are selected from each stratum.
    - If one knows that all strata have equal distributions of the statistic of interest on the sampling frame, an equal allocation will create the highest level of precision in the sample estimate.

- Optimal allocation.
  - An allocation that produces the highest precision (i.e., narrowest confidence intervals) for the sample mean of any statistic of interest.
  - The sampler needs accurate estimates of the distributions of the frame elements for each stratum on the statistic of interest.

Cluster Sampling.

- Within-cluster homogeneity:
  - When selecting people, it is important to consider that people within a cluster tend to be more similar than people across clusters because of:
    - Environment.
    - Self-selection.
    - Interaction with one another.
  - Since elements within a cluster tend to be alike, we receive less new information about the population when we select another element from that cluster rather than from another cluster. This lack of new information makes a cluster sample less precise than a stratified or even simple random sample. The rate of homogeneity (*roh*) is a way to measure this clustering effect.

Design effect

- A survey's design effect is defined as the ratio of the sampling variance under the complex design to the sampling variance computed as if a simple random sample of the same sample size had been selected. The purpose of the design effect is to evaluate the impact of the complex survey design on sampling variance measured to the variance of simple random sampling as the benchmark.
- For a cluster sample, the design effect is the effect of having chosen sampled clusters instead of elements. Due to within-cluster homogeneity, a clustered sample cannot assure representation of specified population subgroups as well as SRS, and will tend to have a design effect greater than one. On the other hand, stratification tends to generate design effects less than one since it ensures that specified population groups will be allocated at least one sample selection.
  - In general, clustering increases the design effect, while stratification decreases it.
  - Formulas:

Stratified designs

$$d_{eff} = \frac{\mathrm{var}(\bar{y}_{complex})}{\mathrm{var}(\bar{y}_{SRS})}$$

where d_{eff} is the design effect;

$\mathrm{var}(\bar{y}_{complex})$ is the variance of the complex sample design, whether it be stratfied only, clustered only, or a stratified cluster design;

$\mathrm{var}(\bar{y}_{SRS})$ is the variance of a SRS design, with the same sample size

$$d_{eff} = 1 + (b-1)roh$$

where d_{eff} is the design effect;

b is the number of subselections within a selected cluster;

and roh is the rate of homogeniety

- In order to estimate the design effect for a new study, the roh is calculated from an earlier survey on a similar topic within a similar target population.

- Subsampling within selected clusters (multi-stage sampling).
  - *n = a*b,* where *n* is the sample size, *a* is the number of clusters selected and *b* is the number of selections within each cluster.
  - Pros: reduces the design effect and makes estimates more precise.
  - Cons: increases total costs because need to send interviewers to more areas.

Probabilities Proportional to Size (PPS)

- Situations where clusters are all of equal size rarely occur. PPS can control the sample size while ensuring that each element on the sampling frame has an equal chance of selection.

- Probabilities at either the first or second stage can be changed to ensure equal probabilities of selection for all elements.

- Imagine a two-stage cluster design where the clusters were blocks and the elements were housing units (HUs). The PPS formula would be:

$$f = f_{block} * f_{hu} = \frac{\alpha B_\alpha}{\sum B_\alpha} * \frac{b}{B_\alpha}$$

where
f is the overall probability of selection of the element,
f_{block} is the probability of selection of the cluster, and
f_{hu} is the probability of selection of the element within the cluster,
α is the number of cluster sections,
B_{α} is the number of elements within the selected sections α on the frame,
$\sum B_{\alpha}$ is the number of elements on the frame,
b is the number of elements selected within cluster α.

Example

Block #	Housing Units in Block	Cumulative Housing Units
1	25	25
2	30	55
3	35	90
4	40	130
5	20	150

- The sampler has the above list of blocks and wants to select three blocks (*a*), keep the sample size constant at 15 HUs and ensure that each HUs has the same probability of selection of one in ten (*f*=15/150). Using cumulative totals, numbers can be assigned to each block. Block 1 is assigned numbers 1-25, Block 2 26-55, Block 3 56-90, Block 4 91-130 and Block 5 131-150. From here, systematic sampling can be used to obtain a simple, without replacement sample of blocks based on the HUs within each block. Based on the frame size of 150 ($\sum B_{\alpha}$) and the number of selections being three, the selection interval is 50. Suppose the sampler chooses a random start of 29. In this case, the selection numbers would be 29, 79, and 129 corresponding to selections of Block 2, Block 3 and Block 4. To determine the selection probability of the HUs within Block 2 (f_{hu}), use the formula:

$$f = f_{block2} * f_{hu}$$

$$\frac{1}{10} = 3\left(\frac{30}{150}\right) * f_{hu}$$

$$f_{hu} = \frac{1}{10} * \frac{150}{90} = \frac{1}{6}$$

Since the selection probability of HUs within Block 2 is 1/6, the number of HUs selected within Block 2 (*b*) will be 30*1/6 or 5. Going through the same calculations for Blocks 3 and 4 will show that each block will have five selections.

- Potential problems and solutions with PPS sampling.
  - Problem: The same cluster may be chosen more than once.
    Solution: Use systematic selection with PPS (Kish, 1965).
  - Problem: Some of the clusters may not be large enough to produce subsamples of the required size.
    Solution: Link clusters to create new clusters that are all of sufficient size.
  - Problem: Some of the clusters are too large and the probability of selecting the cluster is greater than one.
    Solution: Remove the cluster from the list and choose elements from it directly.

Two-Phase Sampling

- Suggested steps (Groves et al., 2009):
  - Phase 1 – Conduct a survey on a probability sample, using a relatively inexpensive data collection method subject to higher nonresponse rates than more expensive methods (see Data Collection).
  - Once the survey is completed, select a probability subsample of the nonrespondents to the Phase 1 survey.
  - Phase 2 – Use a more expensive method that generally produces lower nonresponse on the subsample.
  - Combine the results of the two surveys, with appropriate selection weights to account for unequal probabilities of selection between the selected respondents.

Panel Designs

- Three concerns about panel designs:
  - The effort and costs of tracking and locating respondents who move over the duration of the panel survey.
  - The change in the elements on the sampling frame over time. For example, in a cross-cultural panel survey of persons age 65 and older, some members of the original sampling frame will die, while other people will become eligible for selection.
  - The repeated questioning of the same subjects over time may change how the subjects act and answer the questions (i.e., panel conditioning effect).

Appendix D

Sample Size Determination

- Recommended steps.
  - Define how many nested cells will be relevant for the analysis and what should be the minimal number of cases in each cell allowing for substantial analyses.
  - Have the survey sponsor specify the desired level of precision.
  - Convert these 95% confidence intervals into a sampling variance of the mean, $\mathrm{var}(\bar{y})$.
    - Example: The survey sponsor wants a 95% confidence interval of .08 around the statistic of interest. Since the half width of a 95% confidence interval (CI) is $\frac{1}{2}(95\%\ CI) = 1.96(se(\bar{y})).$ This formula can be rearranged with basic algebra to calculate the precision (sampling variance of the mean) from this confidence interval:

$$\mathrm{var}(\bar{y}) = (se(\bar{y}))^2 = \left(\frac{.5(95\%\ \text{Conf. Interval})}{1.96}\right)^2 = \left(\frac{.04}{1.96}\right)^2 = .0004165.$$

  - Obtain an estimate of S^2 (population element variance).
  - If the statistic of interest is not a proportion find an estimate of S^2 from a previous survey on the same target population or a small pilot test.
  - If the statistic of interest is a proportion, the sampler can use the expected value of the proportion (*p*), even if it is a guess, to estimate S^2 by using the formula $s^2 = p(1-p)$.
  - Estimate the needed number of completed interviews for a simple random sample (SRS) by dividing the estimate of S^2 by the sampling variance of the mean.
    - Example: the obtained estimate of S^2 is .6247. Therefore the needed number of completed interviews for an SRS (n_{srs}) is:

$$n_{srs} = \frac{.6247}{.0004165} = 1,499.88 \approx 1,500.$$

  - Multiply the number of completed interviews by the design effect to account for a non- SRS design.
    - Example: the design effect of a stratified clustered sample is 1.25. Taking into account the design effect, the number of completed interviews for this complex (i.e., stratified clustered) sample is: $n_{complex} = n_{srs} * d_{eff} = 1,500 * 1.25 = 1,875$.
  - The sample size must account for three additional factors:

- Not all sampled elements will want to participate in the survey (i.e., response rate).
- Not all sampled elements, given the target population, will be eligible to participate (i.e., eligibility rate).
- The frame will likely fail to cover all elements in the survey population (i.e., coverage rate).
- Calculate the necessary sample size by dividing the number of completed interviews by the expected response rate, eligibility rate, and coverage rate.
- The sampler can estimate these three rates by looking at the rates obtained in previous surveys with the same survey population and survey design.
    - Example: The expected response rate is 75%, the expected eligibility rate is 90%, and the expected coverage rate is 95%. Therefore, the necessary sample size is:

$$n_{final} = \frac{n_{complex}}{\text{Resp rate} * \text{Elig rate} * \text{Cov rate}} = \frac{1875}{.75 * .9 * .95} = 2923.97 \approx 2924.$$

References

Afrobarometer Survey. (2014). Round 6 Survey Manual. Retrieved June 5, 2016, from http://www.afrobarometer.org/sites/default/files/survey_manuals/ab_r6_survey_manual_en.pdf

Asian Barometer Survey (ABS). (2014). Retrieved Nov 28, 2014, from http://www.asianbarometer.org

Barahona, C., & Levy, S. (2006). The best of both worlds: Producing national statistics using participatory methods. *World Development, 35*(2), 326-341.

Bauer, J. J. (2016). Biases in random route surveys. *Journal of Survey Statistics and Methodology*, *4*(2), 263-287.

Bergsten, J. W. (1980). *Some sample survey designs in Syria, Nepal and Somalia.* Paper presented at the Proceedings of American Statistical Association: Survey Research Methods Section. Retrieved March 25, 2010, from http://www.amstat.org/Sections/Srms/Proceedings/papers/1980_013.pdf

Chikwanha, A. B. (2005). *Conducting surveys and quality control in Africa: Insights from the Afrobarometer*. Paper presented at the WAPOR/ISSC Conference, Ljubljana, Slovenia.

China Health and Retirement Longitudinal Study (CHRLS). (2009). Advance team manual*.* Retrieved June 04, 2016 from http://charls.ccer.edu.cn/uploads/document/public_documents/application/Advance_team_manual________.pdf

Chinese Family Panel Studies (CFPS). (2010). *Listing manual (draft)*. Unpublished manuscript.

Cochran, W. G. (1977). *Sampling techniques*. New York, NY: John Wiley & Sons.

De Luca, G., Rossetti, C., & Malter, F. (2015). Sample design and weighting strategies in SHARE Wave 5. In F. Malter & A. Borsch-Supan (Eds.), *SHARE Wave 5: Innovations & Methodology* (pp. 75-84). Munich: MEA, Max Planck Institute for Social Law and Social Policy.

Dunzhu, S., Wang, F. S., Courtright, P., Liu, L., Tenzing, C., Noertjojo, ...Bassett, K. L. (2003). Blindness and eye diseases in Tibet: Findings from a randomised, population based survey. *British Journal of Ophthalmology, 87*(12), 1443-1448.

Eckman, S., Himelein, K., & Dever, J. A. (2017). Sample designs using GIS technology for household surveys in the developing world. In T. P. Johnson, B-E. Pennell, I. A. L. Stoop, & B. Dorer (Eds.), *Advances in comparative survey methods: Multicultural, multinational and multiregional contexts (3MC).* New York, NY: John Wiley & Sons.

European Social Survey (ESS). (2010). Retrieved May 03, 2010, from http://www.europeansocialsurvey.org

European Social Survey (ESS). (2012). *Sampling guidelines*. Retrieved June 10, 2016, from http://www.europeansocialsurvey.org/docs/round6/methods/ESS6_sampling_guidelines.pdf

European Social Survey. (2014). *Sampling for the European Social Survey Round VII: Principles and Requirements*. Retrieved May 5, 2016, from https://www.europeansocialsurvey.org/docs/round7/methods/ESS7_sampling_guidelines.pdf

European Social Survey. (2015). *ESS5 – 2010 Documentation Report*. Retrieved May 5, 2016, from http://www.europeansocialsurvey.org/docs/round5/survey/ESS5_data_documentation_report_e04_0.pdf

Groves, R. M. (1989). *Survey errors and survey costs*. New York, NY: John Wiley & Sons.

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*, 439–457.

Groves, R. M., & Lepkowski, J. M. (1985). Dual frame, mixed mode survey designs. *Journal of Official Statistics, 1*(3), 263-286.

Groves, R. M., Fowler, F. J. Jr., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2009). *Survey methodology* (2nd ed.). Hoboken, NJ: John Wiley & Sons.

Groves, R. M., Kirgis, N., Peytcheva, E., Wagner, J., Axinn, W., & Mosher, W. (2008). *Responsive design for household surveys: Illustration of management interventions based on survey paradata*. Paper presented at the 2008 AAPOR meetings.

Häder, S., & Gabler, S. (2003). Sampling and estimation. In J. A. Harkness, F. J. R. van de Vijver, & P. M. Mohler (Eds.), *Cross-cultural survey methods,* (pp. 117-136). Hoboken, NJ: John Wiley & Sons.

Hartley, H. O. (1962). Multiple-frame surveys. *Proceedings of the Social Statistics Section* (pp. 203-206). Alexandria, VA: American Statistical Association.

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. P. Mohler, … T. W. Smith (Eds.), *Survey methods in multicultural, multinational, and multiregional contexts* (pp. 251–267). Hoboken, New Jersey: John Wiley & Sons.

Hoffmeyer-Zlotnik, J. H. P., & Warner, U. (2008). *Private household: Concepts and their operationalization in national and international social science surveys* (In German). Retrieved February 13, 2010, from http://www.ssoar.info/ssoar/View/?resid=2660

International Social Survey Programme (ISSP) (2015). Retrieved March 9, 2015, from http://www.issp.org/

Kalton, G. (1983). *Introduction to survey sampling.* Newbury Park, CA: Sage Publications.

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

Kish, L. (1965). *Survey sampling.* New York, NY: John Wiley & Sons.

Kish, L. (1994). Multipopulation survey designs: Five types with seven shared aspects. *International Statistical Review, 62*, 167–186.

Koch, A. (2017). Within-household selection of respondents: The last step of sampling in household surveys. In T. P. Johnson, B-E. Pennell, I. A. L. Stoop, & B. Dorer (Eds.), *Advances in comparative survey methods: Multicultural, multinational and multiregional contexts (3MC).* New York, NY: John Wiley & Sons.

Kuper, H., & Gilbert, C. (2006). Blindness in Sudan: Is it time to scrutinize survey methods? *PLoS medicine*, *3*(12), e476.

Living Standard Measurement Study Survey. (1996). *Working paper 126: A manual for planning and implementing the Living Standard Measurement Study Survey*. Retrieved July 10, 2010, from http://www-wds.worldbank.org/external/default/WDSContentServer/WDSP/IB/2000/02/24/000009265_3961219093409/Rendered/PDF/multi_page.pdf

Lohr, S., & Rao, J. N. K. (2006). Estimation in multiple-frame surveys. *Journal of the American Statistical Association*, *101*(475), 1019-1030.

Lynn, P., Häder, S., Gabler, S., & Laaksonen, S. (2007). Methods for achieving equivalence of samples in cross-national surveys: The European Social Survey experience. *Journal of Official Statistics, 23*(1), 107–124.

Mankoff, C. (2009). *Assisting the survey listing process using geography*. [PowerPoint slides]. Retrieved May 15, 2016, from https://www.urbanreproductivehealth.org/sites/mle/files/Assisting_the_Survey_Listing_Process_Web.pdf

Martin, E. (1999). Who knows who lives here? Within-household disagreements as a source of survey coverage error. *Public Opinion Quarterly, 63*(2), 220-236.

Minnesota Population Center. (2015), *Integrated public use microdata series, international: Version 6.4* [Machine-readable database]. Minneapolis: University of Minnesota.

Okafor, R., Adeleke, I., & Oparac, A. (2007*). An appraisal of the conduct and provisional results of the Nigerian Population and Housing Census of 2006*. Paper presented at the Proceedings of American Statistical Association: Survey Research Methods Section. Retrieved March 26, 2010, from http://www.amstat.org/Sections/Srms/Proceedings/y2007/Files/JSM2007-000561.pdf

Survey of Health, Ageing and Retirement in Europe (SHARE). (2010). Retrieved July 10, 2010, from http://www.share-project.org/

Survey Research Center. (1985). *Household survey of Grenada*. Ann Arbor, MI: Institute for Social Research, University of Michigan.

Survey Research Center. (1990). National area sample sampling manual-Listing instructions*.* Ann Arbor, MI: Institute for Social Research, University of Michigan.

Survey Research Center. (2006). *Listing procedures manual*. Ann Arbor, MI: Institute for Social Research, University of Michigan.

Tourangeau, R., Shapiro, G., Kearney, A., & Ernst, L. (1997). Who lives here? Survey undercoverage and household roster questions. *Journal of Official Statistics, 13*, 1-18.

Tucker, C., Lepkowski, J. M., & Piekarski, L. (2002). The current efficiency of list assisted telephone sampling designs. *Public Opinion Quarterly*, *66*(3), 321-338.

United Nations. (1998). *Principles and recommendations for population and housing censuses, revision 1*. Statistical Papers, No. 67/Rev. 1, Sales No. E.98.XVII.8. New York, NY: United Nations. Retrieved March 26, 2010, from http://unstats.un.org/unsd/pubs/gesgrid.asp?ID=127

Üstun, T. B., Chatterji, S., Mechbal, A., & Murray, C. J. L. (2005). Chapter X: Quality assurance in surveys: Standards, guidelines, and procedures. In United Nations Statistical Division, United Nations Department of Economic and Social Affairs (Eds.), *Household surveys in developing and transition countries*. New York, NY: United Nations. Retrieved March 26, 2010, from http://unstats.un.org/unsd/HHsurveys/pdf/Household_surveys.pdf

Vanden-Eng, J. L., Wolkon, A., Frolov, A. S., Terlouw, D. J., Eliades, M. J, Morgah, K., …Hightower, A. W. (2007). Use of handheld computers with global positioning systems for probability sampling and data entry in household surveys. *American Journal of Tropical Medicine and Hygiene*, *77*, 393–399.

World Values Survey (WVS). (2016). Retrieved July 7, 2016, from http://www.worldvaluessurvey.org/wvs.jsp

Yansaneh, I. (2005). Chapter II: Overview of sample design issues for household surveys in developing and transition countries. In United Nations Statistical Division, United Nations Department of Economic and Social Affairs (Eds.), *Household surveys in developing and transition countries.* New York, NY: United Nations. Retrieved March 26, 2010, from http://unstats.un.org/unsd/HHsurveys/pdf/Household_surveys.pdf

Further Reading

<u>Sampling</u>

Cochran, W. G. (1977). *Sampling techniques*. New York, NY: John Wiley & Sons.

Gabler, S., & Häder, S. (2016). Challenges of sampling for comparative surveys. In C. Wolf, D. Joye, T.W. Smith, & Y. Fu (Eds.), *The SAGE Handbook of Survey Methodology*. Thousand Oaks, CA: Sage.

Kalton, G. (1983). *Introduction to survey sampling.* Newbury Park, CA: Sage Publications.

Lohr, S. (2009). *Sampling: Design and analysis* (2nd ed.). Pacific Grove, CA: Duxbury Press.

Tillé, Y., & Matei, A. (2016). Basics of sampling for survey research. In C. Wolf, D. Joye, T.W. Smith, & Y. Fu (Eds.), *The SAGE Handbook of Survey Methodology*. Thousand Oaks, CA: Sage.

Vehovar, V., Toepoel, V., & Steinmetz, S. (2016). Nonprobability sampling. In C. Wolf, D. Joye, T. W. Smith, & Y. Fu (Eds.), *The SAGE Handbook of Survey Methodology*. Thousand Oaks, CA: Sage.

An overview of frame building and associated issues can be found in Household Sample Surveys in Developing and Transition Countries (Section B: Sample Design) http://unstats.un.org/unsd/hhsurveys/sectionb_new.htm

A summary of the challenges faced within Kenya can be found in *List and Go! Building a Sample for an Urban Household Survey in Kenya* (http://csdiworkshop.org/images/2015/Sampling/CSDI2015_HughesFINAL.pdf)

<u>Statistics</u>

Snedecor, G. W., & Cochran, W. G. (1989). *Statistical methods* (8th ed.). Ames, IA: Iowa State University Press.

Questionnaire Design

Janet Harkness with (alphabetically) Ipek Bilgen, AnaLucía Córdova Cazar, Lei Huang, Debbie Miller, Mathew Stange, and Ana Villar, 2010
Updated by Ting Yan, Sunghee Lee, Mingnan Liu, and Mengyao Hu, 2016

Introduction

The International Organization for Standardization (2012) points out that research findings can be affected by wording, question order, and other aspects of questionnaire design. The following guidelines present options for the deliberate design of questions intended for implementation in multinational, multicultural, or multiregional surveys, which we refer to as "3MC" surveys. In this context, "deliberate design" means that the questions have been specifically constructed or chosen for comparative research purposes, according to any of several criteria and strategies (Harkness, Edwards, Hansen, Miller, & Villar, 2010b). The models and strategies discussed here are applicable to a variety of disciplines, including the social and behavioral sciences, health research, and public opinion research.

This chapter presents a basic outline of the approaches available to develop questions for comparative studies, the procedures involved in each, and the advantages and disadvantages of the different approaches.

Although questionnaire design for 3MC surveys is related in various ways to question translation, adaptation, technical instrument design, pretesting, and harmonization, these topics are more fully addressed in other chapters (see Translation: Overview, Adaptation, Instrument Technical Design, and Pretesting).

This chapter borrows terminology from translation studies, which define "source language" as the language translated out of and "target language" as the language translated into. In like fashion, the chapter distinguishes between "source questionnaires" and "target questionnaires." Source questionnaires are questionnaires used as a blueprint to produce other questionnaires, usually on the basis of translation into other languages (see Translation: Overview); target questionnaires are versions produced from the source questionnaire, usually on the basis of translation or translation and adaptation (see Adaptation). Target questionnaires enable researchers to study populations who could not be studied using the source questionnaire.

Guidelines

Goal: To maximize the comparability of survey questions across cultures and languages and reduce measurement error related to question design.

1. **Ensure that questionnaire design follows basic best practice recommendations for general survey research.**

Rationale

There are three general strategies for questionnaire design (Harkness, van de Vijver, & Johnson, 2003):

1.1 To re-use questions which seem suitable that have already been used in other surveys.
1.2 To adapt questions which have been developed for other purposes to suit new needs or populations.
1.3 To write entirely new questions.

Basic questionnaire design requirements need to be met regardless of which one of the three strategies is adopted and whether the project is comparative or not.

The procedural steps presented here identify fundamental aspects of questionnaire design with which researchers should be familiar before beginning work on any questionnaire and certainly before attempting comparative design. The steps do not provide guidance on each facet of design identified or on general design issues. A wealth of survey literature addresses these topics (e.g., see Bradburn, Sudman, & Wansink, 2004; Converse & Presser, 1986; Fowler, 1995; Groves, Fowler, Couper, Lepkowski, Singer, and Tourangeau, 2009; Willimack, Lyberg, Martin, Japec, & Whitridge, 2004).

Procedural steps

1.1 Review survey methods literature and research on basic aspects of general questionnaire design. Theories contributing to question/questionnaire design include:
    1.1.1 Cognition and survey research, including theories of survey response (Schwarz, 1996; Tourangeau & Rasinski, 1988; Tourangeau, Rips, & Rasinski, 2000).
    1.1.2 Measurement error and other sources of observational errors (Biemer & Lyberg, 2003; Groves, 1989).
    1.1.3 Response styles and response bias (Baumgartner & Steenkamp, 2001; Johnson & van de Vijver, 2003; Schwarz, Oyserman, & Peytcheva, 2010; Vaerenbergh & Thomas, 2012; Yang, Harkness, Ching, & Villar, 2010).
    1.1.4 Functional equivalence and comparability (Berry, 1969; Johnson, 1998a; Verba, 1969; Mohler and Johnson, 2010).

1.2 Review literature and research on the kinds of questions that can be asked (Bradburn et al., 2004; Converse & Presser, 1986; Dillman, Smyth, & Christian, 2009; Fowler, 1995; Groves, et al., 2009; Payne, 1980). Some of the kinds of questions listed below may overlap; for example, a factual judgment question may be about behavior or may ask for socio-demographic details.

1.2.1 Knowledge questions. Knowledge questions assess the respondent's familiarity, awareness, or understanding of someone or something, such as facts, information, descriptions, or skills.

Example: *Who is the President of the United States?*

1.2.2 Factual judgment questions. Factual judgment questions require respondents to remember autobiographical events and use that information to make judgments (Tourangeau, et al., 2000). In principle, such information could be obtained by other means of observation, such as comparing survey data with administrative records, if such records exist. Factual judgment questions can be about a variety of things, such as figure-based facts (e.g., date, age, weight), events (e.g., pregnancy, marriage), and behaviors (e.g., smoking, media consumption).

Example: *During the past two weeks, how many times did you see or talk to a medical doctor?*

1.2.3 Socio-demographic questions. Socio-demographic questions typically ask about respondent characteristics such as age, marital status, income, employment status, and education. For discussion of their design and interpretation in the comparative context, see Granda, Wolf, & Hadorn (2010), Hoffmeyer-Zlotnik & Wolf (2003), and the International Organization for Standardization (2012). See also Translation: Overview and Adaptation.

Example: In *what year and month were you born?*

1.2.4 Behavioral questions. Behavioral questions ask people to report on things they do or have done.

Example: *Have you ever smoked cigarettes?*

1.2.5 Attitudinal questions. Attitudinal questions ask about respondents' opinions, attitudes, beliefs, values, judgments, emotions, and perceptions. These cannot be measured by other means; we are dependent on respondents' answers.

Example: *Do you think smoking cigarettes is bad for the smoker's health?*

1.2.6 Intention questions on behavior. Intention questions ask respondents to indicate their intention regarding some behavior. They share features with attitudinal questions.

Example: *Do you intend to stop smoking?*

1.2.7 Expectation questions. Expectation questions ask about respondents' expectation about the chances or probabilities that certain things will happen in the future. They are used in several cross-national surveys, such as Survey of Health, Ageing and Retirement in Europe (SHARE).

Example: *Thinking about the next ten years, what are the chances that you will receive any inheritance, including property and other valuables?*

1.3 Review literature and research on question formats.

1.3.1 Closed-ended question format. In closed-ended question formats, the survey question provides a limited set of predefined answer categories from which respondents choose.

Example: *Do you smoke?*

Yes ___
No ___

1.3.2 Open-ended question format. Open-ended question formats require respondents to answer questions in their own words.

Example: *What is your occupation?*
(Please write in the name or title of your occupation)

1.4 Review literature and research on response scales. Response scales are predefined sets of answer categories for a closed question from which respondents are asked to select a response. Common response scales include rating, ranking, and frequency scales.

1.4.1 Rating uses an ordered scale of response options and requires the respondent to select one position on the scale.

Example: *To what extent do you agree or disagree with the following statement?*
It is a good idea to ban smoking in public places.

Strongly agree
Somewhat agree
Neither agree nor disagree
Somewhat disagree
Strongly disagree

1.4.2 Ranking is a response format where respondents express their preferences by ordering persons, brands, etc., from top to bottom, generating a rank order of a list of items or entities. Ranking can be partial, where a longer list of responses is presented, and respondents are requested to rank a limited number.

Example*: Listed below are possible disadvantages related to smoking cigarettes. Please enter the number 1, 2, 3, or 4 alongside each possible disadvantage to indicate your rank ordering of these. 1 stands for the greatest disadvantage, 4 for the least disadvantage.*

___ *Harmful effects on other people's health*
___ *Stale smoke smell in clothes and furnishings*

___ *Expense of buying cigarettes*
___ *Harmful effects on smoker's health*

Example of partial ranking: *Out of these 13 qualities in children, please rank the 5 qualities you think are most desirable in children* (Kohn, 1969)*.*

___ *Has good manners*
___ *Tries hard to succeed*
___ *Is interested in how and why things happen*
___ *etc.*

Card sorts are another ranking technique, wherein words, statements, graphics, etc., are written onto cards which the respondent arranges according to some dimension.

1.4.3 Frequency scales are a response format where respondents are required to select the option that best describes the frequency in which certain behaviors occur.

Example*: How often did you attend live music events in past year?*

Never
Rarely
Sometimes
Often
Always

1.5 Review literature and research on types of data. Data from rating, ranking, and frequency scales can be categorized as nominal, ordinal, and numeric (Fink, 2003).

1.5.1 Nominal dada. Data are nominal or categorical when respondents are asked to name, or categorize, their answer. In nominal data, there is no numeric way to rate, rank, or otherwise differentiate response categories.

Example: *Which of these political parties did you vote for in the last national election?*

Republican Party
Democratic Party
Socialist Party
Libertarian Party

1.5.2 Ordinal data. Data are ordinal when respondents are asked to rate or order items on a list. In ordinal data, there are no real numeric values associated with the categories, and there is no way to measure distance between categories. However, the categories can be ranked or rated from one end of a scale to the other.

Example: *How much do you disagree or agree with the statement: Democracy is the best form of government*.

Strongly agree
Agree

Disagree
Strongly disagree

1.5.3 Numeric data. There are two types of numeric data: interval data and ratio data, although statistically these two types of data tend to be treated the same. With interval data, the distances between the response values (i.e., numbers) have real meaning.

Example: *The Fahrenheit temperature scale, where a 10-point difference between 70F and 80F is the same as a 10-point difference between 40F and 50F.*

Measurements of physical properties have characteristics whose quantity or magnitude can be measured using ratio scales. Ratio measurements have a true zero, unlike interval data, and comparisons between data points can be made accordingly.

Example: *The Kelvin temperature scale, where 50 kelvins is half as warm as 100 kelvins.*

Other examples include time (measured in seconds, minutes, hours), meters, kilograms.

1.6 Review literature and research on mode (i.e., the means by which data are collected) (de Leeuw, 2008; Dillman, et al., 2009; Smith, 2005). The choice of mode will affect the design options for various aspects of questionnaire and survey instrument design (e.g., length of the questionnaire, layout of instruments, and application of visual stimuli). (See Study Design and Organizational Structure and Instrument Technical Design.)

1.6.1 In terms of the standard literature, “mode” is related to whether an interviewer enters the data (as in telephone and face-to-face interviews) or the respondent enters the data (as in web surveys and paper-and-pencil surveys).

1.6.2 A second relevant aspect is the channel of communication (visual, oral, aural, tactile).

1.6.3 A third is the sense of privacy. Usually, self-administered survey modes create a greater sense of privacy than interviewer administered modes.

1.7 Review literature on techniques can be used in survey questionnaire design. Random-response technique (RRT) is a method designed to elicit reliable responses to sensitive survey items, although it is only useful for a very limited number of yes/no questions in any given survey. In RTT, respondents are randomly assigned to answer one of two yes/no questions: one sensitive and the other non-sensitive and with a known probability. The interviewer is unaware of which question is given to the respondent, and only records the answer. Based on the probability of selecting the sensitive question, the probability of respondents who answer yes to the nonsensitive question, and the proportion of respondents who answer yes to the RRT question, the is technique can be used to calculate the proportion of respondents who give an affirmative answer to the sensitive question. In a study of abortion behavior among Mexican women, RRT generated the most reliable data regarding induced abortion, when compared to face-to-face surveys and both audio and paper-based self-administered modes (Lara, Strickler, Olavarrieta, & Ellertson, 2004).

2. **Become familiar with the comparative design options available and the advantages and disadvantages of each.**

Rationale

Knowledge of the different approaches available for comparative design for surveys in multiple cultures, languages, or countries enables researchers to make informed choices for their projects.

Procedural steps

2.1 Read relevant literature (and, if possible, talk to primary researchers) to become familiar with the advantages and disadvantages of the major approaches to questionnaire design for 3MC surveys. 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 (Harkness, 2008b; Harkness, et al., 2010b; Harkness et al., 2003). See Translation: Overview for more information on methods of translation, and Adaptation for more examples of ADQ.

2.1.1 Ask the same questions and translate (ASQT). In this approach to question design, researchers ask a common set of questions of all populations studied.

- The most common way to do this is by developing a source questionnaire in one language and then producing other language versions, usually on the basis of translation or translation and adaptation. A TRAPD (Translation, Review, Adjudication, Pretesting, and Documentation) team translation model is suggested. See Translation: Overview for more information.
- Decentering is a second way to "ask the same questions." With decentering, the same questions are developed simultaneously in two languages -- there is no source questionnaire or target language questionnaire. The decentering process removes culture-specific elements

from both versions (Harkness, 2008b). However, decentering is only suitable for two language projects and its use is restricted in many ways (Harkness, 2008b; Harkness, et al., 2010b; Werner & Campbell, 1970). It is also very work intensive and there is little information about recent experiences using this technique. See Harkness, et al. (2010b) for a more detailed explanation of decentering.

- The key advantage of the ASQT approach is standardization of the stimuli across cultures; the main disadvantage of the ASQT approach is that a literal or near close translation may not be culturally suitable and appropriate or may not be possible. For example, the European Social Survey (ESS) had no issue using ASQT to translate, "Do you consider yourself as belong to any particular religion or denomination?" into multiple languages; however, ASQT was not appropriate for, "Do you have difficulty walking several blocks?" Using ASQT for the latter was judged problematic in study countries where neighborhoods are not organized into blocks
- Anchoring vignettes allow researchers to make adjustments when respondents from different cultures, countries, or ethnic groups interpret questions in different ways (King, Murray, Salomon, & Tandon, 2004). Participants are asked to provide assessments both for themselves and for several hypothetical people. Anchoring vignettes assume vignette equivalence (i.e., the hypothetical situations portrayed in the vignettes are viewed equivalent across cultures to be compared) and reporting consistency (i.e., respondents rate their own and hypothetical vignette persons in a consistent way). They can be analyzed by both nonparametric methods and model-based parametric methods. However, the profile of hypothetical persons in the vignettes can potentially affect the adjustments (Grol-Prokopczyk, 2014). Methods have been developed for the evaluation and selection of anchoring vignettes for a diverse range of topics (King & Wand, 2007). See King (n.d.) for more information and resources.

2.1.2 Ask different questions (ADQ). In this approach, researchers ask the most salient questions for each population on a given common construct or conceptual domain. The different questions and, possibly, different indicators used in each location are assumed to tap a construct that is shared across populations. For example, the following questions may all be effective indicators of the concept of intelligence for individual populations. However, characteristics of intelligence may be more or less salient, depending on local context and ADQ may be the best strategy (Harkness, 2008b). See Adaptation for more information.

Is she quick-witted?
Does she give considered responses?
Is she good at knowing whom to ask for help?
Is she good at finding solutions to urgent problems?

- This approach emphasizes the standardization of meanings and strives for functional equivalence.
- The downside of this approach is that the item-by-item analyses across populations are more difficult to justify since the questions are not the same across different groups.

2.1.3 A mixed approach that combines ASQT and ADQ. Many 3MC surveys use a mix of ASQT and ADQ questions.

- Some questions blend a common part (ASQT) and country-specific parts (ADQ). Socio-demographic questions on education, for example, are often asked in terms of a shared question stem (such as “What is the highest level of education you have completed?”), accompanied by local/national categories of educational level or qualification (ADQ). These are then mapped onto an international standard (see Translation: Harmonization).
- For cross-cultural surveys, cultural adaptation of instruments along with translation improves measurement comparability (Georgas, Weiss, van de Vijver, & Saklofske, 2003; Hambleton, Merenda, & Spielberger, 2005). See the Adaptation chapter for more details.

2.2 Weigh the advantages and disadvantages of each approach in terms of the study design (see overview in Appendix A).

2.3 Decide on the most viable approach for the study within a quality framework that addresses survey error related to questionnaire design (see Survey Quality).

2.4 Match the question style (responses) to respondent recall style. For example, incorporate calendar techniques (e.g., event history calendars or life history calendars; see Data Collection: General Considerations) for people who identify time by events (Yount & Gittelsohn, 2008). Of course, this requires researchers with substantive knowledge of each cultural group in the survey.

Lessons learned

2.1 Not all options will be available for every study. The study design, the target population, and the mode required may all impose constraints. For example, if questions from other studies are to be used again ("replicated"), only an ASQT model (perhaps with adaptation) is possible for these questions. The chosen data collection method, the sample design, the fielding schedules, and available funds or stipulations on the use of funds can all limit options (see Study Design and Organizational Structure, Data Collection: General Considerations, Sample Design, and Tenders, Bids, and Contracts).

2.2 Cross-cultural questionnaire design literature can be hard to locate, unclear, or very sparse on details. Even detailed study reports might be clear to people involved in a project but not clear enough for "outside" readers. Detailed and transparent documentation of the questionnaire design process is critical for cross-cultural survey research in order for other data users to understand the data collection procedures in each country and to evaluate the data quality in a comparative manner.

2.3 Researchers are usually most familiar with the ASQT approach, but may not be aware of the limitations and constraints of this approach (Behr, 2010; Harkness, et al., 2010b; Harkness et al., 2003; Harkness, Villar, & Edwards, 2010a). In addition, pressures to replicate questions might over-promote the ASQT approach. Please see Appendix A for the pros and cons of ASQT.

2.4 Comparability or equivalence is sometimes judged on the basis of similar wording across questionnaires. This is, indeed, what is often targeted in ASQT approaches. However, even nominally "accurate" translations do not necessarily produce comparable data (see Translation: Overview). For example, a close translation of the English question "Does he like adventures?" in French is more likely to be understood as "Does he like amorous adventures?" Bilingual or multi-lingual researchers with substantive knowledge of two or more cultures and languages are essential in this approach. In addition, qualitative study and cognitive testing are critical for questionnaire translations. After all, the mutual understanding among the respondents is the goal.

2.5 It is difficult to find examples of surveys with most substantive questions based on an ADQ approach. There are examples of research that analyzes different questions from different studies and takes them to reflect aspects of a given common construct (Van Deth, 1998).

2.6 Change of question formats or adapted questions can radically affect respondents' answers in cross-cultural surveys, such as in the International Social Survey Programme (ISSP) (Smith, 1995) (see also Adaptation).

2.7 Researchers also need to be aware of the negative consequences associated with inappropriate standardization (see Harkness & Behr, 2008, and Lynn, Japec, & Lyberg, 2006).

2.8 Researchers need to be aware of cross-cultural differences in the relevance, saliency, and social desirability of survey questions. For instance, certain questions may not be relevant or salient for a given population and that population may not have the information necessary to answer those questions. In addition, questions considered innocuous in one culture may be threatening or taboo in other cultures. For example, respondents in Islamic countries would find questions about alcohol use or the number of children born to an unmarried respondent offensive (e.g., see Smith, 2002).

2.9 Respondents' social reality and cultural framework shape their perceptions and survey responses in a variety of ways (see Braun & Mohler, 2003, and Yang, et al., 2010).

3. **Establish a lead team or working group responsible for questionnaire design, and appoint a coordinator responsible for organizing scheduling, communication channels and rules, and the design deliverables.**

Rationale

Good questionnaires can rarely be developed by a single person. This is especially true for 3MC research. In accordance with a quality assurance framework for design, a team is needed that provides the spread of knowledge, diverse skills, and cultural backgrounds for which successful comparative design calls (Lyberg & Stukel, 2010).

Procedural Steps

3.1 Decide, as appropriate, on the lingua franca and communication mediums to be used in the overall project and in the work of the questionnaire design team.

3.2 Identify a lead person in the design team who is also responsible for coordinating with other groups in the project (such as the coordinating center, if one exists – see Study Design and Organizational Structure).

3.3 Decide on appropriate communication channels (e.g., in-person and telephone meeting, or video-conferencing, including online meetings). Meet regularly to communicate progress.

3.4 Identify the various skills required in the team.
  3.4.1 These include all the skills needed for questionnaire design in general, including but not limited to 3MC research.
  3.4.2 They also include special expertise or skills relevant for designing a comparative instrument (e.g., understanding design models such as ASQT and ADQ), understanding the cultural impact of conceptual coverage, cultural norms that affect common ground and response processes, response styles, local population structure and needs, etc.).
  3.4.3 Depending on their roles in the team, members may need to be conversant in whatever lingua franca is used in a multilingual project.

3.5 Ensure that the team members recruited are from a reasonable spread of the countries, locations, or cultures participating in the study.

3.6 Ensure that the members recruited for the questionnaire design team have the skills and abilities needed for good questionnaire design. A questionnaire design team should consist of 1) comparativists, including area/cultural specialists, 2) substantive/subject area experts, 3) linguistic experts and 4) survey research experts (Mohler, 2006).
  3.6.1 If the cultural and linguistic experts in the project lack fundamental knowledge in survey research, it is important to provide training to them or include survey methodologists on the team.

3.7 If qualitative components are included, involve an interdisciplinary decision-making team with training in both qualitative and quantitative methods (Massey, 1987).

3.8 Identify the responsibilities of each member at an appropriate level of detail.

3.9 Recruit collaborators and external experts, as necessary and feasible, from the different populations involved. This ensures the availability of expertise on given topics and local knowledge. A drafting team might need specific and short-term input from an expert on a substantive area in the questionnaire. For example, if input on pensions is needed, an expert on the topic may be brought in exclusively for the development of pension-specific questions.

Lessons learned

3.1 In addition to the lead team, each cultural group can benefit from strong input from local participants who are similar to the intended sample population. Ways should be found to have any groups who are participating in the project, but are not directly part of the core development team, to contribute to the development of the questionnaire. This can be a less formal team of local participants who can help guide questionnaire development from the ground up, along the lines of "simultaneous [questionnaire] development" (Harkness, et al., 2010b; Harkness et al., 2003). Another option is for the working group to specify target variables while allowing local participants to specify the particular questions (Granda, et al., 2010). It will be helpful for local participants to be familiar with survey research methods.

3.2 Qualitative methods such as focus groups and cognitive interviews can be used to gain insights into the local community and experiences of the target population, which researchers alone may not be able to recognize or capture. Findings from qualitative methods can be used to inform questionnaire design and subsequent interpretation of quantitative results (Habashi & Worley, 2009).

4. Establish the procedures and protocols for questionnaire development and for testing at different stages in this development.

Rationale

Clear identification of the procedures and the protocols to be followed is essential to inform all those involved and to effectively implement and assess the chosen design process.

While different studies follow different design models (ASQT, ADQ, mixed approaches), this guideline identifies some of the key generic elements to be considered.

Procedural steps

4.1 Establish which design and related procedures are to be used (e.g., ASQT source questionnaire and translation).
4.2 Develop the protocols relevant for the chosen design model and the processes it calls for (e.g., protocol for questionnaire development of a source questionnaire intended for use in multiple locations and cultures/languages).

4.3 Create a schedule and budget for the milestones, deliverables, and procedures involved in the chosen design model. In the ASQT model this would include schedules and a budget for producing draft source questionnaires, review by participating cultures or groups, deadlines for feedback, schedules for pretesting, schedules for language harmonization, schedules for translation (See Translation: Scheduling), and subsequent assessment and pretesting. The participation of team members from all countries throughout the process is essential in ensuring that the questions are developed, translated, and tested appropriately among all target populations.

4.4 Create a framework of quality assurance and quality control to ensure compliance with protocols and the adequacy of outputs (see Survey Quality).

4.5 Create communication channels and encouragements which ensure that participants can and do make feedback on draft designs they are asked to review.

Lessons learned

4.1 Not all participating groups in a project will be confident that their input in the developmental process is (a) valuable in generic terms for the entire project, (b) accurate or justified, and (c) welcomed by perceived leading figures or countries in either the design team or the larger project. It is important to make clear to participating groups that every contribution is valuable. Sharing feedback across the project underscores the value of every contribution and explains to participating groups why their suggestions are or are not incorporated in design modifications.

5. Pretest source and target questionnaires.

Rationale

Questionnaires need to be pretested before they are used. The source questionnaire needs to be assessed for its suitability as a source questionnaire for multiple other versions, rather than as a questionnaire for a single population. Pretesting often relies on expert review, particularly for reviewing the suitability of the source questionnaire in other cultural groups.

The other versions produced—most likely on the basis of translation or translation and adaptation—also need to be pretested for suitability, ideally with every target population in the study. Various qualitative and quantitative approaches can be taken to pretest the target questionnaires in the target population. An often-cited recommendation is "If you do not have the resources to pilot-test your questionnaire, don't do the study." (Sudman & Bradburn, 1982, p. 283).

Procedural steps

(For detailed information about pretesting, see Pretesting.)

Lessons learned

5.1 Pretesting is essential. Even questions previously used in other questionnaires must be tested for their suitability in a new context and for use with new populations.

5.2 Where possible, pretesting of the source questionnaire should be combined with pretesting a spread of other languages representing the diverse target populations in the project (Skevington, 2002).

5.3 Ensuring the quality of questionnaire development prior to pretesting is just as important as pretesting itself. Proper team selection, adequate briefing on requirements and expectations, and good use of documentation will enhance the quality of the questions presented for pretesting so that pretesting serves the monitoring and refining purposes it should have.

5.4 Combine both quantitative and qualitative techniques to evaluate and test questionnaires.

5.4.1 Question design and statistical modeling "should work in tandem for survey research to progress" (Presser, et al., 2004). In other words, when designing questions, consider how they will be used in analysis.

5.5 Even locations sharing the language of the source questionnaire (e.g., the U.S. and the U.K.) need to review the instrument for local suitability (Jowell, 1998).

6. Establish a quality assurance and quality monitoring framework for questionnaire development.

Rationale

Irrespective of the design approach followed to produce a questionnaire, quality standards must be set. These are critical to establishing quality assurance and quality monitoring steps for the process of developing any questionnaire (International Organization for Standardization, 2012).

Procedural steps

6.1 Be cognizant of possible sources of survey error in the questionnaire design phase and develop quality assurance and quality monitoring steps to address these (see Survey Quality). Possible sources of error in this phase include validity and measurement issues (Groves et al., 2009).

6.2 Acquaint question designers with important quality assurance literature on the topic of question design (e.g., on validity, tests of conceptual coverage, response process, sources of measurement error) (Biemer & Lyberg, 2003; Groves et al., 2009).

6.3 For source questionnaires, form a team in each country or location that meets to discuss the development and assessment of the source questionnaire at each phase. The team should have, or should be provided with, the methodological expertise needed for this task.

6.4 Have such teams document and report any queries or problems to the questionnaire drafting group in a timely fashion during the development phases or, as appropriate, report to the coordinating center (Lyberg & Stukel, 2010).

Lessons learned

6.1 Quality assurance and quality monitoring should be addressed early in the design planning process. Thornton, et al., (2010) describe the explicit procedures and protocols he and his research team followed in designing a multinational study with no pre-existing survey items to measure underlying theoretical concepts of the research question. Because error due to measurement and validity of brand-new measures were of particular concern, all survey question writing (and subsequent translation) was done as a team through a series of weekly meetings, beginning at the inception of the research project. All meetings involved collaborators, experts in a wide variety of relevant disciplines, from each study country. Meeting discussions, item wording decisions, and inconsistent field results pointing to measurement issues were carefully documented and necessary deviations from between-country comparability were detailed in the dataset codebook for future users (Thornton et al., 2010). See also de Jong & Young-DeMarco (2017) for a similar discussion of protocols followed for a cross-national comparative survey in the Middle East.

6.2 Variations in country-level assessment experience, research traditions, and methodological rigor regarding question design need to be thoroughly investigated and understood when setting quality standards. Some locations or countries will need more assistance than others in understanding the relevance of some requirements. They may also need guidance on how products can be assessed in terms of these requirements.

6.3 Some entity, such as a questionnaire drafting group coordinator or a coordinating center, must be appointed to lead on these matters.

6.4 Through their knowledge of their own location and culture, local level collaborators and team members may well provide insights that other team members lack, even if quite experienced in questionnaire design.

7. Develop qualitative and quantitative protocols and procedures for assessing the quality of questions across survey implementations.

Rationale

Identifying standards to be met and establishing the criteria required to meet them, as well as agreeing on the good/best practice procedures to follow, are basic to undertaking quality assurance and quality monitoring.

Procedural steps

7.1 Determine appropriate methods to assess the quality of questions. Consider question standards and survey determinants (e.g., funding and resources), as well as the model of design chosen for the topic (Groves et al., 2009).

7.2 Include qualitative and quantitative methods of assessment (see the Pretesting chapter for a detailed description of assessment methods).

7.2.1 Qualitative options include:
- Various pretesting techniques, such as focus groups and cognitive interviews (see Pretesting).
- Expert appraisals by such groups as target population members, substantive experts, question design experts, or translators.
- Debriefings from any testing (interviewers and respondents).

7.2.2 Quantitative methods of assessment include pilot studies (Biemer & Lyberg, 2003; van de Vijver & Leung, 1997):
- Reliability (e.g., Cronbach's alpha) and validity.
- Exploratory and confirmatory analyses such as variance analysis, factor analysis, multi-trait multi-method, item

response theory, latent class analysis, differential item functioning, or stand-alone or embedded experiments.
- These methods require planning as they require specific types of data.

7.3 When possible, use wording experiments to decide between different candidate question wordings (Fowler, 2004; Moore, Pascale, Doyle, Chan, & Klein Griffiths, 2004). However, little effort has been devoted to comparative experimental study on survey design and translation (for examples, see Harkness, Villar, Kephart, Schoua-Glusberg, & Behr, 2009).

7.4 Consider using advance translations or translatability assessment as part of questionnaire design to minimize later translation problems.

Lessons learned

7.1 Both qualitative and quantitative methods of assessment are necessary. Reliance on one without the other is not advised.

7.2 Do not use pretesting as the main tool for question refinement. Make the questions as well designed as possible before pretesting so that pretesting can be used to find problems not identifiable through other refinement procedures.

7.3 Different disciplines favor and can use different developmental and testing procedures, partly because of their different typical design formats. Social science surveys, for example, often have only one or two questions on a particular domain; psychological and educational scales, on the other hand, might have more than twenty questions on one domain.

8. Develop a documentation scheme for questionnaire design decisions, design implementation, and quality assurance protocols.

Rationale

Documentation aids in producing the questionnaire and can be a tool for quality assurance and monitoring. As indicated in Survey Quality, continual measurement and documentation of the quality targeted and achieved is necessary to identify quality problems. Even if sources of error are not recognized until later, documentation can be used to inform improved designs for future studies.

Procedural steps

8.1 Design the documentation process before question development begins and document question design from the start. This ensures that all decisions are captured and that action can be taken in a timely fashion.

8.2 Standardize documentation requirements and formats across all locations or countries involved in question development. This facilitates feedback in an ASQT model and comparable development in an ADQ model.

8.3 Create flexible documentation templates that allow categories to be added if unforeseen issues arise.

8.4 Create a clear and concise description of the questionnaire design procedures which is user-oriented and user-friendly. Include:
 8.4.1 Conceptualization from concept to questions.
 8.4.2 Operationalization (approach, mode, development across versions, adaptation agreements, annotations, (shared) language harmonization, origin of questions whether new, replicated, adopted, or adapted).
 8.4.3 An analysis plan.

8.5 Record the development of indicators and questions from start to finish (e.g., any modifications made to questions at different stages and why).

8.6 Version control procedures are necessary whenever a source questionnaire is modified across time.
 8.6.1 A version of the source questionnaire will serve as the gold standard, or source version 1. Document any changes made to it over time.

Lessons learned

8.1 Documentation must accompany questionnaire design since it will be used to detect problems in time to address them.

8.2 If documentation is left to the end of questionnaire design (or even later), details will be forgotten and intervention will not be possible. Study monitoring questionnaires for the ISSP (completed well after question design and translation have been completed) sometimes contain documentation on translation challenges for two or three phrases. The templates used in recent German ISSP translation discussions note a myriad of challenges (Behr, 2010).

8.3 Any changes countries make to their design protocols and procedures and any reservations they have about development must be carefully documented. If these are made available in a timely fashion to either the questionnaire drafting coordinator or, as appropriate, the coordinating center, problems can be addressed. For example, feedback to questionnaire drafting groups from countries participating in the ISSP and ESS studies sometimes lead to changes in draft versions of source questions.

8.4 At a later stage, documentation might be helpful in understanding potential differences in the data, either over the course of the study (within a country) or across variables (between countries).

8.5 Providing tools to make the job easier encourages people to engage in the task and ensures better documentation.

8.6 Demonstrating the importance of documentation motivates people to engage in it. Even simple things can help convince and motivate ─ for example, showing how a template can help check for flipped order of answer categories across a range of questions.

Appendix A

Some advantages and constraints on different approaches to question design

Approach	Advantages	Constraints
Ask the same question and translate (ASQT)	If successful, questions and item scales can be compared one-by-one across data sets and thus permit the most sophisticated analyses based on what is sometimes called full scalar equivalence (see van de Vijver & Leung, 1997).	Developing for an ASQT questionnaire may result in reduced specificity of questions used and a resultant loss of saliency and fine-grained information.
	ASQT is the least complicated approach to organize and implement. This is not to suggest it does not involve considerable effort as reflected in this chapter and Translation: Overview.	Conceptual coverage for all or some of the populations in the study may thus be reduced and not comparable across populations.
	Researchers engaged in it and clients requesting or expecting it may feel more like they are on familiar territory with this model than with others.	At the translation stage, those implementing the ASQT may have inappropriate goals for the translation product and produce poor target language versions.
	ASQT potentially permits replication of existing questions—provided their basic suitability for translation and fielding with the target populations is ensured.	Replicated questions encourage close translation and may not be optimal for one or more target populations.
		ASQT does not work well at all for some kinds of questions (e.g., background variables such as education).

Approach	Advantages	Constraints
		ASQT and adapt approaches call for expertise in question development and translation in areas still requiring basic research and/or training.
Decentering	Allows two questionnaires to be developed in collaboration and creates the potential for full scalar equivalence (see van de Vijver & Leung, 1997).	May result in questions with low saliency for either culture since anything that constitutes a problem in the course of development is removed or altered. This would be an obstacle to full scalar equivalence.
	Avoids the situation where the needs of one questionnaire and language/culture dominate.	Decentering is not viable for projects involving more than a handful of languages.
	Can be useful in developing comparable questions for very disparate cultures.	Decentering is very work intensive and there is little information about recent experiences using this technique.
Ask different questions approaches	Avoids the need to base questionnaires for various cultures and languages on translation of a source questionnaire.	Little detailed information is available about recent projects adopting an ADQ approach. Researchers have few guidelines about how to develop the quality assurance and quality control steps needed.
	Researchers can select the indicators and questions considered most salient for a given population provided these produce data which can still be compared across populations.	If different populations are only asked questions developed for them, item-by-item analyses across populations are more difficult to justify.

Approach	Advantages	Constraints
	It is easier for a group joining an ADQ-based study after other groups have developed and fielded their questionnaires to produce a suitable questionnaire for their context than it is for researchers joining an ASQT project after the source questionnaire has been finalized.	Most researchers and clients are unfamiliar with ADQ approaches.
Mixed approaches combining ASQT and ADQ components	These can combine the advantages of ASQT and ADQ.	They increase the number and kind of procedural steps to be implemented and assessed.
		They call for expertise in areas still requiring basic methodological research.

References

Baumgartner, H., & Steenkamp, J. B. (2001). Response style in marketing research: A cross-national investigation. *Journal of Marketing Research, 38*(2)*,* 143–156.

Behr, D. (2010). *Translationswissenschaft und international vergleichende Umfrageforschung: Qualitätssicherung bei Fragebogenübersetzungen als Gegenstand einer Prozessanalyse* [Translation research and cross-national survey research: quality assurance in questionnaire translation from the perspective of translation process research]. Mannheim, Germany: GESIS.

Berry, J. W. (1969). On cross-cultural comparability. *International Journal of Psychology*, *4*(2), 119-128.

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

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*. San Francisco, CA: Jossey-Bass.

Braun, M. & Mohler, P. P. (2003), Background variables. In J. A. Harkness, F. J. R. van de Vijver, & P. P. Mohler (Eds*.), Cross-Cultural Survey Methods* (pp. 101-115). Hoboken, NJ: John Wiley & Sons.

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

De Jong, J., & Young-DeMarco, L. (2017). Best practices: Lessons from a Middle East survey research program. In M. Gelfand & M. Moaddel (Eds.), *The Arab spring and changes in values and political actions in the Middle East: Explorations of visions and perspective.* Oxford, U. K.: Oxford University Press.

de Leeuw, E.D. (2008). Choosing the method of data collection. In E. D. de Leeuw, J. J. Hox, & D. A. Dillman (Eds.), *International handbook of survey methodology* (pp. 113–135). New York/London: Lawrence Erlbaum Associates.

Dillman, D. A., Smyth, J. D., & Christian, L. M. (2009). *Internet, mail, and mixed-mode surveys: The tailored design method* (3rd ed.). New York, NY: John Wiley & Sons.

Fink, A. (2003). *How to ask survey questions* (2nd ed.). Thousand Oaks, CA: Sage Publications.

Fowler, F. J. Jr. (1995). *Improving survey questions: Design and evaluation* (Vol. 38, Applied social research methods series). Thousand Oaks, CA: Sage Publications.

Fowler, F. J. (2004), The case for more split-sample experiments in developing survey instruments. In S. Presser, J. M. Rothgeb, M. P. Couper, J. T. Lessler, E. Martin, J. Martin, et al. (Eds.), *Methods for Testing and Evaluating Survey Questionnaire* (pp. 453-172). Hoboken, NJ: John Wiley & Sons.

Georgas, J., Weiss, L. G., van de Vijver, F. J. R., & Saklofske, D. H. (Eds.). (2003), *Culture and Children's Intelligence: Cross-Cultural Analysis of the WISC-III*. Amsterdam, Boston: Academic Press.

Granda, P., Wolf, C., & Hadorn, R. (2010). Harmonizing survey data. In J. A. Harkness, M. Braun, B. Edwards, T. Johnson, L. E. Lyberg, P. Ph. Mohler, et al. (Eds.), *Survey methods in multinational, multicultural and multiregional contexts* (pp. 315-334). Hoboken, NJ: John Wiley & Sons.

Grol-Prokopczyk, H. (2014). Age and sex effects in anchoring vignette studies: methodological and empirical contributions. *Survey Research Methods*, *8*(1), 1-17.

Groves, R. M. (1989). *Survey errors and survey costs*. New York, NY: John Wiley & Sons.

Groves, R. M., Fowler, F. J. Jr., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2009). *Survey methodology* (2nd ed.). Hoboken, NJ: John Wiley & Sons.

Habashi, J., & Worley, J. (2009). Child geopolitical agency: A mixed methods case study. *Journal of Mixed Methods Research*, *3*, 42-64. doi:10.1177/1558689808326120

Hambleton, R. K., Merenda, P. F., & Spielberger, Ñ D. (Eds.). (2005), *Adapting educational and psychological tests for cross-cultural assessment*. Mahwah, N.J.: L. Erlbaum Associates.

Harkness, J. A. (2008b). Comparative survey research: Goals and challenges. In E. D. de Leeuw, J. J. Hox & D. A. Dillman (Eds.), *International handbook of survey methodolog*y (pp. 56-77). New York, NY/London: Lawrence Erlbaum Associates.

Harkness, J., & Behr, D. (2008). *How instrument quality affects translation: Insights from a film of a team translation review*. International Sociological Association Research Committee 33, Naples, Italy, September.

Harkness, J. A., Edwards, B., Hansen, S. E., Miller, D. R., & Villar, A. (2010b). Designing questionnaires for multipopulation research. 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, multicultural, and multiregional contexts* (pp. 33–58). Hoboken, NJ: John Wiley & Sons.

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. M. Mohler (Eds.), *Cross-cultural survey methods* (pp. 19-34). Hoboken, NJ: John Wiley & Sons.

Harkness, J. A., Villar, A., & Edwards, B. (2010). Translation, adaptation, and design. In J. A. Harkness, M. Braun, B. Edwards, T. P. Johnson, L. E. Lyberg, P. Ph. Mohler, et al. (Eds.), *Survey methods in multinational, multicultural and multiregional contexts* (pp. 117-140). Hoboken, NJ: John Wiley & Sons.

Harkness J. A., Villar, A., Kephart, K., Schoua-Glusberg, A., & Behr, D. (2009). *Survey translation evaluation: Back translation versus expert review*. Paper presented at the American Association for Public Opinion Research meeting (AAPOR), Hollywood, Florida.

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

International Organization for Standardization (ISO). (2012). *Market, opinion and social research — Vocabulary and service requirements* (ISO 20252). Retrieved March 18, 2016, from https://www.iso.org/obp/ui/#iso:std:iso:20252:ed-2:v1:en

Johnson, T. P. (1998a). Approaches to equivalence in cross- cultural and cross-national survey research. *ZUMA-Nachrichten Spezial*, (3), 2–40. http://www.gesis.org/fileadmin/upload/forschung/publikationen/zeitschriften/zuma_nachrichten_spezial/znspezial3.pdf

Johnson, T. P. (1998b). Approaches to establishing equivalence in cross-cultural and cross-national survey research. In Janet Harkness (Ed.), *Cross-cultural survey equivalence*. Mannheim, Germany: ZUMA.

Johnson, T. P., & van de Vijver, F. (2003). Social desirability in cross-cultural research. In J. A. Harkness, F. van de Vijver, & P. Ph. Mohler (Eds.), *Cross-cultural survey methods* (pp.195–204). Hoboken, NJ: John Wiley & Sons.

Jowell, R. (1998). How comparative is comparative research? *American Behavioral Scientist*, *42*(2), 168-177.

King, G., Murray, C. J. L., Salomon, J. A., & Tandon, A. (2004). Enhancing the validity and cross-cultural comparability of measurement in survey research. *American Political Science Review*, *98*(February), 191-207.

King, G., & Wand, J. (2007). Comparing incomparable survey responses: Evaluating and selecting anchoring vignettes. *Political Analysis*, *15*(1), 44-66. Oxford: Oxford University Press.

Kohn, M. L. (1969). *A study of values*. Homewood, IL: Dorsey.

Lara, D., Strickler, J., Olavarrieta, C. D., & Ellertson, C. (2004). Measuring induced abortion in Mexico A comparison of four methodologies. Sociological Methods & Research, 32(4), 529-558.

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, et al. (Eds.), *Survey methods in multinational, multicultural, and multiregional contexts* (pp. 227–250). Hoboken, NJ: John Wiley & Sons.

Lynn, P., Japec, L., & Lyberg, L. (2006). What's so special about cross-national surveys? in J. A. Harkness (Ed.), *Conducting Cross-National and Cross-Cultural Surveys*, pp. 7-20. ZUMA-Nachrichten Spezial, 12, Mannheim: ZUMA.

Massey, D. S. (1987). The ethnosurvey in theory and practice. *International migration review*, *21*(4), 1498-1522.

Mohler, P. Ph. (2006). Sampling from a universe of items and the de-machiavellization of questionnaire design. In M. Braun & P. Ph. Mohler (Eds.), *Beyond the horizon of measurement - Festschrift in honor of Ingwer Borg* (ZUMA-Nachrichten Spezial,10). Mannheim, Germany: ZUMA. Retrieved April 28, 2010, from http://www.gesis.org/fileadmin/upload/forschung/publikationen/zeitschriften/zuma_nachrichten_spezial/znspezial10.pdf

Mohler, P. P., & Johnson, T. P. (2010). Equivalence, comparability, and methodological progress. survey methods in multinational, multiregional, and multicultural contexts. In J. A. Harkness, M. Braun, B. Edwards, T. P. Johnson, L. E. Lyberg, P. Ph. Mohler, et al. (Eds.), *Survey methods in multinational, multicultural and multiregional contexts* (pp. 17–29). Hoboken, NJ: John Wiley & Sons.

Moore, J., Pascale, J., Doyle, P., Chan, A., & Klein Griffiths, J. (2004). Using field experiments to improve instrument design: The SIPP Methods Panel Project. In S. Presser, et al. (Eds.), *Methods for Testing and Evaluating Survey Questionnaires*. Hoboken, NJ: John Wiley & Sons.

Payne, S. L. (1980). *The art of asking questions.* Princeton, NJ: Princeton University Press.

Presser, S., Rothgeb, J. M., Couper, M. P., Lessler, J. Ò., Martin, E. A., Martin, J., et al. (Eds.), (2004). *Methods for testing and evaluating survey questionnaires*. Hoboken, NJ: John Wiley & Sons.

Schwarz, N. (1996). *Cognition and communication: Judgmental biases, research methods, and the logic of conversation.* Mahwah, NJ: Lawrence Erlbaum Associates.

Schwarz, N, Oyserman, D., & Petycheva, E. (2010). Cognition, communication, and culture: Implications for the survey response process. 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, multicultural, and multiregional contexts* (pp. 177–190). Hoboken, NJ: John Wiley & Sons.

Skevington, S. M. (2002). Advancing cross-cultural research on quality of life: Observations drawn from the WHOQOL development. *Quality of Life Research*, *11*, 135-144.

Smith, E. E. (1995). Concepts and categorization. In E. E. Smith & D. N. Osherson (Eds.), *Thinking (An invitation to cognitive science)*, Vol. 3 (pp. 3-34). Cambridge, MA: MIT Press.

Smith, T. W. (2002). *CNR 22 Constructing cross-national surveys*. Cross-National Report. Chicago: NORC.

Sudman, S. & Bradburn, N. M. (1982), *Asking questions: A practical guide to questionnaire design.* San Francisco, CA: Jossey-Bass.

Thornton, A., Achen, A., Barber, J. S., Binstock, G., Garrison, W. M., Ghimire, D. J., ... & King, K. (2010). Creating questions and protocols for an international study of ideas about development and family life. In J. A. Harkness, M. Braun, B. Edwards, T. P. Johnson, L. E. Lyberg, P. Ph. Mohler, et al. (Eds.), *Survey methods in multinational, multiregional, and multicultural contexts* (pp. 59-74). Hoboken, NJ: John Wiley & Sons.

Tourangeau, R., & Rasinski, K. (1988). Cognitive processes underlying context effects in attitude measurement. *Psychological Bulletin*, *103* (May), 299-314.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response.* Cambridge: Cambridge University Press.

Vaerenbergh, Y. V., & Thomas, T. D. (2012). Response styles in survey research: A literature review of antecedents, consequences, and remedies. *International Journal of Public Opinion Research*, *25*(2), 195–217.

van de Vijver, F. J. R., & Leung, K. (1997). *Methods and data analysis for cross-cultural research*. Thousand Oaks, CA: Sage Publications.

Van Deth, J. W. (1998). *Comparative politics: The problem of equivalence.* London: Routledge.

Verba, S. (1969), The uses of survey research in the study of comparative politics: Issues and strategies. In S. Rokkan, S. Verba, J. Viet, & E. Almasy (Eds.), *Comparative survey analysis*, pp. 56-106. The Hague: Mouton.

Werner, O. & Campbell, D. T. (1970). Translating, working through interpreters, and the problem of decentering. In R. Naroll & R. Cohen (Eds.), *A handbook of method in cultural anthropology* (pp. 398–420). New York, NY: American Museum of Natural History.

Willimack, D.K., Lyberg, L. E., Martin, J., Japec, L., & Whitridge, P. (2004). Evolution and adaptation of questionnaire development, evaluation, and testing methods for establishment surveys. In S. Presser, J. M. Rothger, M. P. Couper, J. T. Lessler, E. Martin, J. Martin, & E. Singer (Eds.), *Methods for testing and evaluating survey questionnaires* (pp. 385–408). Hoboken, NJ: John Wiley & Sons.

Yang, Y., Harkness, J. A., Ching, T-Y., & Villar, A. (2010). Response styles and culture. In J. A. Harkness, M. Braun, B. Edwards, T. P. Johnson, L. E. Lyberg, P. Ph. Mohler, et al. (Eds.), *Survey methods in multinational, multicultural, and multiregional contexts* (pp. 203–226). Hoboken, NJ: John Wiley & Sons.

Yount, K. M. & Gittelsohn, J. (2008). Comparing reports of health-seeking behavior from the integrated illness history and a standard child morbidity survey. *Journal of Mixed Methods Research*, *2*, 23-62. doi:10.1177/1558689807309083

Further Reading

Arnon, S. & Reichel, N. (2009). Closed and open-ended question tools in a telephone survey about 'the good teacher': An example of a mixed method study. *Journal of Mixed Methods Research*, *3*(2), 172-196.

Battiste, M. (2008). Research ethics for protecting indigenous knowledge and heritage: institutional and research responsibilities. In N. K. Denzin, Y. Lincoln, & L. T. Smith (Eds.), *Handbook of critical and indigenous methodologies* (pp. 497-510). Los Angeles, CA: Sage.

Bernard, H. R. (1998). *Handbook of methods in cultural anthropology*. Walnut Creek, CA: AltaMira Press.

Bernard, H. R. (2006). Research *methods in anthropology: Qualitative and quantitative approaches* (4th ed.). Oxford: AltaMira Press.

Betancourt, H., Flynn, P. M., Riggs, M., & Garberoglio, C. (2010). A cultural research approach to instrument development: The case of breast and cervical cancer screening among Latino and Anglo women. *Health Education Research*, *25*(6), 991-1007.

Camfield, L., Crivello, G., & Woodhead, M. (2009). Wellbeing research in developing countries: Reviewing the role of qualitative methods. *Social Indicators Research*, *90*(1), 5-31.

Cheung, F. M., van de Vijver, F. J. R., & Leong, F. T. L. (2011). Toward a new approach to the study of personality in culture. *American Psychologist*, *56*(7), 593-603. doi:10.1037/a0022389

Creswell, J. W. & Plano Clark, V. L. (2011). *Designing and conducting mixed methods research* (2nd ed.). Thousand Oaks, CA: Sage.

Hantrais, L. (2009). *International comparative research: Theory, methods and practice*. Hampshire, UK: Palgrave and MacMillan.

Hitchcock, J. H., Sarkar, S., Nastasi, B. K., Burkholder, G., Varjas, K., & Jayasena, A. (2006). Validating culture and gender-specific constructs: A mixed-method approach to advance assessment procedures in cross-cultural settings. *Journal of Applied School Psychology*, *22*(2), 13-33.

Ingersoll-Dayton, B. (2011). The development of culturally-sensitive measures for research on ageing. *Ageing and Society*, *31*, 355–370. doi:10.1017/S0144686X10000917

King, G. (n.d.). The anchoring vignettes website. Retrieved July 28, 2012, from http://gking.harvard.edu/vign/

O'Cathain, A., Murphy, E., & Nicholl, J. (2007). Why, and how, mixed methods research is undertaken in health services research in England: A mixed methods study. *BMC Health Services Research*, *7*. doi:10.1186/1472-6963-7-85.

Revilla, M., Zavala, D., & Saris, W. (2016). Creating a good question: how to use cumulative experience. In C. Wolf, D. Joye, T.W. Smith, & Y. Fu (Eds.), *The SAGE Handbook of Survey Methodology*. Thousand Oaks, CA: Sage.

Saris, W. E. & Gallhofer, I. N. (2007). *Design, Evaluation, and Analysis of Questionnaires for* Survey *Research*. Hoboken, NJ: Wiley.

Smyth, J. D. (2016). Designing questions and questionnaires. In C. Wolf, D. Joye, T.W. Smith, & Y. Fu (Eds.), *The SAGE Handbook of Survey Methodology*. Thousand Oaks, CA: Sage.

Thomas, D. C., Stahl, G., Ravlin, E. C., Poelmans, S., Pekerti, A., Maznevski, M., Lazarova, M. B., et al. (2012). Development of the cultural intelligence assessment. *Advances in Global Leadership*, *7*, 155–178.

Varjasa, K., Nastasi, B. K., Moore, R. B., & Jayasena, A. (2005). Using ethnographic methods for development of culture-specific interventions. *Journal of School Psychology*, *43*(3), 241–258.

Winchester, H. P. M. (1999). Interviews and questionnaires as mixed methods in population geography: The case of lone fathers in Newcastle, Australia. *The Professional Geographer*, *51*(1), 60-67.

Instrument Technical Design

Sue Ellen Hansen, Hyun Jung Lee, Yu-chieh (Jay) Lin, and Alex McMillan, 2016

Introduction

The technical design and implementation of a given survey instrument can be viewed separately from questionnaire design (see Questionnaire Design). Instrument technical design focuses less on questionnaire content and more on the design of the actual survey instrument that delivers the questionnaire content. In this sense, technical design includes the format, layout, and other visual aspects of the presentation or context of survey questions. In some instances, survey design, questionnaire design, and technical design overlap. Mode decisions, for example, may shape the technical format of questions as well as their wording.

These guidelines will use the more general terms "survey instrument" or "instrument" when describing procedures or features that apply to the technical design of both paper and computerized instruments, and the term "application"—which suggests the need for at least some programming—when discussing procedures for development of computerized instruments. When there is a need to distinguish between types of computerized instruments, such as computer-assisted (computerized, but not necessarily accessed via the Internet) and Web instruments, reference will be made to the mode-specific type of computerized survey.

Study design decisions related to mode have an impact on instrument technical design requirements (see Study Design and Organizational Structure and Data Collection Implementation: General Considerations). Such decisions include whether the survey is to be self-administered or interviewer-administered and whether it is to be administered on paper or computerized. If the survey is self-administered, a decision must be made about whether it should be a paper or a computerized survey. For a computerized survey, depending on whether it is a computer-assisted self-interviewing (CASI) instrument or a Web instrument, there may be effects on the programming costs and the computer user interface --that is, what respondents see on the computer screen and how the computer interacts with the respondent.

If the survey is interviewer-administered, decisions may have to be made about whether the instrument should be computerized or paper and whether it should be in person or by telephone. There may be technical design considerations associated with each of those decisions, as discussed below.

If the survey is to be administered on the respondent's personal device, the design and layout must adapt to the various possible formats. Web instruments need to be developed with the assumption that respondents may complete them on computers, smartphones, or tablets.

Study design also involves decisions about data output, coding, and data documentation (see Data Processing and Statistical Adjustment and Data Dissemination). Thus, design decisions may have an impact on technical instrument design, primarily affecting survey implementation in three ways:

1. How easy it is for an interviewer or a respondent to use the survey instrument and to provide appropriate responses (the "usability" of the instrument). This can help minimize user burden.
2. How easy it is to program a computerized instrument and to test it.
3. How easy it is to code, output, analyze, and document survey data.

An instrument's technical design can impact measurement error, including error resulting from cognitive processing, context effects, and interviewer effects. In the case of multinational, multicultural, or multiregional surveys, which we refer to as “3MC” surveys, problems in each of the different technical implementations of survey instruments may lead to different errors. For instance, local implementations could increase interviewer or respondent burden that will lead to cognitive processing errors or even terminated interviews. Poor design of survey instruments may also increase nonresponse error at the levels of the household or respondent (unit nonresponse) or the survey question (item nonresponse).

These guidelines are intended to help researchers at coordinating centers and individual survey organizations of 3MC surveys understand instrument technical design requirements and how to approach creating instrument technical design specifications, whether at the centralized level, the local level, or both. Study design may dictate how much is specified at the central level and how much is left to local survey organizations. While there may be flexibility in this regard, it is important that technical design across local surveys leads to survey data that can be compared across cultures and that does not contribute to measurement error. For example, question labels should be consistent across survey implementations. Differences across cultures may lead to adaptations in technical design across surveys. In such cases, it is important to document the reasons for adaptation.

In general, the layout in the source questionnaire should be preserved in subsequent translated versions. That is, the translated version and the original should look exactly the same except for the words. Some examples from previous rounds in the European Social Survey (ESS) where there were differences in layout between the translated version and the source version include: (a) show cards containing the start of the response sentence when the original did not, or vice versa; (b) show cards putting the answer codes in boxes, omitting the numbering of the categories, or drawing arrows to indicate the end points, where that was not the case in the original; and (c) survey items that were formatted as single questions, each with their own answer scale, rather than formatted as batteries of items (Dorer, 2012). All of these types of deviations can contribute to measurement error. Further examples of areas to mitigate differences in instrument design between the source and target questionnaires are detailed in the guidelines below.

Guidelines

Goal: To minimize measurement error, nonresponse error, and respondent and interviewer burden due to technical instrument design, and thus maximize the amount of valid and reliable information obtained within an allotted budget and time and at the specified level of precision.

1. **Ensure that technical instrument design is appropriate to the method of administration and the target population.**

   Rationale

   The design requirements for self-administered surveys differ from the design requirements for interviewer-administered surveys. Self-administered surveys have no interviewer to help repair misunderstandings and there is limited opportunity to "train" respondents on how to respond. Computerized instruments, which involve human-computer interaction, call for design features that facilitate such interaction.

   The design requirements for computerized instruments also differ from the design requirements for paper instruments. Interface design rules (see Guideline 4) and quality assurance procedures (see Guideline 5) for self-administered, interviewer-administered, paper, or computerized surveys should be developed in advance, and implemented and documented (see Guideline 6) throughout the data collection process.

   The characteristics of the target population (education, survey experience, literacy, computer literacy, etc.) should influence instrument design decisions Self-administered surveys are useful only if administered to populations with high literacy rates; computerized surveys require target populations with familiarity with computers, or situations in which data collection can be facilitated by trained interviewers. Technical instrument design specifications should include as many culture-specific (see Guideline 2) and language-specific (see Guideline 3) guidelines as possible. For mode selection and its specific design considerations, please see Study Design and Organizational Structure.

Procedural steps

1.1 Determine whether to develop an interviewer- or self-administered instrument and whether to use a paper or computerized instrument. Some points to consider from a technical design standpoint are:

1.1.1 Interviewer- versus self-administered instrument:

- Self-administered instruments including CASI, audio computer-assisted self-interview (A-CASI), video-computer-assisted self-interview (video-A-CASI), mail, or Web may lead to better data quality for surveys with extremely sensitive questions (drug abuse or sexually deviant behavior, for example) (Tourangeau & Yan, 2007). However, there can be cross-cultural differences. See Data Collection: Face-to-Face Surveys and Study Design and Organizational Structure for more in-depth discussion.
- Self-administered instruments should make it easy for respondents to recognize instructions (such as "Select one"), and to read questions, navigate correctly through the instrument, and enter responses (Dillman, Smyth, & Christian, 2009; Dillman, Gertseva, & Mahon-Haft, 2005). Instructions should appear where they are needed, such as "Start here" before the first question, response entry instructions (e.g., "Tick all that apply") after the question text, and recording responses should be displayed in the order of their likely occurrence. In addition, instructions to skip questions should be avoided or used sparingly in paper self-administered instruments because they can lead to response errors.
- Self-administered components can be combined with interviewer-assisted components of surveys. An interviewer-administered instrument would be better when there is a need to explain concepts and probe responses or when sections of the interview are sensitive (see Guideline 7).
- Interviewer-administered instruments make it easy to perform required tasks in the order in which they are expected to be performed. For example, interviewers' instructions such as referring to show cards or other aids, reading questions, providing definitions, probing responses, and recording responses should be displayed in the order of respondents' likely occurrence. This is true in both paper and computer-assisted instruments.
- Interviewer administered computerized instruments may lead to higher data quality in long and complex surveys (for example, providing consistency checks or preloaded information throughout the whole instrument) or those with

embedded experiments (for example, randomizing the order of questions or response options).
- Whether interviewer- or self-administered, the instrument technical design should help to minimize the burden placed on interviewers and respondents, which increases as instruments increase in length and complexity.

1.1.2 Paper or computerized instrument:
- Paper instruments may be less costly to develop, but entail additional data entry costs after data collection, and may affect the timeliness of data dissemination (see Data Dissemination).
- Computer-assisted and Web instruments require programming, but Web surveys generally are less costly because of lack of interviewer costs, and don't necessarily require professional programmers for basic programming. On the other hand, if not programmed well, they may introduce higher costs during data processing.
- Some countries or regions may not have the professional expertise in place to do computerized surveys. There can be infrastructural constraints in some contexts that make it difficult to collect data with telephone or Web survey instruments (e.g., the lack of sufficient telephone or Internet penetration, or the lack of an adequate frame) (see Sample Design).
- Paper instruments should be less complex than CASI and Web instruments in order to minimize respondent burden, but still allow for embedded experiments.
- Web surveys should be designed to be viewed on phones, tablets, and any other mobile devices in addition to the tradition computer components.

1.2 Determine whether there are additional design considerations related to characteristics of members of the target population, such as children, men, the elderly, or the visually or hearing impaired (de Leeuw, Hox, & Kef, 2003). Ensure that all such considerations are reflected in the technical specifications for the survey instrument (see Guideline 2).

1.2.1 Computerized instruments with images to show response options or color-coded keyboards to enter response options can be alternative design solutions for populations with low rates of literacy (see Appendix F) or computer usage experience (see Appendix G).

1.2.2 Instrument designs for interviewing multiple people within the same household and using the same instrument may need a customized interface with specific instructions to accommodate the flow of the interview for both computerized and paper instruments (see Appendix I).

Lessons learned

1.1 The use of survey computer assisted methods can help camouflage complexity and facilitate the tailoring of instruments to special populations. For example, de Leeuw, et al. (2003) describe the results from a number of Dutch surveys of special populations using computer-assisted interviewing and self-administered components, in which instrument design and administration were tailored to target population needs. For example, a simple but attractive screen layout was used to survey grade school children. In addition, students only needed to use simple keystrokes to answer questions and could stop temporarily when they felt tired. As a result, item nonresponse was reduced compared to a paper questionnaire. They concluded that well-designed computer-assisted instruments both improve the quality of data and minimize the burden experienced by respondents and interviewers.

1.2 Use of computerized instruments is possible even with low literacy rates. For example, Bhatnagar, Brown, Saravanamurthy, Kumar, and Detels (2013) describe a study of poorly educated men and women in rural South India that experimented with an A-CASI instrument and color-coded response options. Although only 10% of participants had ever used a computer before, 80% stated that the instrument was user-friendly and felt comfortable responding to sensitive questions.

1.3 Study design should consider the potential measurement effects that may arise from differences in methods of survey administration. A review of paradata from the ESS and the International Social Survey Programme (ISSP) revealed some differences in results across countries between those that implemented paper self-administered surveys by mail and those that used interviewer-assisted self-administered surveys or face-to-face surveys.

2. Develop complete technical instrument design specifications for the survey instrument, specifying culture-specific guidelines as necessary.

Rationale

Technical instrument design specifications guide formatting or programming of the survey instrument or application. They ensure design consistency across culture-specific instruments (to the extent possible) and facilitate post-production data processing, harmonization, documentation, and analysis (see Data Processing and Statistical Adjustment and Data Harmonization). The following should be taken into consideration:

- The formatting of information and areas for recording responses.
- The formatting of specific text elements, such as question text, response scales, and respondent or interviewer instructions.
- The formatting of specific question and response types.
- The linking of survey instrument information and variables in a dataset, and documentation of the instrument and dataset.
- Rules for the use of numbers, color, graphics, images, maps, and icons.
- Specifications for how question formats may differ across different data collection modes.

A coordinating center's specifications should clearly outline the source questionnaire and its content, provide rules for formatting the survey instrument, and suggest appropriate instrument design adaptation strategies for other cultures. Survey agencies may have to adapt specification rules further to adhere to local standards for design of instruments and staff training and other organizational constraints. Any such adaptations should be documented.

Note that similar guidelines are necessary for a data entry application (see Data Processing and Statistical Adjustment). Generally, this guideline is relevant to formatting of elements in either paper or computerized instruments, although a few may relate to only one or the other. Guideline 4 adds guidelines that are relevant specifically to computerized applications and their interface designs and to self-administered paper instruments.

Procedural steps

2.1 At the beginning of the instrument specifications, provide an overview of the survey instrument, including the order of core chapters and required placement of culture-specific chapters (see an example in Appendix C).Make sure that that formatting adapts for cultural differences (see Adaptation). For example:

2.1.1 Differences in the formatting of information and areas for the recording of responses (Aykin & Milewski, 2005), including:

- Date and time (e.g., 24-hour versus 12-hour clock).
- Calendar, holidays, and start of week.
- Numeric formatting (e.g., thousands, million, and billion, and decimal separators).
- Names and addresses (e.g., last name first or second).
- Telephone numbers (e.g., with or without local prefix).

- Currency and monetary values (e.g., placement of currency symbol and negative sign).
- Sizes and measurement (e.g., metric versus imperial units, Celsius versus Fahrenheit, clothing sizes, etc.).

2.2 Provide rules for the consistent formatting of specific text elements, such as question text, response scales, respondent or interviewer instructions, and so on. These might include, for example (Couper, Beatty, Hansen, Lamias, & Marvin, 2000):
  2.2.1 Display question text more prominently than response options.
  2.2.2 Distinguish interviewer or respondent instructions, for example, in a smaller font of a different color, or italicized in parentheses.
  2.2.3 Place text elements where and in the order they are needed based on interviewer or respondent task demands; for example, in an interviewer-administered instrument, a show card instruction precedes question text and a probe instruction follows it.
  2.2.4 Evenly space response options in a scale, grid, or table, so that they appear of equal weight or prominence.
  2.2.5 Underline question text that should be emphasized.

2.3 Provide rules for the formatting of specific question, response types (for example, opened- versus close-ended), and other information. Also include examples for each rule; these may include:
  2.3.1 Enumerated or fixed choice response options (e.g., 1=Female, 2=Male).
  2.3.2 Tick [Check / Select] all that apply (e.g., additional options like All Above or None should be added for respondents to checked/selected).
  2.3.3 Short or fixed-length text (e.g., the maximum number of words should be listed for respondents to provide answers).
  2.3.4 Open-ended text (e.g., the maximum number of words should be provided as needed).
  2.3.5 Numeric responses (e.g., for computer-assisted instruments, the range check should be provided and built in for quality control).
  2.3.6 Response entry masks (e.g., __/__/____ for dates).
  2.3.7 Multi-part questions and question series; for example:
    - Day / Month / Year (e.g., either numeric or text value examples like 01 or January should be provided).
    - Address / contact information (e.g., instruments should list address info to levels like country, state, county, city, street, zip code, etc.).
    - Demographics question sets.

- Amount-per-unit (e.g., income per day / week / month / year).

2.3.8 Randomly ordered questions, response options, or sections.

2.3.9 Response scales.
- Fully-labeled scale.
- Partially-labeled scale.
- Roster or grid. Rosters are tables used to collect various information in columns about entities in rows. For example gender and age (columns) about persons in a household (rows). Grids are often used for scale ratings (columns) on a number of items (rows).

2.3.10 Text fills (variable question text); for example, question text may vary based on size of household—"you" for respondent in a single-person household, and "you and your family living here" for a household with multiple persons.

2.3.11 Visual or contextual indicators that help respondents or interviewers understand where they are in a question series (for example, indicating above or beside a series of questions which household member, vehicle, or source of income they are about).

2.3.12 Progress indicators (i.e., a visual indicator of where the interviewer or respondent is in the instrument as the survey progresses, applicable only for electronic instruments).
- Progress indicators are speculated to reduce breakoffs, but added graphics associated with the use of a progress indicator increases download time (Couper, 2008; Couper et al., 2000).

2.3.13 Question-level help for use as necessary by the interviewer (question-by-question objectives, including definitions) in paper or computerized surveys.

2.3.14 Validation or consistency checks and post-collection edits. For paper instruments, these should be noted in the instrument technical design specification for use in post processing. In computerized surveys with programmed consistency checks that occur during the survey interview, there is a distinction between a
- Hard consistency check (interviewer or respondent cannot continue until an inconsistency is resolved), and a
- Soft consistency check (interviewer or respondent may continue without resolving the inconsistency).

2.4 Add information to the instrument specifications that facilitates recording responses, the linking of survey instrument information and variables in a dataset (data dictionary), and documentation of the instrument and dataset, traditionally called a codebook (see Data

Dissemination guidelines; see also Appendix C). For example, specify:

2.4.1 How questions are identified in the dataset (variable names and labels), and how response categories are numerically represented and labeled (value labels).

2.4.2 Open question formats; consider the amount of space needed to provide for responses, which may differ across languages.

2.4.3 Pre-coded response options. If necessary, specify international standards for code numbers and classifications, such as occupation, language, country of origin, and religion (for example, specifications for the ESS state that codes for respondents' language(s) are based on the ISO-639-2 code frame, but use alphanumeric codes in the dataset).

2.4.4 Code number conventions (e.g., Yes=1, No=5; Yes=1 or No=2; or No=0, Yes=1). Note that code numbers are generally not shown in self-administered questionnaires. Yes=1 and No=5 is sometimes used instead of Yes=1 and 2=No to minimize error in interviewer-administered surveys. This is because the number 5 is farther away from the number 1 than the number 2 is on a computer keyboard; thus, 2 (No) is less likely to be pressed when the interviewer means to press 1 (Yes).

2.4.5 Categories for missing data categories, such as,
- Not applicable (does not apply to the respondent; question not asked based on prior answer).
- Refusal (respondent refused to answer question).
- Don't know/Can't choose.
- No answer (interviewer or respondent did not provide response, including due to errors in computerized instrument programming). Note that interviewing, coding, or statistical software may constrain labels used to create survey datasets. Specifications should indicate the values required in the final datasets and in final data documentation (codebook).

2.4.6 Data input formats, including scales that use metaphors (such as ladders or thermometers).

2.4.7 Interviewer or respondent instructions.
- Respondent show card instructions.
- Routing (skip or filtering) instructions.
- Response format or data entry instructions.
- Question level flag or mark should be added if the question-level Q by Qs information has been prepared for the interviewer or respondent to use for understanding questions asked.

2.4.8 Universe statements, that is, metadata that indicates a question or question group was asked of a specific sub-group of the survey population (e.g., "Universe [for this question]: Women aged greater than or equal to 45 years").

2.4.9 Variables to construct or recode during post-production.

2.5 Provide rules for the use of numbers, color, graphics, images, maps, and icons.

2.5.1 Ensure that numbers used in response scales visible to respondents do not have specific implications in some cultures. For example, some numbers are considered unlucky in some cultures, such as the number thirteen in the United States.

2.5.2 Ensure that colors used in instruments do not have any negative connotations in specific cultures. Color has different meaning across cultures and research has found there are cultural differences in color preferences. Any choice of colors should be validated by experts on particular cultures (Aykin & Milewski, 2005; Kondratova & Goldfarb, 2007; Russo & Boor, 1993). This may involve harmonization to a set of "culture-neutral" colors across instruments or adaptation of some colors across instruments as necessary. For example,

- Red in China means happiness while it means danger in the Western countries, as well as in Japan (Russo & Boor, 1993).
- White, black, all shades of gray, all shades of blue and a light yellow are preferentially used internationally (Russo & Boor, 1993). However, be aware of any association of specific colors with political groups in some countries.

2.5.3 Ensure that any maps used are drawn to scale.

2.5.4 Ensure that images are displayed using comparable typographical units across survey implementations.

2.5.5 Ensure that graphics, images, and icons convey comparable meaning across cultures and do not have negative connotations in specific cultures, or adapt them as necessary.

2.6 If using multiple data collection methods, include specifications for how question formats would differ across methods. For instance, a survey may be interviewer-administered in multiple modes (paper and computerized, or in-person and by telephone); it may be self-administered in two modes (Web and mail); or it may be self-administered in multiple modes (computer-assisted, paper, and Web). For example:

2.6.1 A computer-assisted self interviewing (CASI) screen might have only one question and input field per screen or have questions with same response scales per screen (to minimize respondent burden), whereas an interviewer-administered computer-assisted screen Fsmight have multiple questions and multiple input fields.

2.6.2 Self-administered instruments may be developed without response codes (the respondent clicks on a response option, or clicks on a radio button, or checks a box), whereas some computer-assisted personal interview (CAPI) surveys may require numbered response options for entry of responses, if numbers are the only possible form of input.

2.6.3 Software constraints may also necessitate alternate specifications, for example, if different software were used for Web and computer-assisted telephone interviewing components.

2.7 Based on the guidelines specified above, as well as the interface design and paper instrument guidelines that follow, prepare a survey instrument specification with all survey contents for the instrument as well as a data dictionary, which represents the contents of the survey dataset. Also specify the codebook metadata before data collection.

Lessons learned

2.1 Seemingly small differences in instrument design across cross-cultural surveys can influence responses across cultures. For example, scales that are not formatted consistently, response options with misaligned check boxes, differences in the relative amount of space allowed for open responses, and differences in the physical placement of follow-up questions have been shown to lead to missing data or unusual response distributions across surveys (Smith, 1993). For example, in the 1987 ISSP there was a question on subjective social stratification. Respondents in nine countries were asked to rate themselves on a scale from 1 to 10 (top to bottom). In all countries respondents tended to rate themselves in the middle, and a small proportion of respondents rated themselves in the bottom. However, the Netherlands had 60% in the middle, compared to 72% to 84% in other countries, and had 37% in the bottom, compared to 6% to 24% in other countries. Dutch respondents did not have such a distinctive distribution on other social inequality measures. On examination, it was found that the Dutch translation was comparable to English, but the visual display of the scale differed (see Appendix D).

2.2 On the other hand, cultural customs and norms may require using different graphic images, icons, colors, etc. For example, in 2007, the ISSP allowed countries to use different graphics for an ideal body shape question. See Appendix D for images used in the Austrian and Philippines questionnaires.

2.3 The layout of scales should not deviate from the source questionnaire, e.g. a horizontal scale should never be changed into a vertical scale. Likewise, the order of response categories should not be reversed, e.g. "extremely happy" – "extremely unhappy" should not become "extremely unhappy – "extremely happy". Such changes can contribute to measurement error (Dorer, 2012).

2.4 When underlining is used to emphasize words or phrases to be stressed by interviewers, the emphasis should be maintained in the target languages questionnaire. This may at times mean that a different word or groups of words will need to be stressed if a close translation has not proved possible (Dorer 2012).

2.5 Hashtags are commonly used on social media platforms such as Twitter, Instagram (IG), Facebook, etc. and multiple data extraction tools or Application Programming Interfaces (APIs) have been developed for researchers to access these data (See Appendix H). Data entry instructions may need to be provided to respondents or social media users to include a # before entering any response, or to record the full response without any space to reduce data processing efforts.

3. Develop language-specific guidelines for the survey instrument as necessary.

Rationale

Different language features across cultures are important in designing survey instruments. Survey instrument designers should consider both languages and countries or cultures when developing language specifications, since there is no one-to-one match in languages and cultures.

Some countries share the same language (e.g., English), but may have different language layout systems, and some use multiple languages in a country (e.g., Belgium and Switzerland). In addition, some countries have more than one script or system of writing (e.g., Japan). Therefore, consider any differences across survey implementations in scripts, character sets, fonts, text directions, spelling, and text expansions when developing instrument technical design specifications (Aykin & Milewski, 2005). This is important for computerized instruments, since software may need to be configured and instruments reprogrammed to display languages in cultures for which they were not originally developed.

Procedural steps

3.1 Provide instrument formatting specifications that facilitate the translation of languages (see Translation: Overview), specifying scripts, character sets, fonts, spacing, and so on, for target languages (Aykin (Ed.), 2005; Aykin, & Milewski, 2005; Jagne & Smith-Atakan, 2006; Russo & Boor, 1993) and the programming of computer-assisted instruments; formatting guidelines should address aspects of design such as:

3.1.1 Language- and region-specific character sets.
- The International Organization for Standardization (ISO) 8859 Character Set has language-specific groupings, for example, ISO 8859-1 for Western Europe and ISO 8859-2 for Central and Eastern Europe.

3.1.2 Differences in languages and scripts; for example:
- Japan has one language, but several scripts, which can be mixed.
- China has one official language, Mandarin (Putonghua), seven major languages, and many dialects. Also, Chinese may be displayed in either Traditional or Simplified script.

3.1.3 Differences in fonts that support different character sets; in general:
- Avoid complex or ornate fonts.
- Provide interline space to ensure clear separation between lines and to accommodate underlining.
- Provide space to accommodate changes in line heights.
- Provide flexibility in layout of the instrument to accommodate expansion or contraction of text during translation. For example, use a larger font and/or margins for an English instrument, if translating from English into other languages would increase the amount of space required for text in culture-specific instruments.

3.1.4 Differences across languages in punctuation (e.g., the different question marks in English and Spanish,? and ¿, respectively).

3.1.5 Language- or culture-specific differences in the way characters are sorted alphabetically, including diacritics (accent marks above or below letters, e.g., É), ligatures (multiple letters treated as single typographical units, e.g., æ, œ, and ß), character combinations (e.g., ch follows h in Czech), and uppercase and lowercase letters. For instance, the Ä sorts after Z in Swedish, but after A in German. This is important for computerized survey software that was designed for one type of culture but used in other cultures or countries that sort lists such as response options differently.

3.2 Consider differences in text or figure directionality and provide application design specifications that can be adapted to translated instruments with differing text or figure directionality; the three types of text or figure directionality are:
   3.2.1 Left-to-right (Latin, Cyrillic, Greek, Thai, and Indic languages).
   3.2.2 Left-to-right and vertical (Chinese, Japanese, and Korean).
   3.2.3 Bi-directional (Arabic and Hebrew characters displayed right to left; Latin characters displayed left to right).
   3.2.4 Text directionality applies to displaying images. For example, in Arabic and Hebrew where, the text is read from right to left, images are also read from right to left (Aykin & Milewski, 2005).

Lessons learned

3.1 In Asian countries, vertical text direction is seldom used for survey questions, but it is sometimes used for response options. In the 2006 East Asia Barometer survey, there were differences across countries in the use of vertical text. Mainland China and Taiwan used vertical text for response options, but Singapore did not. In the ISSP in 2007, Japan and China used vertical text. When vertical text was more than one line, they were displayed from left to right in Japan, although they were displayed from right to left in mainland China (see Appendix E). These differences suggest both that design specifications need to reflect an understanding of how different Asian countries display text both vertically and horizontally, and that it would be desirable to pretest separately questions that differ across countries.

3.2 Tanzer (2005) cautions against administering visual representations to right-to-left readers (of Arabic or Japanese, for example) that are meant to be processed from left-to-right (of English, for example). In studies comparing the results of a pictorial inductive reasoning exercise administered to Arabic-educated Nigerian and Togolese high school students with that of an Austrian calibration sample, researchers found the Arab-educated students exhibited far more difficulty using the left-to-right processing format required by the test than the Austrians because Arabic is read from right-to-left. In a 3MC project in six countries in the Middle East, researchers discovered during the design phase that some Arabic-to-English translations uncovered potential differences between how respondents in the Middle East and respondents in western countries visualize and mentally process rating scales (de Jong & Young-DeMarco, 2017).

4. Develop interface design rules for computerized survey applications, and for self-administered paper instruments.

Rationale

Interface design has an effect on the respondent-computer or interviewer-computer interaction, influences user performance, and may affect data quality. Design should not only minimize respondent and interviewer burden and thus maximize usability, but should also be consistent across survey implementations. Therefore, it is important to provide clear guidelines for design of instructions, questions, error messages, and screen elements for computerized instruments (see Appendix A for an example of basic design guidelines for computer-assisted surveys). Note that similar rules are necessary for data entry applications (see Data Processing and Statistical Adjustment).

Many of the principles for interface design of computerized instruments are also relevant to paper instruments. They can just as easily address the usability of paper instruments, whether they are for interviewer-administered or self-administered surveys. In the procedural steps below, no distinction is made between computerized and paper instruments if a step would apply to both paper and computerized surveys. Where necessary, distinctions are made between computer-assisted and Web interface design.

Procedural steps

4.1 Establish the key principles for design, which should lead to effective assessment of the quality of design (see Guideline 5). These include:
- 4.1.1 Consistency.
- 4.1.2 Visual discrimination among questions and related elements, so that interviewers and respondents quickly learn where different elements are located, and thus where to look for what type of element. For example, interviewer and respondent instructions may appear in a smaller text, a different font, and/or a color, to distinguish them from the question text.
- 4.1.3 Adherence to a culture's normal reading behavior for each language and script, based on issues such as text directionality (see Guideline 3).
- 4.1.4 Display of instructions at points appropriate to associated tasks.
- 4.1.5 Elimination of unnecessary information or visual display of other features that distract interviewers and respondents.

4.2 Provide rules for the layout and formatting of question elements, including:

4.2.1 Question text, which should be the primary focus of a question, and its related information.

4.2.2 Response options, which should have instructions or visual characteristics that convey whether a single mutually-exclusive response or multiple responses are possible. For example, in computerized instruments, radio buttons convey there should be one response, and check boxes convey that there may be multiple responses, which should be reinforced by an instruction (e.g., Select all that apply).

4.2.3 Response input fields should convey the length of the response expected. For example:
- An open-ended response area is as wide and has as many lines as the expected length of response.
- The width of an integer response area should be as many number of character lengths wide as the expected input, that is, one character length for a one-digit integer, a two-character length for a two-digit integer, etc.

4.2.4 Instructions, which should appear as expected in relation to task demands; for example, a reference to a respondent booklet or show card should appear before question text, and a probe or data entry instruction after question text.
- Layout can also play a role when deciding on translations for interviewer or respondent instructions. If the instruction reads “Please tick one box” (as in the self-completion supplementary questionnaires), the translation for “box” should match the symbol that is eventually used, such as “□” or “o”. Equally, the translation for “tick” should match the actual action (tick? mark? touch?), which can depend on whether the questionnaire is computer- or paper-based (Dorer, 2012).

4.2.5 In computerized instruments, the interface should facilitate accessing online help, through clear formatting of help text and design of navigational aids that facilitate opening and closing help text windows.

4.2.6 Error messages, warnings, and consistency checks in computerized instruments should clearly identify the nature of the problem, reflect actual question wording if necessary (e.g., for interviewer probes for more accurate responses), and convey how to resolve the problem (see Murphy, Nichols, Anderson, Harley, & Pressley (2001) for examples and for more detailed guidelines on design of error messages).

4.2.7 Context markers (for example, instrument section labels, household member numbers, and so on).

4.2.8 Additional information may be required for Web self-administered surveys, such as contact information and graphic and/or text identification of the sponsoring organization.

4.2.9 In Web surveys, provide guidance on whether to use a paging versus a scrolling design (Peytchev, Couper, McCabe, & Crawford, 2006). Provide rules for handling cultural differences, for example, differences in paper sizes for paper surveys. In such cases, provide guidance on pagination in order to avoid inadvertent context effects (for example, two related questions appearing together on one page in one country's survey and on separate pages in another).

4.2.10 Provide examples of key question types and elements for all target languages and cultures, and for different types of administration if relevant (see Appendix A for examples of computerized questions and Appendix B for examples of paper questions).

4.2.11 Provide examples of correct formatting of elements, for all question types (see Guideline 1) and all languages and cultures (see Appendix A).

Lessons learned

4.1 There is increasing evidence that the visual design of computer-assisted and Web instruments can impact data quality (Christian, Dillman, & Smyth, 2005; Couper, 2008; Couper et al., 2000; Couper, Traugott, & Lamias, 2001; de Leeuw, et al., 2003). For example, providing an input box or field that allows entry of 10 numbers with no guidance or instruction on input format can lead to poorer data quality than if the survey question more precisely calls for an integer of up to three digits; for example, instead of "20," "90" or "100" in an entry field with a width of three (___), a Web survey respondent enters "40 to 50" in a field with a width of 10 (-_________) can lead to poorer data quality due to possible entry errors.

4.2 Not providing rules for formatting questionnaires printed on different sized paper can lead to poorer comparability of data across countries. For example, in the ISSP one country lost the last item in a scale when copying the scale from A4 size paper (8.27" by 11.69") to letter size paper (8.5" by 11") (Smith, 2005).

5. Establish procedures for quality assurance of the survey instrument that ensures consistency of design, adapting evaluation methods to specific cultures as necessary.

Rationale

As discussed in Guideline 4 above, research shows that instrument technical design can affect data quality in computer-assisted or Web surveys, positively or negatively. This is also true of paper instruments. Thus, it is important that pretesting (see Pretesting) of comparative survey instruments include procedures for assessing the quality of the design of the survey instrument and adaptations for specific culture, languages, and modes, not just the quality of the content. This includes the evaluation of the use of color, graphics, images, maps, and icons. As indicated earlier, such evaluation procedures may require adaptation across cultures.

Procedural steps

5.1 Identify a team with members that have expertise in evaluation of technical instrument design. Such experts may include substantive experts, survey methodologists, linguists, and usability professionals, and should include someone with an understanding of response styles across cultures.

5.2 Provide a clear set of instrument specifications and/or a data dictionary for the instrument and culture-specific adaptations (per rules outlined in Guideline 2), which will facilitate testing and assessment of the instruments. Such documentation would include: question (variable) names and labels; question text; response option values and labels; numeric response formats and ranges, and specifications for the lengths allowed for open-ended question text; interviewer or respondent instructions; missing data values; skip instructions; and so on. It should enable comparison of computerized or formatted paper instruments to instrument design specifications.

5.3 Identify appropriate instrument evaluation procedures for the comparative surveys under evaluation. These may be more or less extensive based on whether survey organizations in the targeted cultures previously have used specific guidelines, instruments, and survey software. Most questionnaire pretesting tools (see Pretesting) may be used to evaluate instrument design as well as questionnaire content and data collection procedures. These include:
5.3.1 Expert review or heuristic evaluation, in which one or more experts evaluates the instrument design against a set of evaluation criteria or heuristics, for example:
- Consistency and adherence to design guidelines.
- Error prevention.
- Usefulness of documentation, definitions, help, error messages, and other feedback to users.
- Ease of navigation.
- Ease of recognition of specific question or instrument elements and actions required.

5.3.2 Review of an instrument, data dictionary, or codebook to ensure adherence to instrument specifications for naming and labeling of variables and response options. This should include comparison across instruments or data dictionaries for all survey implementations.

5.3.3 Laboratory or on-site tests of instrument design with users or participants with similar characteristics to target interviewers or respondents. These are called usability tests when evaluating computer-based instruments, but they also may be used to evaluate paper instruments. Since culture-specific response styles affect how participants respond to questions about usability (Clemmensen & Goyal, 2005) every effort should be made to match tester and participant characteristics, language, and cultural background.

5.3.4 If feasible, incorporate methodological experiments on formatting, to assess whether aspects of formatting affect respondents differentially across cultures.

.

5.4 Test instruments locally on top of central testing.

5.4.1 Field instruments that require Internet connection should be tested for connectivity in field situations. An offline alternative should be established if there are connectivity issues.

5.5 Collect measures from all instrument evaluation procedures that will lead to informed decisions about question- or screen-specific or global design changes that need to be made (see Pretesting). Examples include:

5.5.1 Questionnaire length and section and item timings.

5.5.2 Audit trails for computer-assisted or Web applications, which can include item timestamps, keystrokes, mouse actions, and functions invoked. Gathering some of these requires programming that captures information directly from the respondent's computer Heerwegh (2003) provides sample programming code for capturing such paradata for Web surveys).

5.5.3 Behavior codes or event codes based on video or audio recordings that reflect problems using the survey instrument. Such methods are appropriate for both paper and computer-assisted instruments.

5.5.4 Qualitative analyses of cognitive and usability testing.

5.5.5 Heuristic evaluation or expert review.

Lessons learned

5.1 Research (Couper, 1999; Hansen & Couper, 2004) has shown that techniques for evaluating the effectiveness of paper materials and computer software work very well in the evaluation of the design of survey instruments. For example, usability evaluation methods (commonly used in the development of software to assess the quality of user interfaces) and traditional pretesting methods such as conventional pretests, cognitive interviews, and behavior coding can be used to identify instrument design problems as well as problems related to question content.

5.2 Interviewer and participant interaction may need to be considered for usability tests of instruments used in 3MC surveys. There is evidence that when an interviewer is from the same culture as participants, interviewers give more help, tell more about introductions, and encourage participants more frequently; and participants report more usability problems and give more suggestions than when an interviewer is from a different culture (Sun & Shi, 2007). On the other hand, some research indicates that when interviewers are from cultures speaking different languages, participants explain more about their choices of design elements (Vatrapu & Pérez-Quiñones, 2006).

5.3 Incorporating methodological experiments into cross-cultural surveys, whether for experiments on instrument design or other methodological issues, can be difficult to negotiate. It involves agreement of funding agencies, the central coordinating center (if there is one), and the survey organizations involved. It also requires that clear experimental design specifications are included as part of the development of design specifications prepared for each survey organization (see Guideline 2).

6. **Consider all possible formats and layouts, particularly when a survey is self-administered on devices provided to the respondent or administered on the respondent's personal device or devices that respondents can access to complete surveys in a public setting (See Study Design and Organizational Structure).**

Rationale

A self-administered component may be better when the partial of interview is sensitive, although this varies by social context (see Data Collection: Face-to-Face Surveys for further discussion). When using CASI and A-CASI modes, attention to the details discussed below that facilitate the respondent experience can lead to increased data quality.

Procedural Steps

6.1 Ensure that there is a good fit between the project and the technological device.

6.1.1 Handheld devices such as personal digital assistants (PDAs) or smartphones may be more appropriate for smaller or simpler questionnaires.
6.1.2 An important limitation of PDAs and smartphones is that they are not as suitable for collecting open-ended responses (Escandon, Searing, Goldberg, Duran, & Monterrey Arce, 2008).
6.1.3 Particularly with the use of a PDA or smartphone, researchers need to be aware of the size of the device relative to the interviewer's hand.
6.1.4 Interviewers might lose track of where they are in the sequence of questions (Groves & Mathiowetz, 1984; House & Nicholls, 1988; Couper, 2000) and might find it difficult to retain a comprehensive picture of the instrument since they see only one screen at a time.
6.1.5 Moreover, interviewers might find it more difficult to handle qualitative open-ended questions that require a lot of typing verbatim answers. Handheld devices (e.g., smartphones) are not as suitable for collecting open-ended responses as are laptops (Escandon, et al., 2008).

6.2 Implement a system of work ownership. All personnel can be assigned a code for database entry, supervision, and analysis. Logs can be generated to monitor and control data management and information flow.
6.2.1 Additional attention should be given to non-Latin languages (i.e., Chinese, Arabic, Russian, etc.) when selecting technology and programming software. Not all software packages can support non-Latin script.
6.2.2 Allocate sufficient time to designing and pretesting the electronic questionnaire and to overall testing and debugging the software, or difficulties can arise due to lack of adequate preparatory time (Onono, Carraher, Cohen, Bukusi & Turan, 2011).This is particularly for questionnaires in multiple languages, especially if the survey uses a non-Latin script and/or if the questionnaire is lengthy and complex, as it is crucial to ensure that the question flow and skip patterns function correctly before using them in the field.

6.3 Consider using paper documents for certain aspects of the survey. For example, interviewers in China using handheld computers reported that it was overly time-consuming to read the full consent form on a small screen (Wan, et al., 2013).
6.3.1 In a public health survey in China, interviewers reported that entering Chinese characters using the handwriting recognizer was too time-consuming and entering Chinese characters with

the stylus into the handheld computer was also difficult (Wan et al., 2013).

6.4 When using CASI and A-CASI modes, attend to details that facilitate the respondent experience, leading to improved data quality.

6.4.1 Consider disabling the screen saver and power-saving settings on the device so that screens do not go blank if a participant takes additional time to answer a question (National Institute of Mental Health, 2007).

6.4.2 Graphical and/or audio representations of the response process can help guide the respondent through the interview. In a survey using A-CASI in India, the entry of a response was marked by the change in the color of the corresponding response bar on the screen to grey, along with a "beep" sound. A "Thank you" screen indicated the end of the survey (Bhatnagar, et al., 2013).

6.4.3 If a participant did not answer a question after approximately 60 seconds, consider repeating the question and/or programming additional text can be programmed to appear encouraging participants to answer the item(s) in a truthful manner (National Institute of Mental Health, 2007).

6.4.4 If a keyboard is used, it should be user-friendly.

- Keyboard options can be limited to responses (e.g., YES, NO, and numbers) and larger color-coded keyboard keys could be used (see Appendix G).
- Additional keyboard shortcuts to replay questions can also be marked.

6.4.5 Text on the computer screen should be large enough to be easily legible for respondents

6.4.6 In an A-CASI survey in India, neither the question nor the response texts were displayed on the screen to ensure privacy and confidentiality for the respondents (Bhatnagar et al., 2013)

6.4.7 Touchscreens on A-CASI instruments can be particularly helpful for less-educated populations (Lara, Strickler, Olavarrieta, & Ellertson, 2004).

6.5 Consider the different types of mobile devices that a respondent may use to complete a survey. For example, Web surveys may be accessed through computers, smartphones, or mobile tablets and completed on one or more devices. Bring your own device (BYOD) has become a trend for telephone surveys and surveys can be administered at a time most convenient for the respondent.

6.5.1 To achieve its cost and quality targets and meet its strategic goals for Census 2020, the U.S. Census Bureau continues to explore the public's willingness to be enumerated given a BYOD concept in which interviewers are using their personally owned devices (U.S. Census Bureau, 2012; Holzberg & Eggleston, 2016).

6.6 Consider collecting data using Short Message Service (SMS) text, with reminders sent to mobile phones (Zurovac et al., 2011; West, Ghimire, & Axinn, 2015; Lau, Lombaard, Baker, Eyerman, & Thalij, 2016) or with the use of “apps” (Sonck & Fernee, 2013) for surveys that specifically target respondents that possess smartphones.

Lessons Learned

6.1 Do not underestimate the additional time needed for preparation when using technology. In a survey in Burkina Faso, researchers reported underestimating the amount of work required to program questionnaires, and as a result failed to maximize the use of some of the available options for input checking and other real-time quality control procedures. Village names, for example, were implemented as a text-entry field, but would have been better as a drop-down list to avoid ambiguities of spelling, etc. Combinations of input checks, plus quality control measures at the stage where data were downloaded to portable computers in the field, should have picked up concerns at an earlier and remediable stage (Byass et al., 2008).

6.2 In a Bolivian survey, interviewers reported that longer survey questions disrupted the flow of the interview because of extra scrolling time (Escandon et al., 2008).

6.3 In a Kenya study, A-CASI had much lower rates of missing data than the paper self-administered questionnaire; and similar rates to the standard interviewer-administered paper questionnaire. Use of computers in rural populations was sometimes met with suspicion and opposition.

6.4 In a Malawi study (Mensch, Hewett, Gregory, & Helleringer, 2008), reporting for "ever had sex" and "sex with a boyfriend" is higher in the face-to-face (FTF) mode than self-administered A-CASI. Instead, reporting about other partners as well as multiple lifetime partners, however, is consistently higher with A-CASI than FTF. Overall, the FTF mode produced more consistent reporting of sexual activity between the main interview and a subsequent interview. The association between infection status and reporting of sexual behavior is stronger in the FTF mode, although in both modes a number of young women who denied ever having sex test positive for STIs/HIV in associated biomarker collection.

6.5 Comparisons with alternate administration modes suggest that the audio self-administered questionnaire mode strongly increased reporting of socially undesirable behaviors. Further analyses suggest that when self-administration is combined with the use of earphones the threat of bystander disapproval (as opposed to interviewer disapproval) is reduced by effectively isolating respondents from their social environment.

6.6 In Kenya, each text message reminder included a quote that was up to 40 characters long and was unrelated to the topic of the survey, malaria case-management, but was designed to be motivating, entertaining, or merely attention-getting, to increase the probability that health workers would read the messages and respond to the survey (Zurovac, et al., 2011).

See Data Collection: Face-to-Face Surveys for further literature on the use of CASI and A-CASI.

7. **Maintain complete documentation of source and target language or culture-specific instruments, including specification and design guidelines, and provide comprehensive summaries of the same for data dissemination and analysis.**

Rationale

Comprehensive documentation of survey instruments or applications is an essential component of study documentation and comes into play at all stages of the survey lifecycle (questionnaire development, pretesting, data collection, post processing, and data dissemination and analysis). Complete and consistent rules for specifying and designing instruments are important (although not sufficient) to ensuring survey data meet the quality requirements of users (see Survey Quality). Documentation of instrument design specifications also plays a significant role in this regard. In 3MC surveys, it also facilitates the assessment of comparability of survey data across cultures. The rapid increase in computer-assisted data collection methods makes it increasingly possible to provide well-documented survey data. Based on study design, the study coordinating center, the survey agency, or both would be responsible for maintaining documentation related to technical instrument design.

Procedural steps

7.1 Maintain documentation of the rules specified for technical instrument design.

7.2 Maintain documentation of quality assessments of the survey instruments, and the outcomes of decisions made to revise the instrument design.

7.3 Maintain specifications for the final source instruments, based on Guideline 1, Guideline 2, Guideline 3, and Guideline 4 above. These should include the instrument specifications and data dictionaries developed by the coordinating center and/or survey organizations.

7.4 Maintain alternative specifications for target languages or cultures as necessary. For example, if the source instrument is computer-assisted, but it is necessary to develop a paper instrument for one or more locations, separate specifications should be developed for paper instruments.

7.5 Maintain paper and/or electronic copies of all culture-specific instruments or adaptations of instruments, to facilitate comparison of technical design across culture-specific surveys.

7.6 Maintain question-level metadata (question text, response options, instructions, text fills, population universes, definitions, etc.) in an electronic format to facilitate linking and comparing metadata for all survey instruments (e.g., eXtensible Markup Language (XML) data files). If feasible, this should be part of a centralized documentation system that links question metadata and formatting with data codebooks for data disseminated. Some computer-assisted data collection software now makes this possible.

7.7 Provide comprehensive documentation of survey instruments, based on all of the sources of documentation listed above.

Lessons learned

7.1 Survey instrument design and documentation of design rules and specifications can affect the quality of data produced and disseminated, and the ability of users to effectively analyze survey data. Hert (2001) conducted studies of users "interacting" with statistical data in order to understand how to better meet their needs. In one study she found that the completeness and quality of available question-level survey instrument documentation and metadata affected users' selection of variables for analysis. In particular, she found that users used a number of mechanisms for identifying appropriate variables for analysis, including what they knew about variable naming conventions, how particular questions relate to other questions, and even coding categories, if the question text did not provide enough information for selection. These findings reinforce the need for clear documentation of technical design guidelines and instrument specifications, and for these to be readily available to data users.

Appendix A

Technical Design Standards

Following are some basic standards or rules for design of interviewer-administered computer-assisted instruments using Blaise interviewing software (Survey Research Center, 2007) which were based on initial research on developing guidelines for Computer-Assisted Personal Interviews (CAPI) (Couper, 2001). These are included to convey the types of information to include in such standards; for example, display instructions in a smaller font of a different color than question text. Standards for 3MC studies should reflect the requirements for design of instruments across cultures, which could dictate choice of fonts, colors, and so on. The referenced standards included examples of basic screen types formatted according to the standards (see Figures A1 and A2 for selected question type examples).

Text Characteristics

- Display question text on a light background color (cream), in mixed case, and in 12-point Arial, black.
- Display instructions in **11-point Arial bold blue**.
- Display response categories:
  - Those read out to the respondent, in 12-point Arial black.
  - Those not read out to the respondent, in **11-point Arial bold blue**.
- Use underline for emphasis, sparingly.
- Place optional text in (parentheses).
- Display in-text references to function keys and numbers to type in mixed case within square brackets, for example, [Enter], [1], [F12], and [Ctrl-R].

On-Screen Instructions and Other Information

- Place references to interviewer aids (e.g., an event history calendar or show card instruction) and the question text in the upper left corner of the screen, above the question text.
- Place instructions that precede the question flush left with the question;
- Use icons to distinguish special instructions:
  - **Page 1**, for respondent booklet instruction,
  - **Calendar**, for event history calendar instruction, and,
  - **Interviewer Checkpoint**.
- "Bullet" all other interviewer instructions with an 11-point bold blue diamond (♦ **Enter [1] to continue**).
- Single space within an instruction and double space between instructions.
- Place an online help indicator ([F1]-Help) above the question on the right margin, for questions with "question -by-question objectives" (QxQ's).
- Indent instructions that follow the question.

- Place any context-related information below the question-level help indicator on the right margin (for example, changing person-level information as the interviewer navigates a household roster or grid).
- Display instructions in the order associated with required interviewer tasks.
- Include an actual question in explicit interviewer checkpoints, displayed in **11-point Arial bold blue**.
- Capitalize only key task-related action verbs (ASK, READ, ENTER, and PROBE), and only at the beginning of instructions.
- Keep instructions simple and concise.
    - Put long instructions or those not directly related to asking questions or entering responses into online help (question-by-question objectives).
- Conditional instructions start with the conditional phrase, not the action verb, and the action verb is not capitalized (e.g., conditional probes and data entry instructions).
- In probe instructions, place text to be read to the respondent in Arial black.
- Place references to respondent answers in quotation marks.

Examples of Formatted Questions

Figure A1. Example of multiple response questions

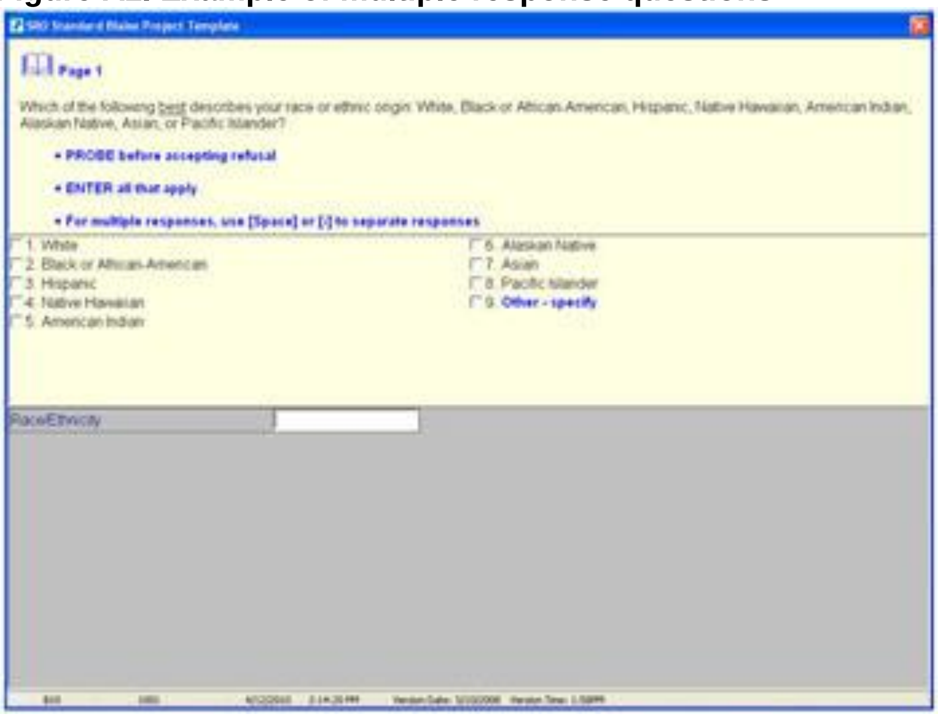
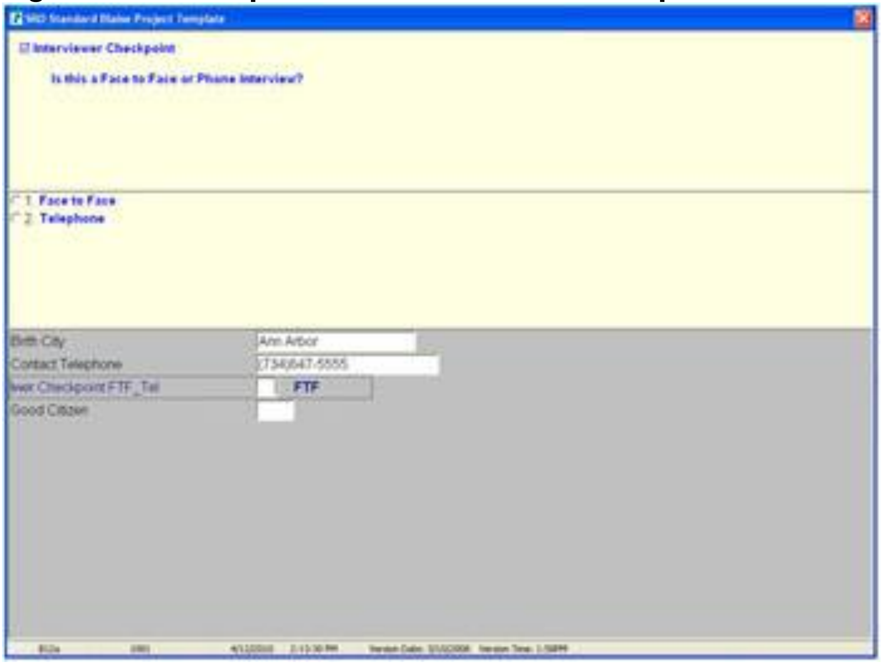


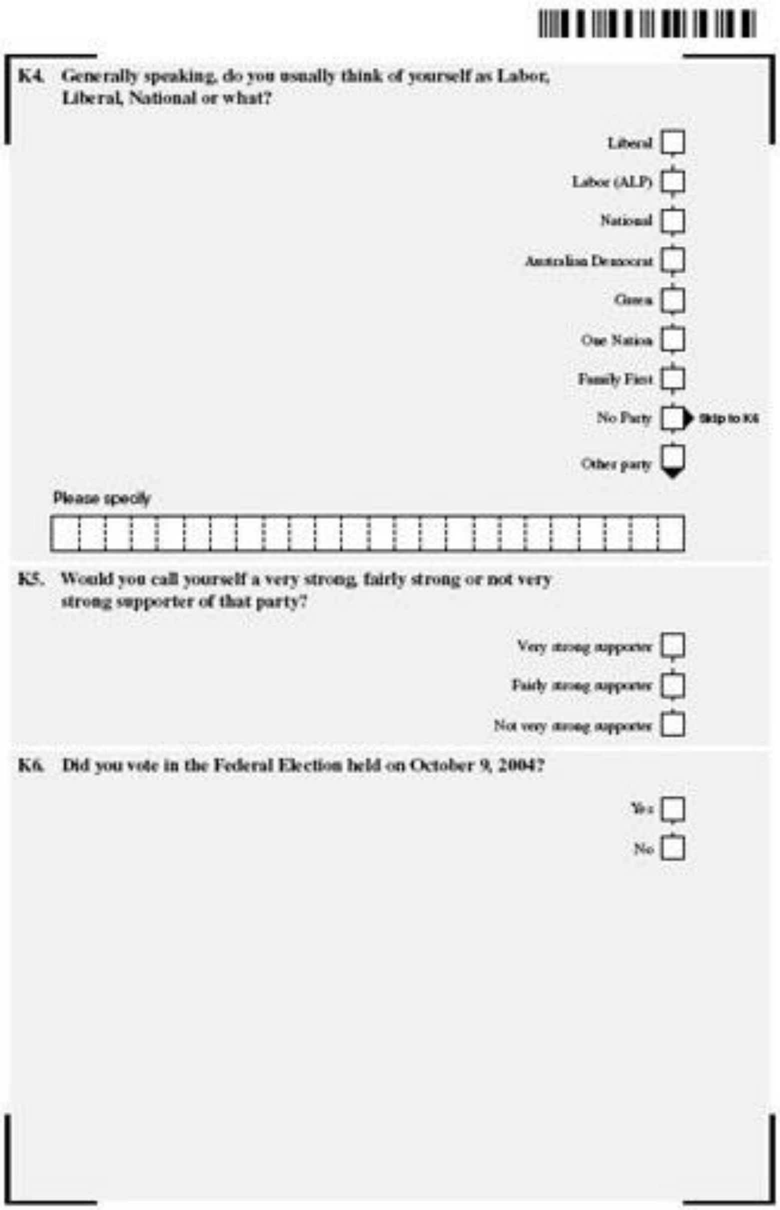
Figure A2. Example of an interviewer checkpoint



Appendix B

Following are examples taken from the ISSP 2007 and the U.S. Census 2010 self-administered paper questionnaires. They both show instructions to the respondent, including skip instructions.

Figure B1. Example of self-administered questions from the Australian ISSP 2007



K4. Generally speaking, do you usually think of yourself as Labor, Liberal, National or what?

Liberal ☐
Labor (ALP) ☐
National ☐
Australian Democrat ☐
Green ☐
One Nation ☐
Family First ☐
No Party ☐ ► Skip to K6
Other party ☐

Please specify

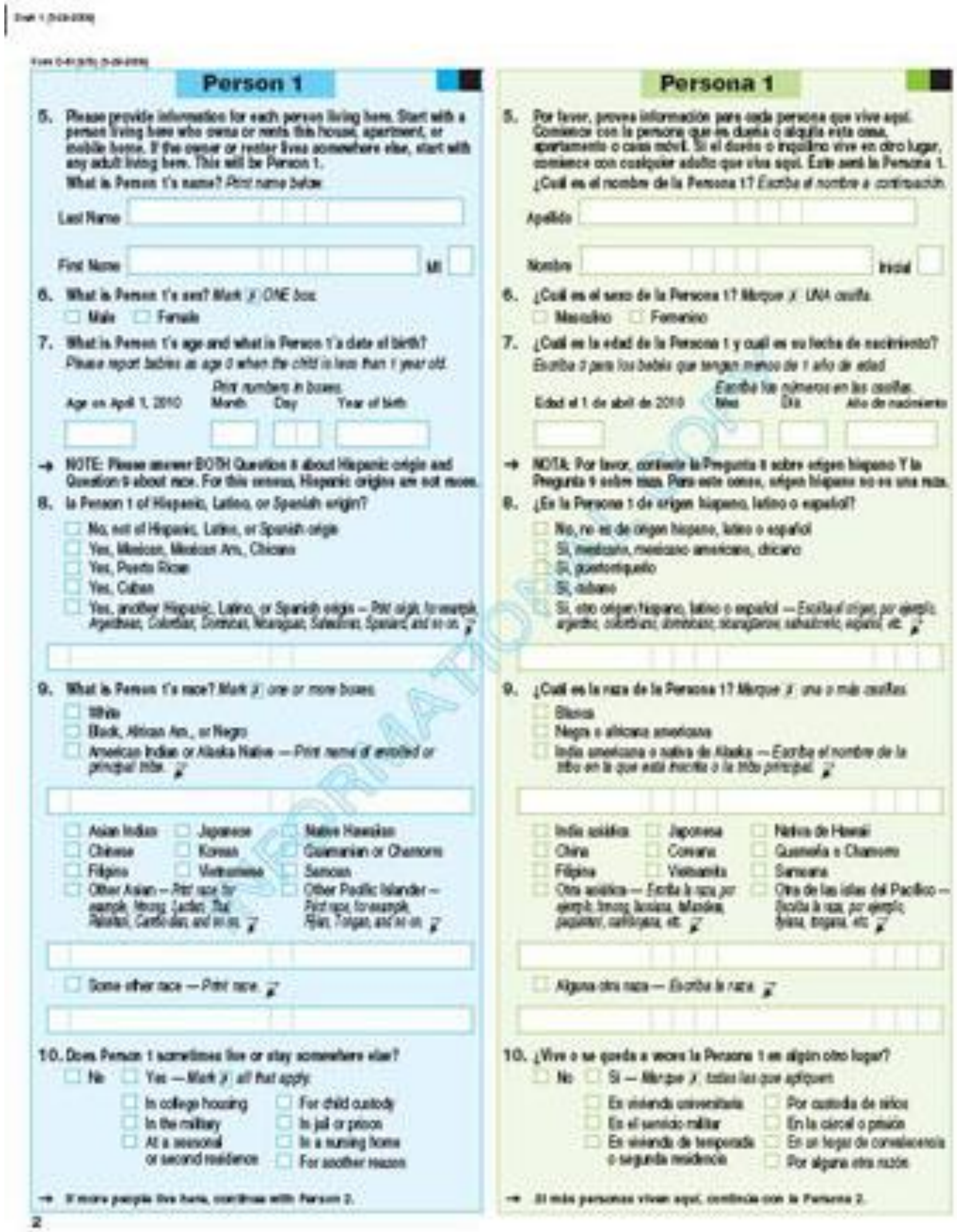
K5. Would you call yourself a very strong, fairly strong or not very strong supporter of that party?

Very strong supporter ☐
Fairly strong supporter ☐
Not very strong supporter ☐

K6. Did you vote in the Federal Election held on October 9, 2004?

Yes ☐
No ☐

Figure B2. Example of self-administered questions from the U.S. Census 2010 Bilingual (English and Spanish) paper self-administered questionnaire



Person 1

5. Please provide information for each person living here. Start with a person living here who owns or rents this house, apartment, or mobile home. If the owner or renter lives somewhere else, start with any adult living here. This will be Person 1.
What is Person 1's name? *Print name below.*
Last Name
First Name MI

6. What is Person 1's sex? *Mark X ONE box.*
☐ Male ☐ Female

7. What is Person 1's age and what is Person 1's date of birth?
Please report babies as age 0 when the child is less than 1 year old.
Print numbers in boxes.
Age on April 1, 2010 Month Day Year of birth

→ NOTE: Please answer BOTH Question 8 about Hispanic origin and Question 9 about race. For this census, Hispanic origins are not races.

8. Is Person 1 of Hispanic, Latino, or Spanish origin?
☐ No, not of Hispanic, Latino, or Spanish origin
☐ Yes, Mexican, Mexican Am., Chicano
☐ Yes, Puerto Rican
☐ Yes, Cuban
☐ Yes, another Hispanic, Latino, or Spanish origin — *Print origin, for example, Argentinean, Colombian, Dominican, Nicaraguan, Salvadoran, Spaniard, and so on.*

9. What is Person 1's race? *Mark X one or more boxes.*
☐ White
☐ Black, African Am., or Negro
☐ American Indian or Alaska Native — *Print name of enrolled or principal tribe.*
☐ Asian Indian ☐ Japanese ☐ Native Hawaiian
☐ Chinese ☐ Korean ☐ Guamanian or Chamorro
☐ Filipino ☐ Vietnamese ☐ Samoan
☐ Other Asian — *Print race, for example, Hmong, Laotian, Thai, Pakistani, Cambodian, and so on.* ☐ Other Pacific Islander — *Print race, for example, Fijian, Tongan, and so on.*
☐ Some other race — *Print race.*

10. Does Person 1 sometimes live or stay somewhere else?
☐ No ☐ Yes — *Mark X all that apply.*
☐ In college housing ☐ For child custody
☐ In the military ☐ In jail or prison
☐ At a seasonal or second residence ☐ In a nursing home
☐ For another reason

→ If more people live here, continue with Person 2.

2

Persona 1

5. Por favor, provea información para cada persona que vive aquí. Comience con la persona que es dueña o alquila esta casa, apartamento o casa móvil. Si el dueño o inquilino vive en otro lugar, comience con cualquier adulto que viva aquí. Este será la Persona 1.
¿Cuál es el nombre de la Persona 1? *Escriba el nombre a continuación.*
Apellido
Nombre Inicial

6. ¿Cuál es el sexo de la Persona 1? *Marque X UNA casilla.*
☐ Masculino ☐ Femenino

7. ¿Cuál es la edad de la Persona 1 y cuál es su fecha de nacimiento?
Escriba 0 para los bebés que tengan menos de 1 año de edad.
Escriba los números en las casillas.
Edad el 1 de abril de 2010 Mes Día Año de nacimiento

→ NOTA: Por favor, conteste la Pregunta 8 sobre origen hispano Y la Pregunta 9 sobre raza. Para este censo, origen hispano no es una raza.

8. ¿Es la Persona 1 de origen hispano, latino o español?
☐ No, no es de origen hispano, latino o español
☐ Sí, mexicano, mexicano americano, chicano
☐ Sí, puertorriqueño
☐ Sí, cubano
☐ Sí, otro origen hispano, latino o español — *Escriba el origen, por ejemplo, argentino, colombiano, dominicano, nicaragüense, salvadoreño, español, etc.*

9. ¿Cuál es la raza de la Persona 1? *Marque X una o más casillas.*
☐ Blanca
☐ Negra o africana americana
☐ India americana o nativa de Alaska — *Escriba el nombre de la tribu en la que está inscrita o la tribu principal.*
☐ India asiática ☐ Japonesa ☐ Nativa de Hawaii
☐ China ☐ Coreana ☐ Guameña o Chamorro
☐ Filipina ☐ Vietnamita ☐ Samoana
☐ Otra asiática — *Escriba la raza, por ejemplo, hmong, laosiana, tailandesa, paquistaní, camboyana, etc.* ☐ Otra de las islas del Pacífico — *Escriba la raza, por ejemplo, fiyiana, tongana, etc.*
☐ Alguna otra raza — *Escriba la raza.*

10. ¿Vive o se queda a veces la Persona 1 en algún otro lugar?
☐ No ☐ Sí — *Marque X todas las que apliquen.*
☐ En vivienda universitaria ☐ Por custodia de niños
☐ En el servicio militar ☐ En la cárcel o prisión
☐ En vivienda de temporada o segunda residencia ☐ En un hogar de convalecencia
☐ Por alguna otra razón

→ Si más personas viven aquí, continúe con la Persona 2.

Appendix C

Following are examples taken or adapted from the ESS Round 4 (European Social Survey, 2010) that could be included in a coordinating center or data collection agency study rules to demonstrate instrument technical specifications for different information and question types. These can be applicable to either paper or computerized instruments.

Instrument Overview

	Q#	Topics
Core	A1 –A10	Media; social trust
Core	B1 – B40	Politics, including: political interest, efficacy, trust, electoral and other forms of participation, party allegiance, socio-political orientations
Core	C1 – C36	Subjective well-being, social exclusion; religion; perceived discrimination; national and ethnic identity
Rotating module	D1-D50	Welfare includes attitudes towards welfare provision, size of claimant groups, views on taxation, attitudes towards service delivery and likely future dependence on welfare.
Rotating module	E1-E55	Ageism covers attitudes towards and experiences of ageism, age related status, stereotypes, experience of discrimination and contact with people in other age groups.
Core	F1 – F73	Socio-demographic profile, including: household composition, sex, age, type of area, education & occupation of respondent, partner, parents, union membership, income, marital status
Supplementary	Section G	Human values scale
Supplementary	Section H	Test questions
Interviewer questionnaire	Section I	Interviewer self-completion questions

Missing Value Definitions

Not applicable: 6, 66, 666 etc.	respondent has been routed away from the question
Refusals: 7, 77, 777 etc.	respondent has explicitly refused
Don't know: 8, 88, 888 etc.	respondent has explicitly said "don't know"
No answer: 9, 99, 999 etc.	Missing data not elsewhere explained

Common Question types

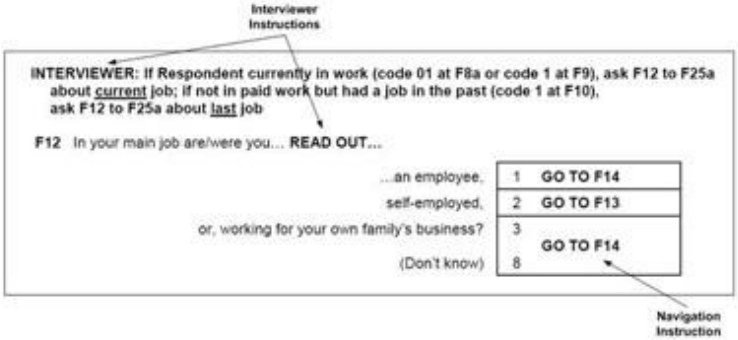
1. Interviewer checkpoints:

- With masks showing date format separators [❑❑ /❑❑ /❑❑] and response format instructions [(dd/mm/yy)]:

INTERVIEWER ENTER START DATE: ❑❑ /❑❑ /❑❑ (dd/mm/yy)

INTERVIEWER ENTER START TIME: ❑❑❑❑ (Use 24 hour clock)

2. Closed question with enumerated response options:

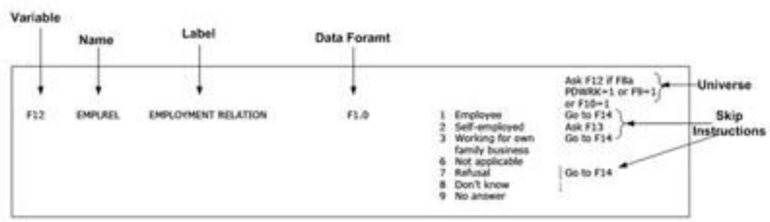


INTERVIEWER: If Respondent currently in work (code 01 at F8a or code 1 at F9), ask F12 to F25a about current job; if not in paid work but had a job in the past (code 1 at F10), ask F12 to F25a about last job

F12 In your main job are/were you… READ OUT…

…an employee,	1	GO TO F14
self-employed,	2	GO TO F13
or, working for your own family's business?	3	GO TO F14
(Don't know)	8	

- Data dictionary elements for question F12 [variable ID F12; variable name EMPLREL; variable label EMPLOYMENT RELATION; one-digit integer format with zero decimal places; universe (Ask F12 if F8a PDWORK = 1 or F9=1); response options and codes; and skip instructions]:



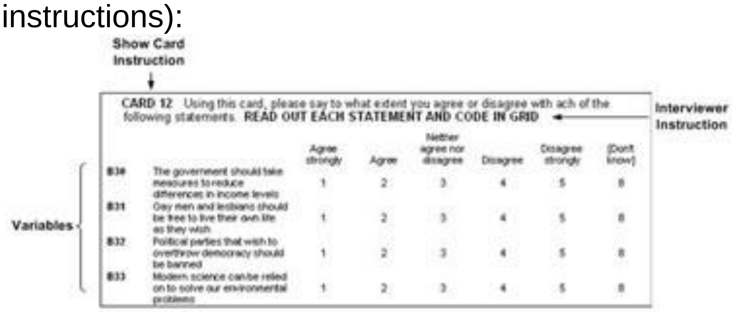
F12 EMPLREL EMPLOYMENT RELATION F1.0

Ask F12 if F8a PDWRK=1 or F9=1 or F10=1

1 Employee — Go to F14
2 Self-employed — Ask F13
3 Working for own family business — Go to F14
6 Not applicable
7 Refusal
8 Don't know — Go to F14
9 No answer

3. Scale Questions in Grid:

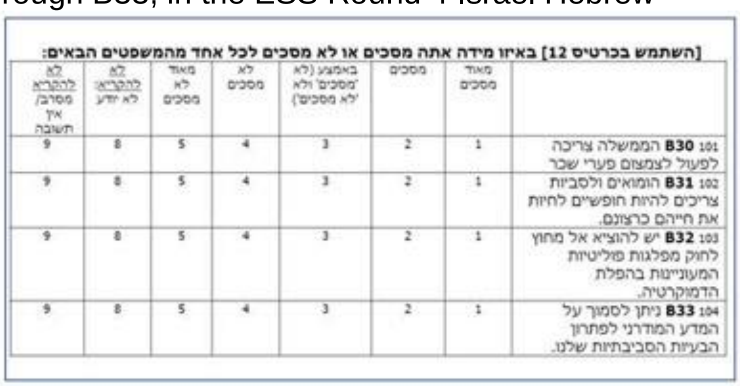
- Questions B30 through B33 [show card (CARD 12) and interviewer instructions):



CARD 12 Using this card, please say to what extent you agree or disagree with ach of the following statements. READ OUT EACH STATEMENT AND CODE IN GRID

		Agree strongly	Agree	Neither agree nor disagree	Disagree	Disagree strongly	(Don't know)
B30	The government should take measures to reduce differences in income levels	1	2	3	4	5	8
B31	Gay men and lesbians should be free to live their own life as they wish	1	2	3	4	5	8
B32	Political parties that wish to overthrow democracy should be banned	1	2	3	4	5	8
B33	Modern science can be relied on to solve our environmental problems	1	2	3	4	5	8

- Questions B30 through B33, in the ESS Round 4 Israel Hebrew questionnaire:



[השתמש בכרטיס 12] באיזו מידה אתה מסכים או לא מסכים לכל אחד מהמשפטים הבאים:

	מאוד מסכים	מסכים	באמצע (לא 'מסכים' ולא 'לא מסכים')	לא מסכים	מאוד לא מסכים	לא להקריא: לא יודע	לא להקריא: מסרב/ אין תשובה
101 B30 הממשלה צריכה לפעול לצמצום פערי שכר	1	2	3	4	5	8	9
102 B31 הומואים ולסביות צריכים להיות חופשיים לחיות את חייהם כרצונם.	1	2	3	4	5	8	9
103 B32 יש להוציא אל מחוץ לחוק מפלגות פוליטיות המעוניינות בהפלת הדמוקרטיה.	1	2	3	4	5	8	9
104 B33 ניתן לסמוך על המדע המודרני לפתרון הבעיות הסביבתיות שלנו.	1	2	3	4	5	8	9

- Show card (CARD 12, used for questions B30 through B33):

Question(s) B30, B31, B32, B33

CARD 12

Agree strongly
Agree
Neither agree nor disagree
Disagree
Disagree strongly

- Show card (CARD 12) in the ESS Round 4 Israel Hebrew questionnaire:

כרטיס 12

מאוד לא מסכים	לא מסכים	באמצע (לא 'מסכים' ולא 'לא מסכים')	מסכים	מאוד מסכים
5	4	3	2	1

- Data dictionary (data protocol) for scale questions in grid [variables B30 through B33; variable names GINCDIF, FREEHMS, PRTYBAN, SCNSENV; variable labels (e.g., GOVERNMENT SHOULD REDUCE DIFFERENCES IN INCOME LEVELS); single-digit integer with no decimal places; universe; response options and codes]:

						B30-B33: Ask All
B30	GINCDIF	GOVERNMENT SHOULD REDUCE DIFFERENCES IN INCOME LEVELS	F1.0	1	Agree strongly	B30-B33: Same format, values and Categories
				2	Agree	
B31	FREEHIS	GAYS AND LESBIANS FREE TO LIVE LIFE AS THEY WISH		3	Neither agree nor disagree	
B32	PRTYBAN	BAN POLITICAL PARTIES THAT WISH OVERTHROW DEMOCRACY		4	Disagree	
				5	Disagree strongly	
B33	SCNSENV	MODERN SCIENCE CAN BE RELIED ON TO SOLVE ENVIRONMENTAL PROBLEMS		7	Refusal	
				8	Don't know	
				9	No answer	

4. Country-Specific Questions

- ESS highlights country-specific questions in gray in the source questionnaire specifications, for example, variable B12:

B11 Some people don't vote nowadays for one reason or another.
Did you vote in the last [country] national[#] election in [month/year]?

Yes	1	**ASK B12**
No	2	
Not eligible to vote	3	**GO TO B13**
(Don't know)	8	

IF YES AT B11 (code 1)

B12 Which party did you vote for in that election?
[Country-specific (question and) codes]

Conservative	01
Labour	02
Liberal Democrat	03
Scottish National Party	04
Plaid Cymru	05
Green Party	06
Other (**WRITE IN**) ______________	07
(Refused)	77
(Don't know)	88

- Country-specific question B12 in the ESS Round 4 Latvian questionnaire:

B11	**Mūsdienās cilvēki tādu vai citu iemeslu dēļ mēdz nepiedalīties vēlēšanās. Vai Jūs balsojāt pēdējās Saeimas vēlēšanās 2006. gada oktobrī?**	Jā	1
		Nē	2 →B13
		Nav balsstiesību	3 →B13
		GP	8 →B13
B12	*JA B11 TIKA ATBILDĒTS JĀ (B11=1)* **Par kuru partiju Jūs balsojāt šajās Saeimas vēlēšanās?**	Tautas partija	01
		Zaļo un Zemnieku savienība	02
		Jaunais laiks	03
		"Saskaņas Centrs"	04
		Latvijas Pirmās partijas un partijas "Latvijas Ceļš" vēlēšanu apvienība	05
		Apvienība "Tēvzemei un Brīvībai"/LNNK	06
		Politisko organizāciju apvienība "Par cilvēka tiesībām vienotā Latvijā"	07
		Par citu partiju (*IERAKSTIET*)	08
		Atteicās atbildēt	77
		GP	88

- Data dictionary (variables B11 and B12; variable names VOTE and PRTVTxx; variable labels; one- and two-digit integer formats; response options and codes; universes; and skip instructions):

B11	VOTE	VOTED LAST NATIONAL ELECTION	F1.0	1 Yes 2 No 3 Not eligible to vote 7 Refusal 8 Don't know 9 No answer	B11: Ask all Ask B12 Go to B13
B12	PRTVTxx	PARTY VOTED FOR IN LAST NATIONAL ELECTION, [COUNTRY]	F2.0	66 Not applicable 77 Refusal 88 Don't know 99 No answer	Ask B12 if B11=1 B12: Country-specific question, see section E.1.1
	PRTVxx1 PRTVxxN	PARTY VOTED FOR IN LAST NATIONAL ELECTION 1, [COUNTRY] PARTY VOTED FOR IN LAST NATIONAL ELECTION N, [COUNTRY]	F2.0	66 Not applicable 77 Refusal 88 Don't know 99 No answer	

Appendix D

Figures D1 through D3, taken from Dutch questionnaires of the International Social Survey Programme (ISSP), shows how different visual scales might result in distinctive response distributions. In 1987, the Dutch questionnaire used the scale that displayed a truncated pyramid, while other countries used a scale with 10 vertically stacked squares. As a result, the response distribution from the Dutch question differed from that of other countries and did not correlate well with other Dutch measures (Smith, 1993). The Dutch scale for the social ladder question was later changed to more closely resemble the visual display used by other countries (Figure D3). Figures D4 and D5 show differences in graphics used for a body shape question in the ISSP 2007 Austrian and Philippines questionnaires.

Figure D1. Social ladder in ISSP 1987 Dutch questionnaire

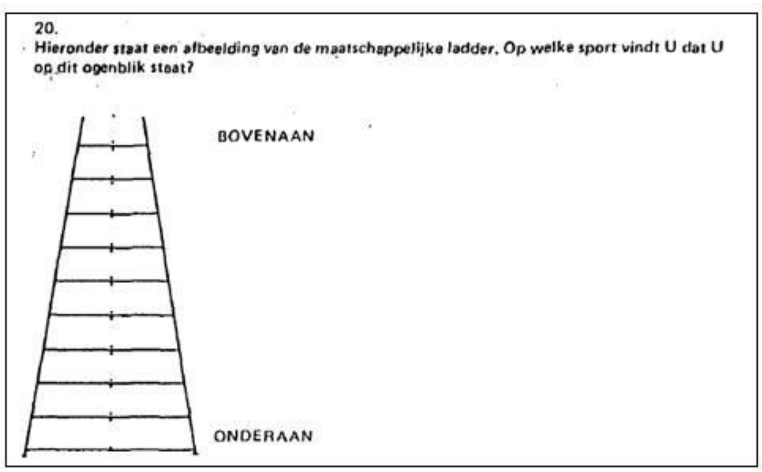


Figure D2. Social ladder in ISSP 1987 Great Britain questionnaire

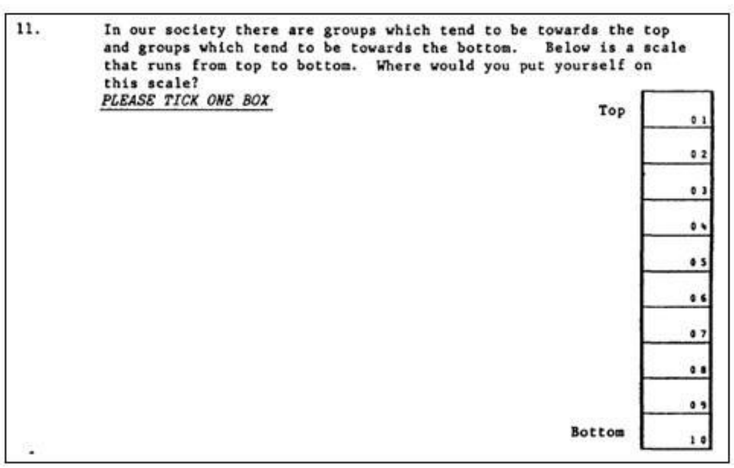


Figure D3. Social ladder in ISSP 2004 Dutch questionnaire

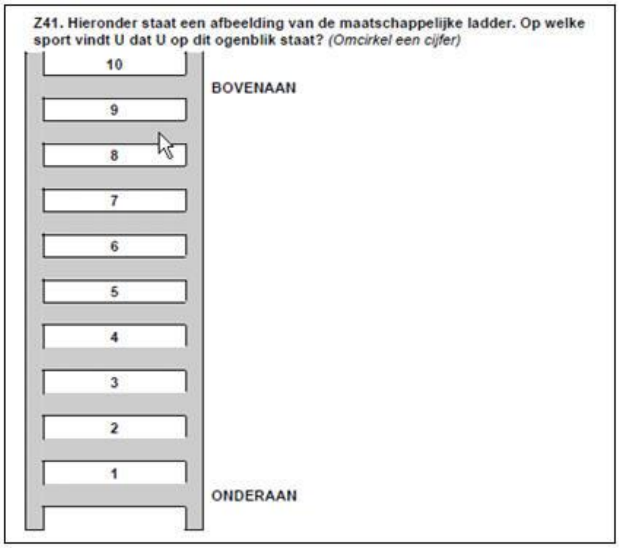


Figure D4. Ideal shape question in ISSP 2007 Austrian survey

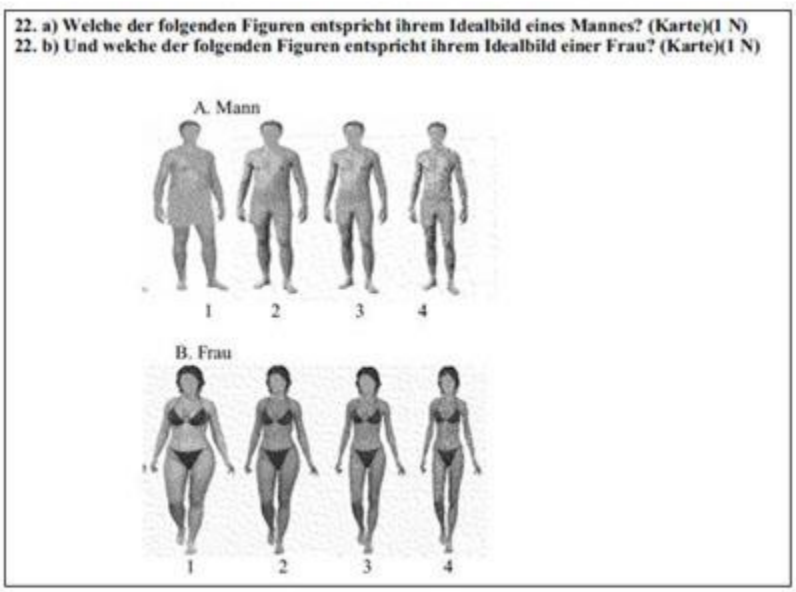
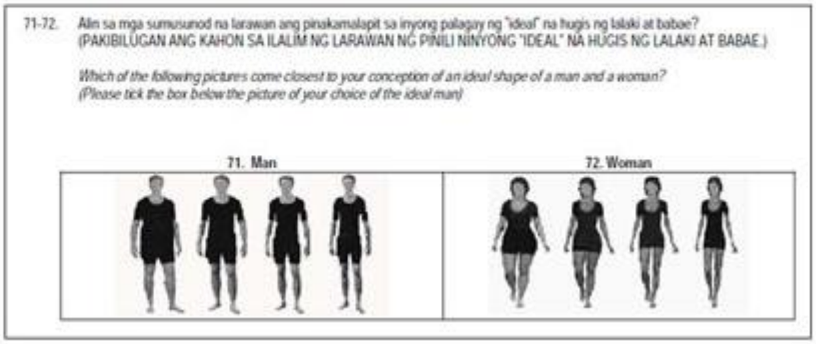


Figure D5. Ideal shape question in ISSP 2007 Philippines survey



71-72. Alin sa mga sumusunod na larawan ang pinakamalapit sa inyong palagay ng "ideal" na hugis ng lalaki at babae?
(PAKIBILUGAN ANG KAHON SA ILALIM NG LARAWAN NG PINILI NINYONG "IDEAL" NA HUGIS NG LALAKI AT BABAE.)

Which of the following pictures come closest to your conception of an ideal shape of a man and a woman?
(Please tick the box below the picture of your choice of the ideal man)

71. Man

72. Woman

Appendix E

Following are examples of text direction used by various countries in Asia, taken from the East Asia Barometer (EAB) survey in 2006 and the ISSP in 2007. These suggest both that design guidelines need to reflect an understanding of how different Asian countries display text both vertically and horizontally, and that it would be desirable to pretest separately questions that differ across countries.

Figure E1. The 2006 EAB Singapore questionnaire: horizontal response option column headers read from top to bottom

F. ACCESS TO PUBLIC SERVICE 公共服务

44-47. Based on your experience, how easy or difficult is it to obtain the following services? Or do you never try and get these services from government? (**Do not read**: Can't choose & Decline to answer)
根据您的经验，请问您觉得下列政府提供的公共服务，容不容易获得？或者，您从未获得下列服务？

(SHOWCARD) 【访员出示卡片】	Very Easy 非常 容易	Easy 容易	Difficult 困难	Very Difficult 非常 困难	Never Try 从来 没有	Can't choose 无法 选择	Decline to answer 不回答
44. An identity document (such as a birth certificate or passport) 申办证件服务（例如身份证、护照等）	4	3	2	1	5	8	9
45. A place in public primary school for a child 为小孩申请入学	4	3	2	1	5	8	9
46. Medical treatment at a nearby clinic 医疗服务（在附近的医疗诊所看病）	4	3	2	1	5	8	9
47. Help from the police when you need it 要求警察帮助与服务	4	3	2	1	5	8	9

Figure E2. 2006 EAB Taiwan questionnaire: vertical response option column headers

F. 公共服務

44-47. 根據您的經驗，請問您覺得下列政府提供的公共服務，容不容易獲得？(或者，您從未獲得下列服務)

	很容易	容易	困難	很困難	從來沒有	不適用	無法選擇	不回答	
	4.	3.	2.	1.	5.	0.	8.	9.	
(44) 申辦證件服務 (例如戶籍謄本、護照等)	□	□	□	□	□	□	□	□	□68
(45) 為小孩申請入學	□	□	□	□	□	□	□	□	□69
(46) 醫療服務 (去附近的醫療診所看病)	□	□	□	□	□	□	□	□	□70
(47) 要求警察幫助與服務	□	□	□	□	□	□	□	□	□71

Figure E3. The 2006 EAB Mainland China questionnaire: vertical response option column headers, read from right to left

I5. 基于您以往的经验，您觉得获得下列服务有困难吗？您曾经尝试从政府机构获得类似的服务吗？

	非常容易	容易	困难	很困难	从没试过	[illegible] [不读]	[illegible] [不读]
I5a. 办理身份证、出生证、护照	4	3	2	1	5	8	9
I5b. 孩子在公立学校注册上学	4	3	2	1	5	8	9
I5c. 在附近医院看病或拿药	4	3	2	1	5	8	9
I5d. 在需要的时候获得警察的帮助	4	3	2	1	5	8	9

Figure E4. The 2007 ISSP Japan questionnaire: vertical response option column headers, read from left to right

問3 あなたは、次にあげるA～Dの余暇活動を、どのくらい楽しんでいますか。最もあてはまるものに**1つだけ**○をつけてください。（**○はそれぞれ1つずつ**）

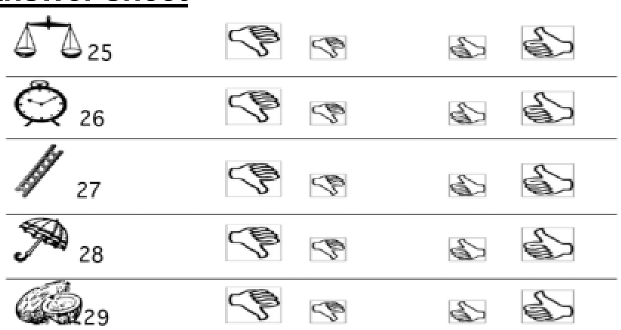
	1 まったく楽しんでいない	2 あまり楽しんでいない	3 まあ楽しんでいる	4 かなり楽しんでいる	5 非常に楽しんでいる	6 していない
A. 読書をる →	1	2	3	4	5	6
B. 友人とう →	1	2	3	4	5	6
C. 運動をする（スポーツをする、ジムに行く、散歩をする →	1	2	3	4	5	6
D. テレビ、DVD、ビデオを見る →	1	2	3	4	5	6

Appendix F

Images can be used to facilitate the self-administered questionnaire mode for respondents in low-literacy settings. A survey on sensitive caste-related attitudes in rural India followed the protocol below (Chauchard, 2013):

1. Respondents were provided a basic MP3 player and headphones.
2. The audio track on the MP3 presented them with a number of first-person statements made by “respondents like [them].”
3. Respondents entered responses on an answer sheet using simple shapes and logos and placed their form in a bolted ballot box to enhance privacy (Lowndes, et al., 2012).
4. Each question number was designated by an image (e.g., scale, clock, etc.).
5. The prerecorded voice uses these images to indicate to the respondent which question to answer.
6. Instructions from interviewer and prerecorded voice detail what each “thumb” means: Clearly disagree, somewhat disagree, somewhat agree, clearly agree.

Figure F1: Images are used as question label and response options on the answer sheet



Appendix G

Identifying relevant keyboard buttons with colors can reduce respondent burden if using CASI or A-CASI. The following protocol was used in an A-CASI survey of sensitive sexual and illegal behaviors in rural Kenya and Malawi (Hewett, Mensch & Erulkar, 2004; Mensch, et al., 2008; Rathod, Minnis, Subbiah, & Krishnan, 2011; Bhatnagar et al., 2013):

1. The survey instrument was programmed and run from a laptop, with external mini-keyboard and headphones connected to laptop
2. Respondents used the audio headphones to listen to the questions and response options
3. Respondents entered responses via an external mini-keypad linked to the A-CASI program on the laptop
4. Keyboard had numbers and color-coded keys for replaying and entering options; majority of questions dichotomous
5. Respondents were able to use A-CASI with minimal assistance

Figure G1: Color-coded keys are used for an A-CASI survey of sensitive sexual and illegal behaviors

Appendix H

As part of carrying out the winning research proposals, participants from 2014 American Association for Public Opinion Research (AAPOR) ResearchHack 1.0 (RH 1.0) helped execute a series of foundational research projects to prove the concept and offer a methodological solution to a real-world challenge for Feeding America and their nationwide network of feeding programs (AAPOR ResearchHack, 2015; Brittany 2015).

This study was to assess whether the Instagram (IG) app is a viable data collection tool to learn how individuals in need of food assistance find feeding programs within the Feeding America network. The proof-of-concept evaluated whether the IG app and data collectors can effectively collect the data points, including photos of clients' food items, service location, hashtags, and clients' general or specific comments, using the IG features (mainly mages, hashtags, location) at one selected feeding program (Lin, Morgan, & Lomelinol, 2015).

A usability study was conducted and gave insights on potential technical and methodological issues for using Instagram as a data collection tool (Kelley, Krishna, & Lai, 2015). More specifically, the usability study answered the following research questions:

1. Do users understand what all the buttons on the bottom of the Instagram home screen are for?
2. Do users know how to upload and share a photo on Instagram?
3. What potential usability issues would users experience when writing the captions, tagging people or places, and identifying hashtags for the data collection task?

A pilot data collection was conducted in five Feeding America food agencies. The study trained food pantry volunteers how to use the Instagram app to collect the responses of food pantry recipients. Using a "show card" with the survey questions and answer options written out, respondents selected their answer which was converted to a hashtag. Below are two example questions the food pantry volunteers used photo and hashtags to record respondents answers.

- Question #1 asked: "Out of all the food you received (or being offered) today, is there a food item you will not use or cannot eat?"

  To respond to this question, data collectors took a photo of the food item.

- Question #2 posed: "Why not?" and offered the following reasons:
  a. Food allergy or intolerance
  b. Dietary restriction
  c. Religious dietary practice
  d. Not familiar with, don't know how to cook or don't have equipment

e. Don't like or children don't like
f. Nothing (will use everything)
g. Not healthy
h. Other

For example, if a recipient answered that they did not take peanut butter because they are allergic to peanuts, the volunteer uploaded a photo of the peanut butter with the hashtag #q2d to illustrate it was answering the second question using the first answer option (d).

Figure F1 shows an example that the food pantry volunteer logged in Instagram App using “rhfa2015” account and conducted interview at “Glen Ellyn Food Pantry.” The food pantry volunteer recorded that the respondent will not use or cannot eat mushroom (photo) because the respondent is not familiar with this item, don't know how to cook it, or don't have equipment to cook it (#q2d). Figure 1 also shows the food pantry volunteer asked another question (#q3a) and recorded his/her first and last name initials (#db).

Figure H1: An example of using Instagram App to collect data in photo and hashtag format

The study also provided the instructional materials for data extract using the Instagram API. Please see the AAPOR ResearchHack 1.0 Web site (http://aaporresearchhack.tumblr.com/).

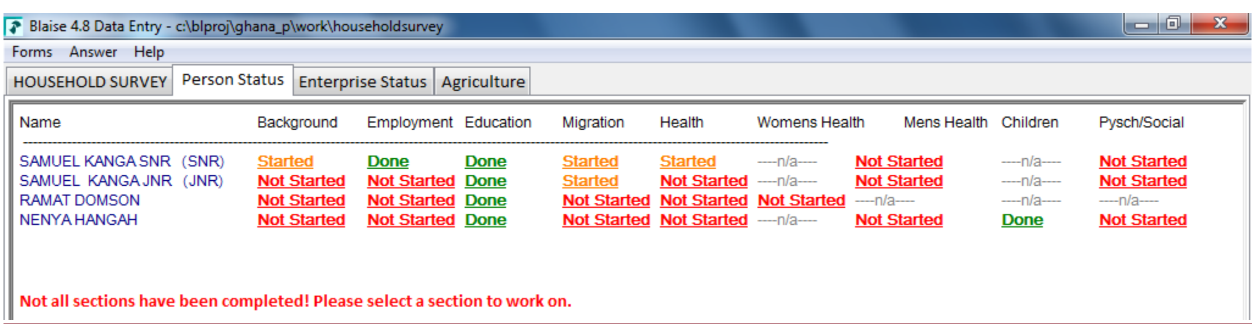
Appendix I – Blaise parallel blocks

Given the complexity of interviewing dynamics and survey topics, the EGC-ISSER Ghana Panel Survey instrument was programmed into parallel blocks with separate tabs (Household, Person, Enterprise, Agriculture tabs) and matrix hyperlinks to different instrument sessions and displayed as interviewing status dashboard with color coding (Not Started, Started, Done, Not Applicable). With the designed instrument layout, interviewers had flexibility to jump in/out from different sections within the same instrument depending on the availability of the respondents and were able to track the real-time status of interviewing progress on multiple respondents within the same instrument (Kwaiser, Williams, Cheung, & Lin, 2015).

Figure I1 shows an example of the personal tab within the EGC-ISSER Ghana Panel Survey instrument. The selected household has four eligible respondents and the interviewing progress of different sessions to be asked by respondent. For example, the interviewer has interviewed Samuel Kanga Sr. on most sessions within this tab. Among all eligible sessions, Employment and Education sessions were done, Background, Migration, Health sessions are in progress, Psychology and Social sessions are not started yet, and Woman Health and Children sessions are not eligible for this respondent. With this color coding, interviewers can have a quick idea about the overall interviewing status.

Given the complexity of the study instrument and on average it took more than 5 hours average interview completion time per household, the project staff used key stroke data recorded by the Blaise software to analyze interviewers' instrument navigation behavior throughout the whole interview and came up the optimal instrument interface design with suggested interviewing order among multiple respondents within the same household (Kwaiser et al., 2015).

Figure I1: Personal tab instrument was programmed with parallel blocks format among all household respondents



References

American Association for Public Opinion Research (AAPOR). (2015). ResearchHack 2015. http://aaporresearchhack.tumblr.com/rhfa2015

Aykin, N. (Ed.). (2005). *Usability and internationalization of information technology*. Mahwah, NJ: Lawrence Erlbaum Associates.

Aykin, N., & Milewski, A. E. (2005). Practical issues and guidelines for international information display. In N. Aykin (Ed.), *Usability and internationalization of information technology* (pp. 21-50). Mahwah, NJ: Lawrence Erlbaum Associates.

Bhatnagar, T., Brown, J., Saravanamurthy, P. S., Kumar, R. M., & Detels, R. (2013). Color-coded audio computer-assisted self-interviews (C-ACASI) for poorly educated men and women in a semi-rural area of south India: "Good, scary and thrilling". *AIDS and Behavior, 17*(6), 2260-2268.

Brittany, M. (2015). *Leveraging social media to help better feed the people in need*. Huffpost Impact Blog (http://www.huffingtonpost.com/brittany-morgan/leveraging-social-media-t_1_b_8136412.html)

Byass, P., Hounton, S., Ouédraogo, M., Somé, H., Diallo, I., Fottrell, E., & Meda, N. (2008). Direct data capture using hand held computers in rural Burkina Faso: Experiences, benefits and lessons learnt. *Tropical Medicine & International Health, 13*(s1), 25-30.

Chauchard, S. (2013). Using MP3 players in surveys: The impact of a low-tech self-administration mode on reporting of sensitive attitudes. *Public Opinion Quarterly*, *77*(S1), 220-231.

Christian, L. M., Dillman, D. A., & Smyth, J. D. (2005). Helping respondents get it right the first time: The influence of words, symbols, and graphics in Web surveys. *Public Opinion Quarterly, 71*(1), 1-13.

Clemmensen, T., & Goyal, S. (2005). *Cross-cultural usability testing—The relationship between evaluator and test user* (Working paper no. 06-2005). Frederiksberg Municipality, Denmark: Department of Informatics, Copenhagen Business School.

Couper, M. P. (1999). The application of cognitive science to computer-assisted interviewing. In M. G. Sirken, D. J. Herrmann, S. Schechter, N. Schwarz, J. M. Tanur, & R. Tourangeau (Eds.), *Cognition and survey research*. New York, NY: John Wiley & Sons.

Couper, M. P. (2000). Usability evaluation of computer-assisted survey instruments. *Social Science Computer Review*, *18*(4), 384-396.

Couper, M. P. (2001). *Web surveys: The questionnaire design challenge*. Paper presented at the proceedings of the 53rd session of the International Statistical Institute.

Couper, M. P. (2008). *Designing effective Web surveys*. New York, NY: Cambridge University Press.

Couper, M. P., Beatty, P., Hansen, S. E., Lamias, M., & Marvin, T. (2000). *CAPI design recommendations*. Report submitted to the Bureau of Labor Statistics.

Couper, M. P., Traugott, M. W., & Lamias, M. J. (2001). Web survey design and administration. *Public Opinion Quarterly, 65*(2), 230-253.

De Jong, J., & Young-DeMarco, L. (2017). Best practices: Lessons from a Middle East survey research program. In M. Gelfand & M. Moaddel (Eds.), *The Arab spring and changes in values and political actions in the Middle East: Explorations of visions and perspective*. Oxford, U. K.: Oxford University Press.

de Leeuw, E., Hox, J., & Kef, S. (2003). Computer-assisted self-interviewing tailored for special populations and topics. *Field Methods, 15*(3), 233-251.

Dillman, D. A., Gertseva, A., & Mahon-Haft, T. (2005). Achieving usability in establishment surveys through the application of visual design principles. *Journal of Official Statistics, 21*, 183-214.

Dillman, D. A., Smyth, J. D., & Christian, L. M. (2009). *Internet, mail, and mixed-mode surveys: The tailored design method*. (3rd ed.). New York, NY: John Wiley & Sons.

Dorer, B. (2012) *Round 6 translation guidelines*. Mannheim: European Social Survey, GESIS. Retrieved July 2, 2014, http://www.europeansocialsurvey.org/docs/round6/methods/ESS6_translation_guidelines.pdf .

EGC-ISSER Ghana Panel Survey – http://www.econ.yale.edu/~egcenter/egc_isser_overview.html

Escandon, I. N., Searing, H., Goldberg, R., Duran, R., & Monterrey Arce, J. (2008). The use of PDAs to collect baseline survey data: Lessons learned from a pilot project in Bolivia. *Global Public Health, 3*(1), 93-104.

European Social Survey (ESS). (2010). Retrieved May 03, 2010, from http://www.europeansocialsurvey.org

Groves, R. M., & Mathiowetz, N. A. (1984). Computer assisted telephone interviewing: Effects on interviewers and respondents. *Public Opinion Quarterly, 48*(1B), 356-369.

Hansen, S. E., & Couper, M. P. (2004). Usability testing as a means of evaluating computer assisted survey instruments. In S. Presser, J. M. Rothgeb, M. P. Couper, J. T. Lessler, E. Martin, J. Martin, & E. Singer (Eds.), *Methods for testing and evaluating survey questionnaires*. New York, NY: John Wiley & Sons.

Heerwegh, D. (2003). Explaining response latencies and changing answers using client-side paradata from a web survey. *Social Science Computer Review, 21*(3), 360-373.

Hert, C. A. (2001). Studies of metadata creation and usage. In *Proceedings of the Federal Committee on Statistical Methodology*. Washington, DC: U.S. Department of Commerce.

Hewett, P. C., Mensch, B. S., & Erulkar, A. S. (2004). Consistency in the reporting of sexual behaviour by adolescent girls in Kenya: A comparison of interviewing methods. *Sexually Transmitted Infections, 80*(suppl 2), ii43-ii48.

Holzberg, J. L., & Eggleston, C. M. (2016). *Implementing bring your own device in a survey organization*. Paper presented at the American Association for Public Opinion Research, Austin, Texas..

House, C. C., & Nicholls, W. L. (1988). Questionnaire design for CATI: Design objectives and methods. In R. M. Groves (Ed.), *Telephone survey methodology* (pp. 421-436). New York, NY: John Wiley Company,

Jagne, J., & Smith-Atakan, A. S. G. (2006). Cross-cultural interface design strategy. *Universal Access in the Information Society*, *5*(3), 209-305.

Kelley, J., Krishna, K. S., & Lai, J. (2015, May 14-17). *Designing a new data capture method: Usability study of the Instagram app as a data collection tool*. Presentation given at 2015 American Association for Public Opinion Research (AAPOR) Conference, Hollywood, FL, USA.

Kondratova, I., & Goldfarb, I. (2007). Color your website: Use of colors on the web. In N. Aykin (Ed.), *Usability and internationalization: Global and local user interfaces* (pp. 123-132). Berlin, Germany: Springer.

Kwaiser, K., Williams, J, Cheung, G., & Lin, Y. (2015, March 26-28). *Using paradata to monitor interviewers' behavior: A case study from a national household survey in Ghana.* Presentation given at the 2015 International Workshop on Comparative Survey Design and Implementation (CSDI), London, United Kingdom.

Lara, D., Strickler, J., Olavarrieta, C. D., & Ellertson, C. (2004). Measuring induced abortion in Mexico: A comparison of four methodologies. *Sociological Methods & Research, 32*(4), 529-558.

Lau, C., Lombaard, A., Baker, M., Eyerman, J., & Thalij, L. (2016). *The validity of collecting data using short messge service (SMS): Experimental evidence from four African countries*. Paper presented at the American Association for Public Opinion Research, Austin, Texas.

Lin, Y., Morgan, B., & Lomelino, L. (2015, May 14-17). *Instagram app for data collection: Can images, hashtags and location tagging be the new data formats?* Presentation at the 2015 American Association for Public Opinion Research (AAPOR) conference, Hollywood, FL.

Lowndes, C. M., Jayachandran, A. A., Banandur, P., Ramesh, B. M., Washington, R., Sangameshwar, B. M., … Alary, M. (2012). Polling booth surveys: A novel approach for reducing social desirability bias in HIV-related behavioral surveys in resource-poor settings. *AIDS & Behavior*, *16,* 1054–1062.

Mensch, B. S., Hewett, P. C., Gregory, R., & Helleringer, S. (2008). Sexual behavior and STI/HIV status among adolescents in rural Malawi: An evaluation of the effect of interview mode on reporting. *Studies in Family Planning, 39*(4), 321-334.

Murphy, E. D., Nichols, E. M., Anderson, A. E., Harley, M. D., & Pressley, K. D. (2001). *Building usability into electronic data-collection forms for economic censuses and surveys*. Paper presented at the Federal Committee on Statistical Methodology Research Conference, Arlington, VA.

National Institute of Mental Health Collaborative HIV/STD Prevention Trial Group. (2007). The feasibility of audio computer-assisted self-interviewing in international settings. *AIDS, 21*(S49-S58).

Onono, M., Carraher, N., Cohen, R., Bukusi, E., & Turan, J. (2011). Use of personal digital assistants for data collection in a multi-site AIDS stigma study in rural south Nyanza, Kenya. *African Health Sciences, 11*(3), 464–473.

Peytchev, A., Couper, M. P., McCabe, S. E., & Crawford, S. D. (2006). Web survey design. *Public Opinion Quarterly*, *70*(4), 596-607.

Rathod, S. D., Minnis, A. M., Subbiah, K., & Krishnan, S. (2011). ACASI and face-to-face interviews yield inconsistent estimates of domestic violence among women in India: The Samata Health Study 2005-2009. *Journal of Interpersonal Violence*, *26*(12), 2437-2456.

Russo, P., & Boor, S. (1993). *How fluent is your interface? Designing for international users*. In Proceedings of the INTERACT '93 and CHI '93 conference on Human Factors in Computing Systems. New York, NY: Association for Computing Machinery.

Smith, T. W. (1993). *Little things matter: A sampler of how differences in questionnaire format can affect survey responses*. GSS Methodological Report No. 78. Chicago: NORC.

Smith, T. W. (2005). The laws of studying societal change. *Survey Research*, *36*(2), 1-5.

Sonck, N., & Fernee, H. (2013). *Using smartphones in survey research: A multifunctional tool. Implementation of a time use app; A feasibility study*. Netherlands Institute for Social Research.

Sun, X., & Shi, Q. (2007). Language issues in cross cultural usability testing: A pilot study in China. In N. Aykin (Ed.), *Usability and internationalization: Global and local user interfaces* (pp. 274-284). Berlin, Germany: Springer.

Survey Research Center. (2007). *SRC Blaise standards*. Ann Arbor, MI: The University of Michigan.

Tanzer, N. K. (2005). Developing tests for use in multiple languages and cultures: A plea for simultaneous development. In R. K. Hambleton, P. F. Merenda, & C. C. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment*. Mahwah, NJ: Lawrence Erlbaum Associates.

Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin, 133*(5), 859-883.

U.S. Census Bureau (2012). *United States Census 2020: Strategic plan for the 2020 Census*. November 28, 2012, Version 3.0.

Vatrapu, R., & Pérez-Quiñones, M. A. (2006). Culture and usability evaluation: The effects of culture in structured interviews. *Journal of Usability Studies*, *1*(4), 156-170.

Wan, X., Raymond, H. F., Wen, T., Ding, D., Wang, Q., Shin, S. S., ... & Novotny, T. E. (2013). Acceptability and adoption of handheld computer data collection for public health research in China: A case study. *BMC Medical Informatics and Decision Making*, 13(1), 1.

West, B. T., Ghimire, D., & Axinn, W. G. (2015). Evaluating a modular design approach to collecting survey data using text messages. In *Survey Research Methods* (Vol. 9, No. 2, p. 111). NIH Public Access.

Zurovac, D., Sudoi, R. K., Akhwale, W. S., Ndiritu, M., Hamer, D. H., Rowe, A. K., & Snow, R. W. (2011). The effect of mobile phone text-message reminders on Kenyan health workers' adherence to malaria treatment guidelines: A cluster randomised trial. *The Lancet, 378*(9793), 795-803.

Translation: Overview

Peter Mohler, Brita Dorer, Julie de Jong, and Mengyao Hu, 2016
(2010 Version: Janet Harkness, Brita Dorer, and Peter Ph. Mohler)

Introduction

Translation is the process of expressing the sense of words or phrases from one language into another. It is also known as one type of asking the same questions and translating (ASQT) as discussed in Questionnaire Design (where we also discuss asking different questions (ADQ) and its correspondence to Adaptation). Another type of ASQT is decentering (See Questionnaire Design). Given that the former approach is more commonly used in cross-cultural research, in this chapter, we mainly focus on translation from one language to another.

Following terminology used in the translation sciences, this chapter distinguishes between "source languages" used in "source questionnaires" and "target languages" used in "target questionnaires." The language translated out of is the source language; the language translated into is the target language.

Translation procedures play a central and important role in multinational, multicultural, or multiregional surveys, which we refer to as “3MC” surveys. Although good translation products do not assure the success of a survey, badly translated questionnaires can ensure that an otherwise sound project fails because the poor quality of translation prevents researchers from collecting comparable data.

The guidelines in Translation: Overview provide an overview of the translation process. In addition, there are six other sets of guidelines focusing on specific aspects of the translation process: Translation: Management and Budgeting, Translation: Team, Translation: Scheduling, Translation: Shared Language Harmonization, Translation: Assessment, and Translation: Tools.

Total Survey Error (TSE) is widely accepted as the standard quality framework in survey methodology (Groves & Lyberg, 2010; Biemer, 2010; Pennell, Cibelli Hibben, Lyberg, Mohler, & Worku, 2017): “The total survey error (TSE) paradigm provides a theoretical framework for optimizing surveys by maximizing data quality within budgetary constraints. In this article, the TSE paradigm is viewed as part of a much larger design strategy that seeks to optimize surveys by maximizing total survey quality; i.e., quality more broadly defined to include user-specified dimensions of quality.” (Biemer, 2010). See Survey Quality for more information. Seen from a TSE perspective, successful translation is a cornerstone of survey quality in 3MC surveys and comparative research.

A successful survey translation is expected to do all of the following: keep the content of the questions semantically similar; keep the question format similar within the bounds of the target language; retain measurement properties, including the range of response options offered; and maintain the same stimulus (Harkness, Edwards, Hansen, Miller, & Villar, 2010). Based on growing evidence, the guidelines presented below recommend a team translation approach for survey instrument production (Harkness, 2008a; Harkness, 2008b; Harkness, Pennell, & Schoua-Glusberg, 2004; Pan & de la Puente, 2005; Willis et al., 2010). Other approaches, such as back translation, although recommended in the past, do not comply with the latest translation research.

As discussed in Questionnaire Design, there are three major approaches to questionnaire development for 3MC surveys: asking the same questions and translating (ASQT), adapt to new needs and asking different questions (ADQ), or use a mixed approach that combines ASQT and ADQ. That is to say, to design cross-culturally comparable surveys, the translation team needs to closely collaborate with other teams such as an adaptation team. See Questionnaire Design and Adaptation for more information.

The guidelines address, at a general level, the steps and protocols recommended for survey translation efforts conducted using a team approach. The guidelines and selected examples that follow are based on two principles:

- Evidence – recommendations are based on evidence from up to date literature
- Transparency – examples given should be accessible in the public domain

Many examples draw on the European Social Survey (ESS) which is the current leader in research on, and the implementation of, modern translation procedures and transparent documentation, including national datasets. Thus it serves as a model for these guidelines.

Team translation

In a team approach to survey translation, a group of people work together. Translators produce, independently from each other, initial translations, reviewers review translations with the translators, one (or more) adjudicator decides whether the translation is ready to move to detailed pretesting (See Pretesting chapter) and also decides when the translation can be considered to be finalized and ready for fielding.

Figure 1 below presents the TRAPD (Translation, Review, Adjudication, Pretesting, and Documentation) team translation model. In TRAPD, translators provide the draft materials for the first discussion and review with an expanded team. Pretesting is an integral part of the TRAPD translation development. Documentation of each step is used as a quality assurance and monitoring tool (Harkness, 2008a; Harkness, 2003; Harkness, 2007; Harkness, Villar, & Edwards, 2010a).

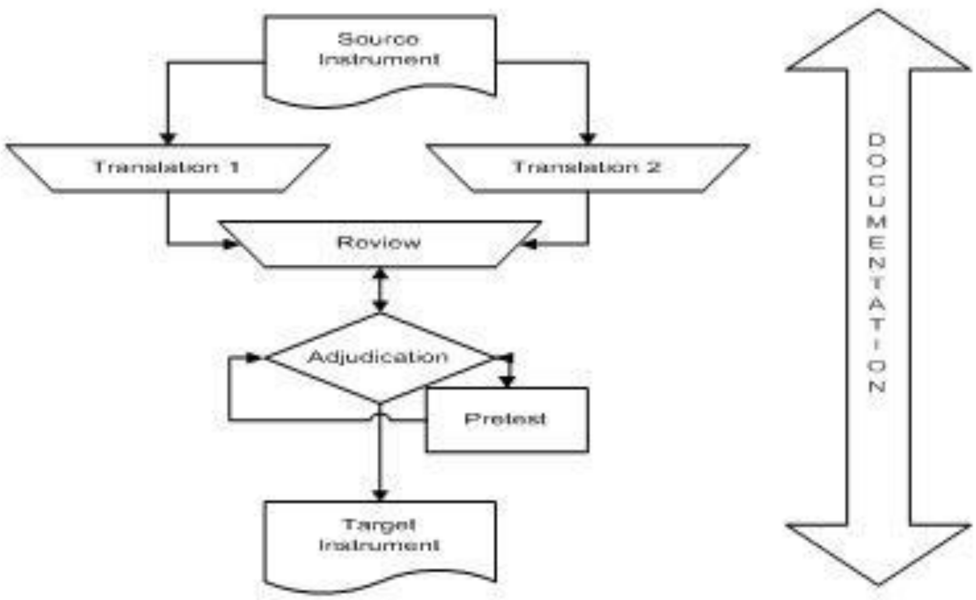


Figure 1. The TRAPD Team Translation Model

Procedures are partially iterative in team translation. The review stage reviews and refines initial parallel translation. Adjudication, often a separate step from review, can lead to further modifications of the translation before it is signed off for pretesting (see Pretesting). Pretesting may again result in modifications before the adjudicator signs off on the version for final fielding.

Team approaches to survey translation and translation assessment have been found to be particularly useful in dealing with the fairly unique challenges of survey translation. The team can be thought of as a group with different talents and functions, bringing together the mix of skills and discipline expertise needed to produce an optimal version in the survey context where translation skill alone is not sufficient. Team translation counteracts the subjective nature of translation and assessment procedures that do not deliberate translation outcomes in a professional team. In doing so team translation can achieve systematic intersubjective agreement as required in standard methodology. In addition, while providing a combined approach which is qualitatively superior, it is not a more expensive or more complicated procedure.

There are a number of other advantages to the team approach as well. The ability for each member of the translation team to document steps facilitates adjudication and provides information for secondary analysis which can inform versions for later fieldings. Additionally, the team approach allows for a considered but parsimonious production of translations which share a language with another country. All or some of these procedures may need to be repeated at different stages (see Figure 2). For example, pre-testing and debriefing sessions with fielding staff and respondents will lead to revisions; these then call for further testing of the revised translations.

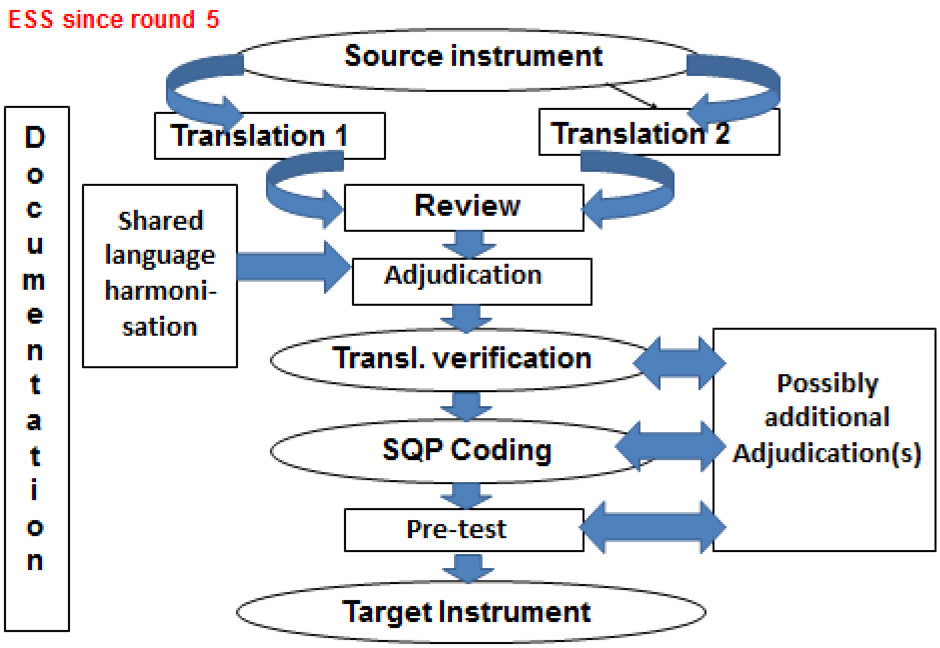


Figure 2. European Social Survey Translation Process

Team approaches to survey translation and assessment have been found to provide the richest output in terms of (a) options to choose from for translation and (b) a balanced critique of versions (Acquadro, Jambon, Ellis, & Marquis, 1996; Behr, 2009; Guillemin, Bombardier, & Beaton, 1993; Harkness & Schoua-Glusberg, 1998; McKay et al., 1996). The team should bring together the mix of skills and disciplinary expertise needed to decide on optimal versions. Collectively, members of this team must supply knowledge of the study, of questionnaire design and of fielding processes (Johnson et al., 1997; Van de Vijver & Hambleton, 1996). The team is also required to have the cultural and linguistic knowledge needed to translate appropriately in the required varieties of the target language (e.g., Acquadro et al.,1996; McKay et al.,1996). Further consideration of advantages that team efforts have over other approaches can be found in Harkness (2008a), Harkness (2008b), Harkness (2010a), Harkness et al. (2004), and Harkness & Schoua-Glusberg (1998).

Each stage of the team translation process builds on the foregoing steps and uses the documentation required for the previous step to inform the next. In addition, each phase of translation engages the appropriate personnel for that particular activity and provides them with relevant tools for the work at hand. These tools (e.g., documentation templates; see Appendix A) increase process efficiency and make it easier to monitor output. For example, translators producing the first, independent translations ('T' in the TRAPD model) are required to keep notes about any queries they have on their translations or the source text. These notes are considered along with the translation output during the next review stage in which reviewers work together with the translators (Harkness, 2008a; Harkness, 2003; Harkness, 2007).

Team translation efforts work with more than one translator. Translators produce translation material and attend review meetings. Either each translator produces a first, independent translation of the source questionnaire (double or full/parallel translation) or each translator gets parts of the source questionnaire to translate (split translation) (Harkness, 2008a; Harkness & Schoua-Glusberg, 1998; Schoua-Glusberg, 1992). The double translations or the sections of the split translation are refined in the review stage and possibly again after subsequent steps, as just described.

Whenever possible, translation efforts that follow a team approach work with more than one initial version of the translated text. A sharing of these initial versions and discussion of their merits is a central part of the review process. Two initial translations, for example, can dispel the idea that there is only one "good" or "right" translation. They also ensure that more than one translation is offered for consideration, thus enriching the review discussion. This encourages a balanced critique of versions (Acquadro et al.,1996; Harkness, 2008a; Harkness et al., 2004; McKay et al., 1996). Contributions from more than one translator also make it easier to deal with regional variance, idiosyncratic interpretations, and translator oversight (Harkness, 2008a; Harkness, 2007; Harkness, et al., 2010b).

Survey translations also often call for sensitivity for words people speak rather than words people write. Apart from ensuring the needed range of survey expertise and language expertise, the discussion that is part of team approaches (the Review session, that is 'R' in the TRAPD scheme) is more likely to reveal vocabulary or vocabulary level/style (register) problems which might be overlooked in a review made without vocalization. Pretesting may, of course, reveal further respondent needs that "experts" missed.

As noted, team-based approaches aim to include the translators in the review process. In this way, the additional cost of producing two initial/parallel translations would be offset by the considerable contributions the translators can bring to review assessments. Since they are already familiar with the translation challenges in the texts, they make the review more efficient. Split translation arrangements can still capitalize on the advantages of having more than one translator in the review discussion but avoid the cost of full or double translations.

The advantages and disadvantages of each approach are discussed under Guidelines 3 and 4 below (see also Harkness (2008a) and Schoua-Glusberg (1992)).

Guidelines

Goal: To create and follow optimal procedures to standardize, assess, and document the processes and outcomes of survey questionnaire translation.

1. Plan translation as an integral part of the study design.

This planning should include all the elements that will be part of the translation procedures (e.g., selection of team members, language harmonization/shared language arrangements), and should accommodate them in terms not only of procedural steps but with regard to hiring, training, budgeting, time schedules, and the questionnaire and translation production processes.

Rationale

Survey translation efforts are part of the target language instrument development and should be treated accordingly. In addition, when translations are produced in order to take part in a larger comparative project, forethought and a clear direction to planning and implementing translation will help produce translations across multiple locations which comply with project requirements.

Restrictions

Some surveys, such as Eurobarometer, are designed using English and French simultaneously as source languages. That procedure involves complex issues of linguistic equivalence beyond the realm of translation (Harkness, et al., 2010a).

These guidelines only refer to studies using one single source language. Studies using more than one source language would need to implement additional steps that are not discussed in these guidelines.

Procedural steps

1.1 Define the following:
- 1.1.1 The larger vision (e.g., a successfully implemented survey).
- 1.1.2 The concrete goal (e.g., a well-developed translation for the various contexts and populations).
- 1.1.3 Important quality goals (e.g., a population-appropriate translation, comparability with source questionnaire, efficiency

and feasibility of translation procedures, timeliness).

- 1.1.4 Relevant factors (e.g., schedules, budget, personnel available, unexpected events).
- 1.1.5 Tasks involved (e.g., assembling personnel and the translation documents; preparing tools, such as templates; training personnel; producing and reviewing translations; pretesting; copyediting).

1.2 Identify core team members (those people required for the team translation effort). (See Appendix B for specific tasks of each core team member and other team players identified below.)

- 1.2.1 Translators
- 1.2.2 Reviewer(s)
- 1.2.3 Adjudicator(s)

1.3 Identify any other team players who may be required, based upon the size of the project, the mode of data collection, etc.

- 1.3.1 Copyeditor(s)
- 1.3.2 Co-coordinator
- 1.3.3 Substantive experts
- 1.3.4 Programmers
- 1.3.5 Other experts, such as visual design experts, adaptation experts
- 1.3.6 External assessors
- 1.3.7 Back-up personnel

1.4 Determine whether regional variance in a language or shared languages need to be accommodated; decide on strategies for this as needed (see Translation: Shared Language Harmonization).

- 1.4.1 Select, brief, and train personnel (see Translation: Team). In training personnel, consult Appendix C (Causes of Mistranslation) for detail and examples of common causes of mistranslation. Identify the in-house and external staff and consultant needs on the project and follow appropriate selection, briefing, and training procedures for each person or group.
- 1.4.2 Identify, acquire, and prepare the materials for translation. In addition to the source questionnaire, these may include advertising material, interviewer manuals, programmer instructions, and any supporting materials such as “showcards”, as well as statements of informed consent.
- 1.4.3 Clarify payment arrangements for all involved (see Translation: Management and Budgeting).
- 1.4.4 Create a time schedule and identify project phases and milestones for members of the team (see Translation: Management and Budgeting).

1.4.5 Arrange for back-up team members in the event of unavailability or illness.

1.4.6 Decide on the mode and schedule of meetings (face-to-face, web casting, or conference calls) and materials to be used at meetings (e.g., shared templates, software tools, documents deposited in e-room facilities, dictionaries, paper-and-pencil note-taking).

1.4.7 Decide on other communication channels and lines of communication (reporting delays, illness, completion, deadlines).

1.4.8 Decide whether each translator will prepare a full translation (double/parallel translation) or whether the material to be translated will be divided among the translators (split translation).

1.4.9 Decide on deliverables for translation from all study countries (e.g., information on national translation teams, documentation of national versions and translation discussions, etc.).

1.4.10 Translation involves understanding of meaning of the source text and conveying this meaning in the target language with the means of the target language. To this end, identify elements of the source questionnaire that would benefit from the use of translation annotations and explicitly invite countries to point out in advance where they would like annotations. As mentioned in Questionnaire Design, using advance translations or translatability assessment at the beginning of the translation process can effectively minimize later translation problems (See also Appendix D on Annotation. See also Dorer (2011) for carrying out advance translations).

Lessons learned

1.1 Mistaken translation can greatly jeopardize research findings. As reported in the article “World values lost in translation” in the Washington Post (Kurzman, 2014), many translated terms showed different associations from the term used in English. It also shows the changes of translation in later waves of the survey made trend analysis impossible for some countries in the World Value Survey. It thus prevents the analysis on the stability of change in values, which is one of the main goals of the survey.

1.2 It is question development rather than question translation that is the real key to comparative measurement. Questions properly developed for the comparative context give us the chance to measure what we intend to measure and to ask respondents what we intend to ask. At the same time, poorly translated questions (or response categories, instructions, showcards, or explanations) can rob us of that chance – they can mean that respondents are not, in fact, asked what they should be asked. Seen against the costs and effort involved in developing and implementing a comparative study, translation costs are low. On the other hand, the cost of inappropriate versions or mistakes in questionnaire translations can be very high (European Social Survey, 2014).

1.3 In major efforts, the bigger picture must first be considered to confirm which routine or special tasks are vital and which are not. It is easy to focus on procedures which are familiar and thus inadvertently miss other vital elements. For example, if consistency in terminology across versions is not something a project leader has usually considered, procedures to check for this might be overlooked in planning.

1.4 The number of translations required varies among multilingual survey projects. The Afrobarometer Survey, the Asian Barometer Survey, and the ESS Source specify that every language group that is likely to constitute at least 5% of the sample should have a translated questionnaire.

1.5 Planning quality assurance and quality control should go hand-in-hand. When planning the project or procedure, it is also time to plan the quality assurance and quality control steps. For example, in planning the translation of response scales, steps to check that scales are not reversed or a response category omitted can be incorporated into a translation template.

Have two or more translators produce initial, parallel translations.

2. If possible, have each translator produce a full (parallel) translation; if that is not possible, aim to create overlap in the split translation sections each translator produces.

Rationale

Having more than one translator work on the initial translation(s) and be part of the review team encourages more discussion of alternatives in the review procedure. It also helps reduce idiosyncratic preferences or unintended regional preferences. In addition, including the translators who produced the first translations in the review process not only improves the review but may speed it up as well.

Procedural steps

2.1 Determine lines of reporting and document delivery and receipts.

2.1.1 Translation coordinators typically deliver materials to translators. Coordinators should keep records of the delivery of materials and require receipt of delivery. This can be done in formal or less formal ways, as judged suitable for the project complexity and the nature of working relationships.

2.1.2 The project size and complexity and the organizational structure (whether centralized, for example) will determine whether translation coordinators or someone else actually delivers materials and how they are delivered.

2.2 Determine the protocol and format for translators to use for note-taking, asking translation queries and providing comments on source questions, on adaptations needed, and translation decisions. (See Appendix A for documentation templates.)

2.3 Establish deadlines for deliveries, including partial translations (see below), and all materials for the review session.

2.3.1 If working with new translators, consider asking each translator to deliver the first 10% of his/her work by a deadline to the coordinator (senior reviewer or other supervisor) for checking. Reviewing performance quickly enables the supervisor to modify instructions to translators in a timely fashion and enables hiring decisions to be revised if necessary.

2.3.2 Following the established protocol for production procedures and documentation, each translator produces his/her translation and delivers it to the relevant supervisor.

2.4 Where several different translated questionnaires are to be produced by one country, translation begins from the source questionnaire, not from a translated questionnaire (e.g., for a questionnaire with a source language of English and planned translations into both Catalan and Spanish, both the Catalan and Spanish translations should originate from the English version, rather than the Catalan originating from the Spanish translation).

2.5 Any translated components (e.g., instructions, response scales, replicated questions) used in earlier rounds of a survey that are to be repeated in an upcoming round should be clearly marked in what is given to the translators. See also Appendix E (Changes in Existing Translations) regarding material in existing questionnaires. After receiving the translated materials, have the coordinator/senior reviewer prepare for the review session by identifying major issues or discrepancies in advance. Develop procedures for recording and checking consistency across the questionnaire at the finish of each stage of review or adjudication. (See Appendix A for documentation examples.)

Lessons learned

2.1 The more complex the project (e.g., number of translations), the more careful planning, scheduling, and documentation should be (see Translation: Management and Budgeting).

2.2 Since the aim of review is to improve the translation wherever necessary, discussion and evaluation are at the heart of the review process. The senior reviewer or coordinator of the review meetings must, if necessary, help members focus on the goal of improvement. In line with this, people who do not respond well to criticism of their work are not likely to make good team players for a review.

2.3 Review of the first 10% of the initial translation (in case you are working with a new translator) may indicate that a given translator is not suitable for the project because it is unlikely that serious deficiencies in translation quality can be remedied by more training or improved instructions. If this is the case, it is probably better to start over with a new translator. See also Translation: Team for further detail on skill and product assessment.

2.4 The first or initial translation is only the first step in a team approach. Experience shows that many translations proposed in first drafts will be changed during review.

2.5 If translators are new to team translation or the whole team is new, full rather than a split procedure is recommended whenever possible to better foster discussion at the review and avoid fixation on "existing" text rather than "possible" text.

2.6 Not every single word needs to be translated literally as in a word-for-word version. Consider the survey item: "Employees often pretend they are sick in order to stay at home." In this example from ESS Round 4, a country needed to use two words in order to translate "employees" (employees and workers) since a one-word literal translation for "employees" in their language would convey only employees engaged with administrative tasks. The British English word 'employees' covers all those who work for any employer regardless of the type of work they do. Brief documentation may be useful to make it clear to data users and researchers why this addition was needed. This could, for instance, be documented by including a comment in a documentation form; see also examples in Appendix A). However, whenever decisions such as this are made, careful consideration should equally be given to the issue of respondent burden, question length and double-barreled items.

2.7 It is important to inform team members that changes to the initial translations are the rule rather than the exception. The aim of a review is to review AND improve translations. Changes to initial translations should be expected and welcomed.

2.8 Providing templates to facilitate note-taking will encourage team members to do just this. Notes collected in a common template can be displayed more readily for all to see at meetings. The use of a documentation template allows translators to make this documentation while doing the translation (see examples in Appendix A). A few key words suffice; comments do not have to be as fully phrased as in an essay. Review and adjudication can then draw on these comments; review and adjudication become more efficient since reviewers and adjudicators do not have to "reinvent the wheel". It may seem cheaper only to work with one translator and to eschew review sessions, since at face value, only one translator is paid for his or her translation and there are no review teams or team meetings to organize and budget for. In actuality, unless a project takes the considerable risk of just accepting the translation as delivered, one or more people will be engaged in some form of review. When only one translator is involved, there is no opportunity to discuss and develop alternatives. Regional variance, idiosyncratic interpretations, and inevitable translator blind spots are better handled if several translators are involved and an exchange of versions and views is part of the review process. Group discussion (including input from survey fielding people) is likely to highlight such problems. A professional review team may involve more people and costs than an ad hoc informal review but it is a central and deliberate part of quality assurance and monitoring in the team translation procedure. Team-based approaches include the translators in the *review* process. Thus the cost of using two translators to *translate* is offset by their participation in assessment. And since they are familiar with translation problems in the texts, the review is more effective. The team approach is also in line with the so-called 'four eyes principle' requiring that every translation is double-checked by a second equally qualified translator in order to minimize idiosyncrasies in the final translation.

2.9 In addition, even in a team translation procedure, translation costs will make up a very small part of a survey budget and cannot reasonably be looked at as a place to cut costs. Experience gained in organizing translation projects and selecting strong translators and other experts is likely to streamline even these costs (see Translation: Management and Budgeting).The improvements that team translations offer justify the additional translator(s) and experts employed.

2.10 The burden of being the only person with language and translation expertise in a group of multiple other experts can be extreme. If more than one translator is involved in review, their contributions may be more confident and consistent and also be recognized as such.

2.11 When translators simply "hand over" the finished assignment and are excluded from the review discussion, the project loses the chance to have translator input on the review and any discussion of alternatives. This seems an inappropriate place to exclude translator knowledge.

2.12 Relying on one person to provide a questionnaire translation is particularly problematic if the review is also undertaken by individuals rather than a team (these are reasons for working in teams rather than working with individuals).

2.13 Even if only one translator can be hired, one or more persons with strong bilingual skills could be involved in the review process. (The number might be determined by the range of regional varieties of a language requiring consideration for the translation. Bilinguals might not be able to produce a useable translation but could probably provide input at the review after having gone through the translation ahead of the meeting.)

2.14 Translators should ask themselves 'What does this survey item mean in the source questionnaire?' and then put this understanding into words in your own, that is, the target language. They should produce translations that do not reduce or expand the information to the extent that the meaning or the concept of the original source question is no longer kept. It is important that translated items trigger the same stimulus as the source items (this corresponds to the 'Ask-the-Same-Question' approach). However, ensuring a fully equivalent translation may sometimes turn out to be impossible, in particular if two languages do not have terms that match semantically or equivalent concepts at all. In these cases, the best possible approximation should be striven for and the lack of 'full' equivalence clearly noted (European Social Survey, 2014).

2.15 If a country's team comes across interpretation problems that they are unable to solve, they should be encouraged to query the overall coordinator for the project, as the issue may reveal ambiguities that should be clarified for all countries in a multi-country project.

2.16 Translators should be mindful of clarity and fluency. In general, translators should do their best to produce questions that can readily be understood by the respondents and fluently read out by the interviewers, otherwise the measurement quality of the question may be compromised. Writing questions that can be understood by the target population requires not only taking into account usual target language characteristics but also involves taking into account the target group in terms of their age, education, etc. People of various origins should be able to understand the questionnaire in the intended sense without exerting particular effort (European Social Survey, 2014).

2.17 Translators should use words that the average population can understand. Be careful with technical terms. Only use them when you are confident that they can be understood by the average citizen. For example, in one of the ESS translations the ESS item “When you have a health problem, how often do you use herbal remedies?” the technical term “phytotherapie” (“phytotherapy”) was used for “herbal remedies”. This translation was evaluated by an independent assessor as correct but probably not intelligible to most people (European Social, Survey, 2014).

2.18 Translators should try to be as concise and brief as possible in the translation and not put additional burden upon the respondent by making the translation unnecessarily long. Also, if forced by language constraints to spell out things more clearly in the target language than in the source language (e.g. two nouns rather than one noun; a paraphrase rather than an adverb), always keep the respondent burden to the minimum possible (European Social Survey, 2014).

3. If possible, have new teams work with two or more full translations.

Rationale

Having new teams work with two or more full translations is the most thorough way to avoid the disadvantages of a single translation. It also provides a richer input for review sessions than the split translation procedure, reduces the likelihood of unintentional inconsistency, and constantly prompts new teams to consider alternatives to what is on paper.

Procedural steps

3.1 Have several translators make independent full translations of the same questionnaire, following the steps previously described in Guideline 2.

3.2 At the review meeting, have translators and a translation reviewer and anyone else needed at that session go through the entire questionnaire, question by question. In organizing materials for the review, depending on how material is shared for discussion, it may be useful to merge documents and notes in the template (see Appendix A).

Lessons learned

3.1 The translation(s) required will determine whether more than two translators are required. Thus if, for instance, the goal is to produce a questionnaire that is suitable for Spanish-speaking people from many different countries, it is wise to have translators with an understanding of each major regional variety of Spanish required. If, as a result, 4 or 5 translators are involved, full translation can become very costly and splitting the translation material is probably the more viable option.

3.2 Translators usually enjoy not having to carry sole responsibility for a version once they have experienced team work.

4. To save time and funds, have experienced teams produce split translations.

Rationale

Split translations, wherein each translator translates only a part of the total material, can save time, effort, and expense. This is especially true if a questionnaire is long or multiple regional variants of the target language need to be accommodated (Harkness, 2008a; Harkness & Schoua-Glusberg, 1988; Schoua-Glusberg, 1992).

Procedural steps

4.1 Divide the translation among translators in the alternating fashion used to deal cards in many card games.

4.1.1 This ensures that translators get a spread of the topics and possibly different levels of difficulty present in the instrument text.

4.1.2 This is especially useful for the review session—giving each translator material from each section avoids possible translator bias and maximizes translator input evenly across the material. For example, the Survey on Health, Ageing, and Retirement in Europe (SHARE) questionnaire has modules on financial topics, relationships, employment, health, and other topics. By splitting the questionnaire (more or less) page for page, each translator is exposed to trying to translate a variety of topics and better able to contribute directly during review as a result.

4.1.3 Whenever possible, divide the questionnaire up in a way that allows for some overlap in the material each translator receives (see the first two "lessons learned" for this guideline).

4.1.4 Keep an exact record of which translator has received which parts of the source documents.

4.2 Have each translator translate and deliver the parts he/she has been given for the review meeting.

4.3 Use agreed formats or tools for translation delivery for the review session. For example, if a template is agreed upon, then different versions and comments can be entered in the template to make comparison easier during review. (See examples in Appendix A).

4.4 Develop a procedure to check for consistency across various parts of the translation.

4.5 At the review meeting, have translators and the review team go through the entire questionnaire. When organizing materials for the review, depending on how material is shared for discussion, it may be useful to merge documents and notes (see Appendix A).Take steps to ensure that material or terms which recur across the questionnaire are translated consistently. For example, it is conceivable that two translators translate the same expression and come up with suitable but different translations. Source instrument references to a person's (paid) work might be rendered with "employment" by one translator, with "job" by another, and with "profession" by a third.

4.6 Similarly, it is conceivable that two translators translate the same expression and come up with suitable but different translations. Because they are not problematic, they might then not be discussed during review. Consistency checks can ensure that one translator's translation of, say, "What is your occupation?" as something like "What work do you do?" can be harmonized with another translator's rendering as something more like "What job do you have?" (for additional information on consistency, see European Social Survey (2014)).

Lessons learned

4.1 It is often necessary to split the material to address issues of time, budget, or language variety. Even observing the card-dealing division of the material (Harkness, 2008a; Schoua-Glusberg, 1992), there is often no direct overlap in split translations between the material the different translators translate. Translators are thus less familiar with the challenges of the material that they did not translate than the sections they translated. This can reduce the detail of input at the question-by-question review meeting. The senior reviewer must therefore take care to stimulate discussion involving all translators of any section(s) where only one translation version is available.

4.2 Budget and schedules permitting, it is ideal to create some modest overlap in material translated. This allows the review team, including translators, to have an increased sense of whether there are large differences in translating approaches between translators or in their understanding of source text components at the draft production level.

4.3 Giving people time to prepare the materials for the review meeting and making sure that they prepare is important for the meeting's success. Ad hoc suggestions and responses to translations are usually insufficient. Consistency checks can ensure that one translator's translation can be harmonized with another translator's possibly equally good but different rendering.

4.4 In checking for consistency, it is important to remember this procedure must not be only mechanical (for example, using a find function in software). The source text may use one and the same term in different contexts with different meanings, while other language versions may need to choose different terms for different contexts. The opposite may also hold. Automatic harmonization based on "words" is thus not a viable procedure. For example, the English word "government" may need to be translated with different words in another language depending on what is meant. In reverse fashion, English may use different words for different notions which are covered by a single word or phrase in other languages. Examples: English "ready" and "prepared" can in some circumstances be one word in German; "he" and "she" are differentiated in English but not in Turkish or Chinese (see also European Social Survey, (2014)).

4.5 Checks for general tone consistency are also needed: this means that it is important to use the same style in the entire survey instrument, in terms of language register, politeness norms or level of difficulty. There is, for instance, a difference in tone in English between talking about a person's "job" and a person's "profession," or in referring to a young person as a "child" or a "kid."

4.6 Split translations may be helpful in the case of countries with shared languages, where there will be the benefit of input from the other countries' translations. See Translation: Shared Language

Harmonization for further discussion about split translations in countries with shared languages.

5. **Review and refine draft translations in a team meeting.**

Review meetings may be in person, virtual, or a mix of the two. The time involved depends upon the length and complexity of a questionnaire, the familiarity of the group with procedures, and disciplined discussion. The work may call for more than one meeting.

Rationale

The team meeting brings together all those with the necessary expertise to discuss alternatives and collaborate in refining the draft translations—translation reviewers, survey experts, and any others that a specific project requires.

Procedural steps

5.1 Make all the initial translations available to team members in advance of the review meeting(s) to allow preparation.

5.2 Provide clear instructions to members on expected preparation for the meeting and their roles and presence at the meeting.

5.3 Arrange for a format for translations and documentation that allows easy comparison of versions.

5.4 Use the appropriate template to document final decisions and adaptations (see examples in Appendix A). See also Adaptation.

5.5 Appoint a senior reviewer with specified responsibilities.

5.6 Have the senior reviewer specifically prepare to lead the discussion of the initial parallel translations in advance. Prior to the meeting, this reviewer should make notes on points of difficulty across translations or in the source questionnaire and review translators' comments on their translations and the source documents with a view to managing.

5.7 Ask other team members to review all the initial translations and take notes in preparation for the meeting. The time spent on preparation will be of benefit at the meeting.

5.8 Have the senior reviewer lead the discussion.
 5.8.1 The lead person establishes the rules of the review process.
 5.8.2 He/she emphasizes, for example, that most likely the team will change existing translations, and that the common aim is to collaborate towards finding the best solutions.

5.9 Have the senior reviewer appoint two revision meeting note-takers (any careful and clear note-taker with the appropriate language skills, and often the senior reviewer).

5.10 Have the team go through each question, response scale, instruction, and any other components, comparing draft suggestions, and considering other alternatives. Team members aim to identify weaknesses and strengths of proposed translations and any issues that arise such as comparability with the source text, adaptations needed, difficulties in the source text, etc.

5.11 Ensure that changes made in one section are also made, where necessary, in other places. Some part of this may be more easily made after the review meeting on the basis of notes taken.

5.12 Whenever possible, finalize a version for adjudication.
5.12.1 If a version for adjudication cannot be produced, the review meeting documentation should note problems preventing resolution.

5.13 At the end of the translation process (i.e., normally before, and, if needed, after the pretest) copyedit the translation in terms of its own accuracy (consistency, spelling, grammar, etc.).

5.14 Also, copyedit the reviewed version against the source questionnaire, checking for any omissions, incorrect filtering or instructions, reversed order items in a battery or response scale labels, etc.

Lessons learned

5.1 Guidelines are only as good as are their implementation. Quality monitoring plays an essential role. However, evaluation of survey quality begs many issues. Translators asked to assess other translators' work may, for example, be hesitant to criticize or, if not, may apply standards which work in other fields but are not appropriate for survey translation. In the worst instance, they may follow criteria required by people who do not understand survey translation.

5.2 Much remains to be established with regard to survey translation quality. Group dynamics are important. The lead person/senior reviewer leads the discussion. When two suggested versions are equally good, it is helpful to take up one person's suggestion one time and another person's the next time. Given the objectives of the review, however, translation quality obviously takes priority in making decisions about which version to accept.

5.3 Time-keeping is important. The senior reviewer should confirm the duration of the meeting at the start and pace progress throughout. Otherwise much time may be spent on early questions, leaving too little for later parts of the questionnaire.

5.4 It is better to end a meeting when team members are tired and re-convene than to review later parts of the questionnaire with less concentration.

5.5 Practice taking documentation notes on points not yet resolved or on compromised solutions (see Translation: Team).

5.6 Not everyone needs to be present for all of a review meeting. Members should be called upon as needed. Queries for substantive experts, for example, might be collected across the instrument and discussed with the relevant expert(s) in one concentrated sitting.

6. **Complete any necessary harmonization between countries with shared languages before pretesting.**

Rationale

In 3MC surveys, multiple countries or communities may field surveys in the same language. However, the regional standard variety of a language used in one country usually differs to varying degrees in vocabulary and structure from regional standard varieties of the same language used in other countries. As a result, translations produced in different locations may differ considerably. Harmonization should take place before pretesting to avoid unnecessary differences across their questionnaires.

Procedural steps

See Translation: Shared Language Harmonization.

7. **Assess and verify translations, using some combination of procedures discussed in Translation: Assessment, potentially independent of formal pretesting.**

Rationale

Assessment of translation prior to pretesting can identify certain types of errors that are difficult to detect through pretesting alone, and also allow for a more accurate questionnaire for evaluation in the pretest.

Procedural steps

See Translation: Assessment.

8. Have the adjudicator sign-off on the final version for pretesting.

Rationale

Official approval may simply be part of the required procedure, but it also emphasizes the importance of this step and the significance of translation procedures in the project.

Procedural steps

8.1 If the adjudicator has all the skills needed (strong language ability in the source language and target language, knowledge of the study and also survey measurement and design issues), have him or her take part in the review session if this is possible. Even in this case, whenever possible it is advisable to delay official signing-off to another day, thus leaving time for final checking of the decisions taken (Harkness, et al., 2010b).

8.2 If the adjudicator does not have special relevant expertise, have him or her work with consultants to check that all the procedures have been followed, that appropriate people were involved, that documentation was kept, etc., according to procedural requirements. To assess the quality of review outputs, for example, the adjudicator can ask to have a list of all the perceived challenges and request to have concrete examples of these explained.

8.3 If the expertise of the adjudicator lies somewhere between these extremes, consider having him or her review the translation with the senior reviewer on the basis of the review meeting documentation.

8.4 Ensure again that changes made in one section are also made, if necessary, in other places.

Lessons learned

8.1 Emphasizing the value of finding mistakes at any stage in the production is useful. At the same time, a team effort usually shares responsibility. If things are missed, it is best in any instance if no one is made to feel solely responsible.

8.2 If a translation mistake means a question is excluded from analysis in a national study, the costs and consequences are high; in a comparative survey, the costs and consequences are even higher.

Making team members aware of this may help focus attention. For instance, the German mistranslation in a 1985 International Social Survey Programme (ISSP) question regarding participation in demonstrations meant both the German and the Austrian data on this question could not be compared with other countries (Harkness, 2010a). (Austria had used the German translation, complete with the mistranslation).

9. Pretest the version resulting from adjudication.

Rationale

One purpose of pretesting is to test the viability of the translation and to inform its refinement, as necessary, in preparation for final fielding.
All instruments should be pretested before use. The best possible version achievable by the team development process should be targeted before pretesting (see Pretesting).

Procedural steps

See Pretesting.

Lessons learned

9.1 No matter how good the team translation, review, adjudication and any assessment steps are, pretesting is likely to find weaknesses in design and/or translation (Willis et al., 2010).

10. Review, revise, and re-adjudicate the translation on the basis of pretesting results.

Rationale

Pretesting results may show that changes to the translation are needed. Changes can be implemented as described below.

Procedural steps

10.1 Decide on the team required to develop revisions. This will differ depending on the nature and number of problems emerging from the pretest and on whether or not solutions are presented along with the problems.

10.2 If a one- or two-person team is chosen that does not include one of the translators, share any changes (tracked or highlighted) with a translator for final commentary, explaining the purpose of the revision.

10.3 Review the documentation from the pretest, considering comments for each question or element concerned.

10.4 Ensure that changes made in one section are also made, where necessary, in other places.

10.5 Copyedit the version revised after pretesting in terms of its own accuracy (consistency, spelling, grammar, etc.). Target language competence is required for this.

10.6 Copyedit the version revised after pretesting in its final form against the source questionnaire, checking for any omissions, incorrect filtering or instructions, reversed order items or response scale labels, etc. Competence in both target and source language is required for this.

10.7 Check in programmed applications that hidden instructions have also undergone this double copyediting (see Instrument Technical Design).

10.8 Present the copyedited and finalized version for final adjudication. The adjudication procedures for this are as before. Project specifics will determine in part who is involved in the final adjudication.

Lessons learned

10.1 It is extremely easy to overlook mistakes in translations and in copyediting. The review and adjudication steps offer repeated appraisals which help combat this, as do the documentation tools.

10.2 It is often harder to overlook certain kinds of mistakes if one is familiar with the text. It is better if the copyeditors are not the people who produced the texts.

10.3 Although copyediting is a learnable skill, good copyeditors must also have a talent for noticing small details. The senior reviewer should ensure people selected for copyediting work have this ability.

10.4 If the people available to copyedit have helped produce the translations, allow time to elapse between their producing the translation and carrying out copyediting. Even a few days may suffice.

10.5 Problems with incorrect instructions, numbering, filters, and omitted questions are quite common. They are often the result of poor copyediting, cut and paste errors, or inadvertent omissions, rather than "wrong" translation. Thus, for example, reversed presentation of response scale categories is a matter of order rather than a matter of translation. It can be picked up in checking, even if the reversal may have occurred during translation.

10.6 Use a system of checking-off (ticking) material that has itself been tested for efficiency and usability. In iterative procedures such as review and revision, this checking-off of achieved milestones and versions and the assignment of unambiguous names to versions reduces the likelihood of confusing a preliminary review/adjudication with a final one (as an example see the ESS Translation Quality Checklist (European Social Survey, 2014c). Automatic copyediting with Word will not discover typographical errors such as for/fro, form/from, and if/of/off. Manual checking is necessary.

11. Organize survey translation work within a quality assurance and control framework and document the entire process.

Rationale

Defining the procedures used and the protocol followed in terms of how these can enhance the translation refinement process and the ultimate translation product is the most certain way to achieve the translation desired. Full documentation is necessary for internal and external quality assessment. At the same time, strong procedures and protocols do not resolve the question of what benchmarks should be applied for quality survey translation. Harkness (2007) discusses the need for research in this area.

Procedural steps

The steps involved in organizing a team translation are not repeated here. The focus instead is on what can be targeted in terms of translation quality.

11.1 Define survey translation quality in terms of fitness for use:
- 11.1.1 Fitness for use with the target population.
- 11.1.2 Fitness for use in terms of comparability with the source questionnaire.
- 11.1.3 Fitness for use in terms of producing comparable data (avoiding measurement error related to the translation).
- 11.1.4 Fitness in terms of production method and documentation.

11.2 Produce survey translations in a manner that adequately and efficiently documents the translation process and the products for any users of the documentation at any required stage in production (e.g. review, version production control, shared language harmonization, questionnaire design).

Lessons learned

11.1 The effort required to implement a well-structured and well-documented procedure and process will be repaid by the transparency and quality control options it makes possible. Thus even simple Word or Excel templates make it easier to track the development of translations, to check that certain elements have not been missed, and to verify if and how certain problems have been resolved. These might begin with translator notes from the draft productions and evolve into aligned translations in templates for review, later becoming templates for adjudication with translations proposed and comments on these. Dept, Ferrari, & Wäyrynen (2008) provides examples of how Excel templates help guide quality control and assurance steps. An example of such a template used for documenting the whole translation history is the Translation and Verification Follow-up Form (TVFF) used by the ESS since Round 5 (see Appendix A for an example).

11.2 Once procedures become familiar and people gain practice in following protocols, the effort involved to produce documentation is reduced.

12. Translation procedures from the past – no longer recommended.

After in-depth discussion of team translation procedures, other translation procedures often recommended in the past are briefly outlined here. The outlines concentrate on arguments against using such procedures anymore. The chapter briefly outlines other approaches sometimes followed to produce or check survey translations and indicates why these are not recommended here. For discussion see Harkness (2008a), Harkness (2008b), Harkness et al. (2004), and Harkness, et al., (2010b).

12.1 Machine translation: One of the main goals of machine translation is to greatly reduce human involvement in translation production, where word-based matches can be identified, and, it is assumed, cultural and dynamic aspects of meaning are reduced. However, survey questions are a complex text type with multiple functions and components whose complexities cannot be fully recognized by technology (Harkness, 2007; Harkness, 2010a; Harkness & Schoua-Glusberg, 1998; Harkness, et al. (2010b). As a result, any reduction of human involvement in the decision-making process of survey translation through an automatic mechanism is ill-advised (Harkness, et al. (2010b). If a machine translation is used for questionnaire items, then careful review and adjudication of the resultant translation are necessary.

12.2 Do-it-yourself ad hoc translation: It is a mistake to think that because someone can speak and write two languages he or she will also be a good translator for these languages. Translation is a profession with training and qualifications. Translatology (under various names) is a discipline taught at the university level. Students of the translation sciences learn an array of skills and procedures and become versed in translation approaches and theories which they employ in their work. At the same time, as explained in the description of team translation following here, survey translation calls for not only a good understanding of translation but also of the business of survey measurement and how to write good questions. Under normal circumstances, a trained translator should not be expected to have a strong understanding of survey practices and needs, hence the need for a team of people with different skills (Acquadro et al., 1996; Harkness, 2008a; Harkness, 2003; Harkness, 2007; Harkness & Schoua-Glusberg, 1998; Harkness, et al. (2010b).

12.3 Unwritten translation:

12.3.1 Sometimes bilingual interviewers translate for respondents as they conduct the interview acting as interpreters. In other words, there is a written source questionnaire that the interviewers look at but there is never a written translation, only what they produce orally on the spot. This is sometimes called "on sight" translation, "on the fly translation;" or "oral translation."

12.3.2 Another context in which survey translation is oral is when interpreters are used to mediate between an interviewer speaking language A and a respondent speaking language B. The interviewer reads aloud the interview script in language A and the interpreter is expected to translate this into language B for the respondent and, most important, does not change the translation from one interview to the other. The interpreter is also expected to translate everything the respondent says in language B into language A for the interviewer. Research is quite sparse on the process of oral translation in surveys and how this affects interpretation, understanding, and data. Evidence available from recent investigations suggests that these modes of translation must be avoided whenever possible and that extensive training and briefing should take place if they must be used (Harkness, Schoebi, Joye, Mohler, Faass, & Behr, 2008b; Harkness, Villar, Kruse, Branden, Edwards, Steele & Wang, 2009a; Harkness, Villar, Kruse, Steele, Wang, Branden, Edwards, & Wilson, 2009b).

12.4 Translation and back translation: Even today, many projects rely on procedures variously called "back translation" to check that their survey translations are adequate. In its simplest form, this means that the translation which has been produced for a target language population is re-(or back-) translated into the source language. The two source language versions are compared to try to find out if there are problems in the target language text. As argued elsewhere, instead of looking at two source language texts, it is much better in practical and theoretical terms to focus attention on first producing the best possible translation and then directly evaluating the translation produced in the target language, rather than indirectly through a back translation. Comparisons of an original source text and a back-translated source text provide only limited and potentially misleading insight into the quality of the target language text (Harkness, 2003; Harkness & Schoua-Glusberg, 1998; Harkness, et al. (2010b; Harkness, Villar, Kephart, Behr, & Schoua-Glusberg, 2009; Harkness, Villar, Kephart, Schoua-Glusberg, & Behr, 2009).

Appendix A

Translation and Documentation templates

Nowadays, it is recommended to make use of Excel files in order to both carry out and document the whole questionnaire translation processes and histories for each language version.

Below, the Translation and Verification Follow-up Form (TVFF) used by the ESS since Round 5, will be described and discussed as an example of such an excel-based translation and documentation template.

The TVFF as used by the ESS in Round 7 can be downloaded from the ESS translation page:
http://www.europeansocialsurvey.org/methodology/verification.html

Development of the TVFF

The TVFF was created prior to the translation activities of Round 5 of the ESS: in ESS Rounds 1-4, word-based translation templates had been used (see below). In Round 5, ESS used, for the first time, translation verification by the external service provider cApStAn (http://www.capstan.be/) as an additional translation quality assessment step. cApStAn had been using "Verification Follow-up Forms – VFFs" for their verification processes in other projects prior to verifying ESS translations. The TVFF was thus a way to combine the ESS translation templates with the verification templates: this allowed usage of only one template for the whole translation history for each language version. The TVFF is therefore the result of close collaboration between the ESS translation team (http://www.europeansocialsurvey.org/methodology/translation.html) and cApStAn (http://www.capstan.be/ess/).

[For clarification: the abbreviation is sometimes using brackets "(T)VFF" and sometimes not "TVFF": in the case of the ESS, the national teams have the choice to use this excel file for their translations ("T") too – but it is used for verification ("VFF") in all cases; this optional use for translation is mirrored by retaining the T in these guidelines.]

Overview of the TVFF

Figure 1 is an overview of the TVFF used in Round 7 of the ESS: it allows for adequate documentation of the translation process, in addition to documentation of the verification process if used as a method of assessment (see Translation: Assessment). For a more detailed explanation of the ESS Round 7 TVFF, see the ESS Round 7 Verification Instructions 2014). Being an Excel file, the TVFF is highly flexible and can easily be tailored to different projects' needs. Also in the ESS, it is slightly modified from round to round in order to be adapted to new methodological developed in each round.

Figure 1 – Overview of the TVFF

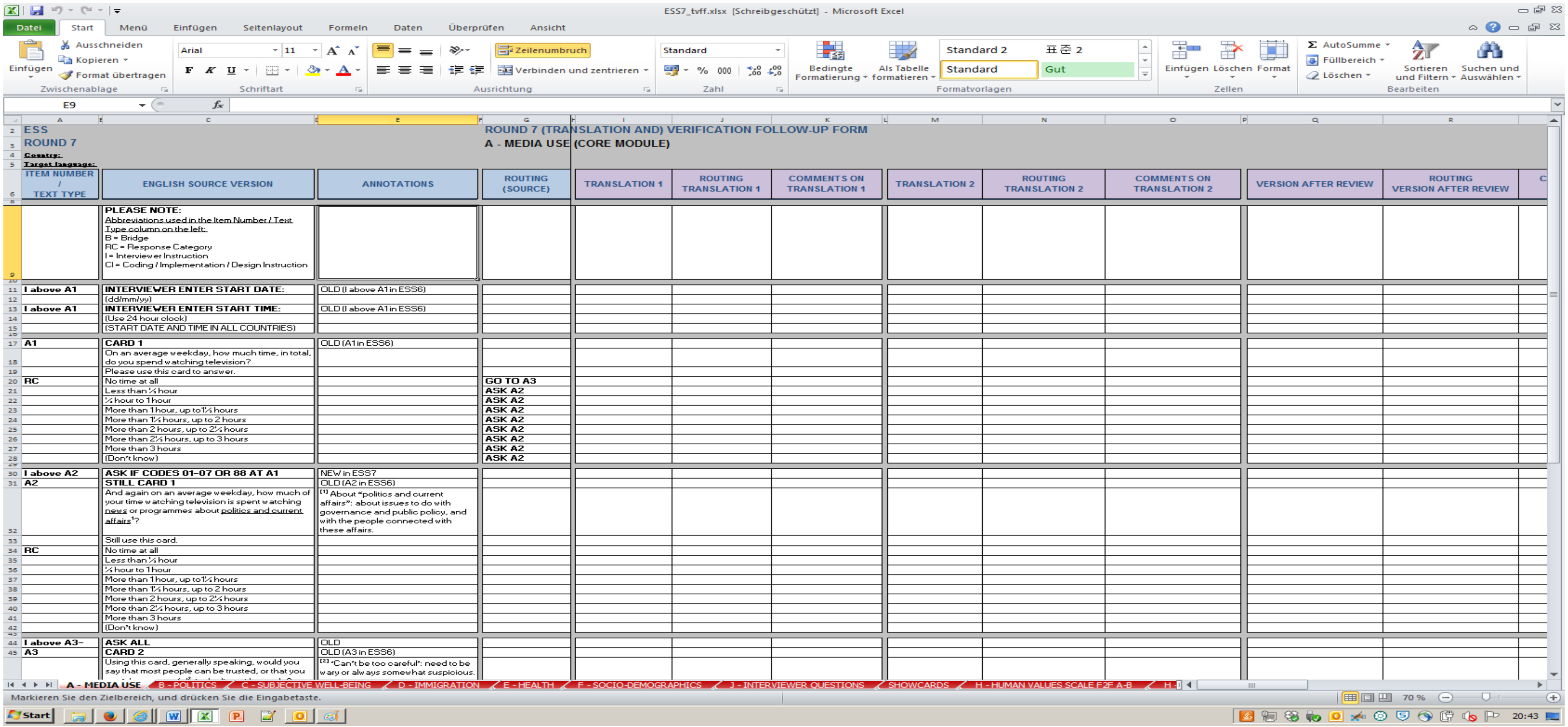


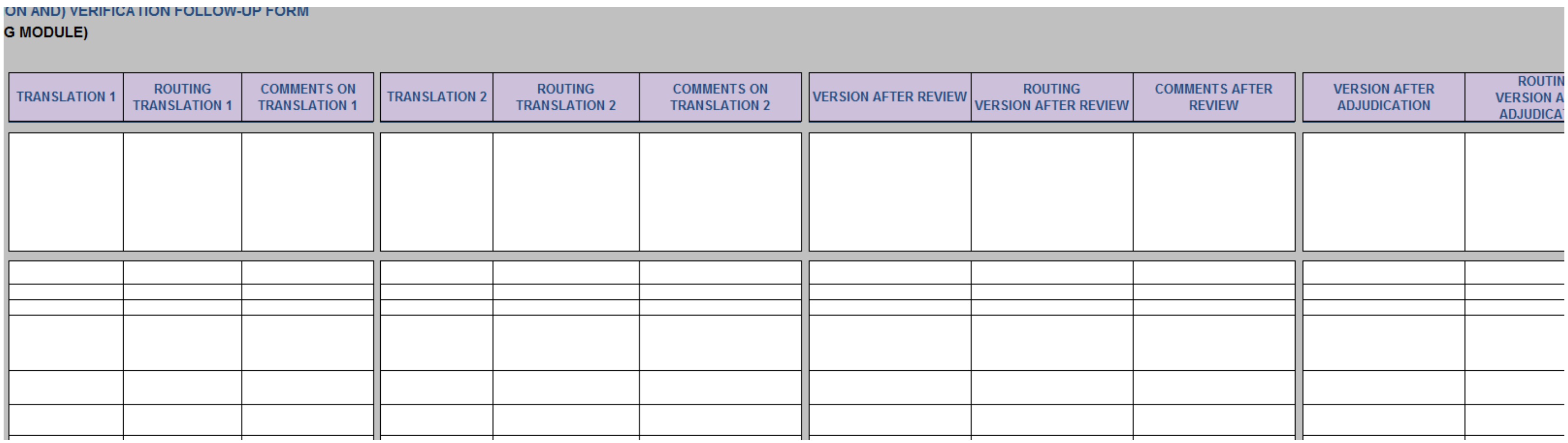
Figure 2 shows the first columns from Figure 1 – with blue headers – in greater detail. The first four columns include (i) the item number; (ii) the English source version; (iii) the ESS annotations, such as the footnotes from the questionnaire, changes between rounds, etc.; and (iv) routing information (e.g., “GO TO”, “ASK item X”, etc.). The columns are populated prior to translation and should not be edited during translation.

Figure 2 – The Source Version Area of the TVFF

ITEM NUMBER/ TEXT TYPE	ENGLISH SOURCE VERSION	ANNOTATIONS	ROUTING (SOURCE)
A2	**STILL CARD 1**	OLD (A2 in ESS5)	
	And again on the average weekday, how much time watching television is spent watching news or programmes about politics and current affairs[1]?	[1] About "politics and current affairs": about issues to do with governance and public policy, and with the people connected with these affairs.	
	Still use this card.		
RC	No time at all		
	Less than 1/2 hour		
	1/2 hour to 1 hour		
	More than 1 hour, up to 1 1/2 hours		
	More than 1 1/2 hours, up to 2 hours		
	More than 2 hours, up to 2 1/2 hours		
	More than 2 1/2 hours, up to 3 hours		
	More than 3 hours		
	(Don't know)		

The columns in Figure 3 are for use by each target country translation team. There are four columns in which to enter the two independent initial translations and the comments relevant to each of these, two columns for the version after the review process and relevant comments after review, and an optional column for issues that arise during discussions between countries that share a language (see also Translation: Shared Language Harmonization). The last two columns are for the version that will be verified in the assessment phase (see Translation: Assessment) and comments arising from the adjudication process.

Figure 3 – Translation Area of the TVFF, reserved for the national teams



ON AND) VERIFICATION FOLLOW-UP FORM
G MODULE)

TRANSLATION 1	ROUTING TRANSLATION 1	COMMENTS ON TRANSLATION 1	TRANSLATION 2	ROUTING TRANSLATION 2	COMMENTS ON TRANSLATION 2	VERSION AFTER REVIEW	ROUTING VERSION AFTER REVIEW	COMMENTS AFTER REVIEW	VERSION AFTER ADJUDICATION	ROUTIN VERSION A ADJUDICA

There may be terms or expressions that are difficult to translate and that have been subject to a lot of discussion during the review and adjudication processes. In such cases it is likely that actors carrying out later translation quality assessments will also stumble over the same issues; it would thus be helpful to document the reasoning behind the choice of word or expression in the comment column.

Figure 4 provides an example of a documented adaptation.

Figure 4 – Example of a documented adaptation

VERSION AFTER REVIEW	COMMENTS AFTER REVIEW	VERSION AFTER ADJUDICATION	COMMENTS AFTER ADJUDICATION
		LISTE 1	
Wie viel Zeit verbringen Sie an einem gewöhnlichen Werktag insgesamt damit fernzusehen?		Wie viel Zeit verbringen Sie an einem gewöhnlichen Werktag insgesamt damit fernzusehen?	
Bitte sagen Sie es mir anhand von Liste 1		Bitte sagen Sie es mir anhand von Liste 1	
Gar keine Zeit		Gar keine Zeit	
Weniger als eine 1/2 Stunde		Weniger als 30 Minuten	We have decided to use '30 minutes' instead of '1/2' hours, because...
30 Minuten bis zu 60 Minuten		30 Minuten bis zu 60 Minuten	See above
Mehr als 1, bis zu 1 1/2 Stunden		Mehr als 1, bis zu 1 1/2 Stunden	
Mehr als 1 1/2, bis zu 2 Stunden		Mehr als 1 1/2, bis zu 2 Stunden	

If verification by an external reviewer is a utilized method of assessment, additional columns can be added to the TVFF (see Translation: Assessment Appendix B).

Translation Templates

Before Excel-based translation templates have started to be in use (see above), translation templates had been Word-based in larger 3MC surveys. However, as Excel templates offer more flexibility, especially in terms of adding additional columns in order to represent the whole translation history within a project, for instance the ESS moved to the Excel-based TVFF in Round 5. Other surveys may still use Word-based templates.

Template 1 is typical of templates used in the ESS in Rounds 1-4 for the initial translations. The source questionnaire was entered in the template in distinct sections. Each translator enters his/her translation in the template and provides commentary. For later stages in the translation process, similar templates retain information from each foregoing stage and add columns for outcomes and comments on the current step (see Template 2).

Template 1: Extract from a translation template from the ESS Round 4 for one initial translation (still called 'draft translation' in earlier ESS rounds)

	Source English Section B	Routing	Draft Translation 1	Comments
B above B1	**Now we want to ask a few questions about politics and government**			
B1	How interested would you say you are in politics – are you...			
I in B1	**READ OUT...**			
RC	very interested,			
	quite interested,			
	hardly interested,			
	or, not at all interested?			
	(Don't know)			
B2	CARD 6			
	How often does politics seem so complicated that you can't really understand what is going on?			
	Please use this card.			
RC	Never			
	Seldom			
	Occasionally			
	Regularly			
	Frequently			
	(Don't know)			

B = Bridge; CI = Coding / Design Instruction; I = Interviewer Instruction; RC = Response Category; RI = Respondent Instruction

Template 2 illustrates possible headings for a template bringing together two initial translations for a review meeting based on Template 1.

Template 2: Headings and columns required for a team review meeting

	Source English Section B	Routing	Draft Translation 1	Comments	Draft Translation 2	Comments	Review version	Comments from review meeting
B above B1	**Now we want to ask a few questions about politics and government**							
B1	How interested would you say you are in politics – are you....							
I in B1	**READ OUT...**							
RC	very interested,							
	quite interested,							
	hardly interested,							
	or, not at all interested?							
	(Don't know)							
B2	**CARD 6**							
	How often does politics seem so complicated that you can't really understand what is going on?							
	Please use this card.							
RC	Never							
	Seldom							
	Occasionally							
	Regularly							
	Frequently							
	(Don't know)							

B = Bridge; CI = Coding / Design Instruction; I = Interviewer Instruction; RC = Response Category; RI = Respondent Instruction

For the "Use of TVFF in Assessment through Verification" see Translation: Assessment Appendix B.

Appendix B

Tasks and qualifications (where specificity is necessary) of personnel in team translation projects

Translators

Tasks

- Prepare individual translations in preparation for the review session.
- Take notes on translation and source texts in preparation for the review session (documentation to inform the review). Specify everything that you think should be discussed or where you think action is needed (such as modifying the source text or providing additional information).
- Participate in review sessions with other members of the review team.
- Consult on any translation revisions at later stages.
- May assess source questionnaires for comparative viability.
- May assess other survey translations.
- May assist in copyediting.

Qualifications

- See Translation: Team

Reviewers

Tasks

- Participate in review sessions at times identified as relevant depending on their role.
- Contribute their individual area of expertise to developing and refining the translated instrument.

Qualifications

- Very good translation skills and language skills in both source and target language.
- Familiarity with questionnaire design principles as well as the study design and topic.
- One reviewing person with linguistic expertise, experience in translating, and survey knowledge is sufficient.
- If one person cannot be found with these skills, two could cover the different aspects.

Adjudicator

Tasks

- Appraise and officially sign off on translations, usually after the review meeting(s), but also after subsequent steps, such as verification, survey quality predictor software (SQP) or the pretesting – depending on the series of steps carried out in each project.
- Appraise the review outputs, if possible in consultation with a senior advisor (the senior reviewer or other consultant) and approve a final version for

pretesting and fielding. If the adjudicator is also the senior reviewer (reviewer-cum-adjudicator), review and adjudication may follow directly upon one another.
- If the senior person on a project who is officially required to sign off on a translation is not appropriate to appraise translation quality and decisions, this nominal adjudicator may delegate adjudication to another senior person better suited for this task. Alternatively, in the same situation, the adjudicator may use consultants for documentation from the review session(s), to work through the translation and document decision points and notes before signing off.

Qualifications
- Proficiency in both target and source languages.
- Familiarity with questionnaire design principles as well as the study design and topic.

Copyeditor(s)

Tasks
- Check for correctness in the target language, including spelling, omissions, wrong formatting, consistency of formulation, and repeated phrases (e.g., "please tick one box"), and for completeness of revision. When multiple versions are in circulation, teams can become unclear, for example, about which version is indeed intended to be the final version. Copyeditors should also check this and report to their adjudicator.
- Check against the source document for such errors, such as inadvertent omissions or additions or question and answer option reversals, mistakes resulting from copy-and-paste activities, misread source questions, and filter numbering correctness.
- Check against the documentation template for any changes that might have been missed.

Co-ordinator

Tasks
- Large translation efforts, centrally organized studies, or efforts conducted within a large organization may have a coordinator to manage the translation effort in an organizational management sense (schedule coordination, personnel identification, budgeting, and so forth). In the case of multinational surveys, this person would typically work at the project and not country-level.
- In other instances, the senior reviewer may organize the translation effort (this would then be at the country-level).

Substantive and other experts

Tasks
- Substantive experts may be needed to provide advice on a variety of matters, such as the suitability of indicators or the formulation of questions with regard to measurement goals.

- Question design experts might be consulted about changes in format necessitated by translation.
- Interviewers might be consulted for fielding matters relevant to translation.
- Visual design experts might, for example, be consulted about cross-cultural aspects of visual presentation.

Programmers

Tasks

- If the questionnaire is computer-assisted, consultation with programmers, or those familiar with programming requirements, is needed to ensure that the translation document or file is marked appropriately. Numerous programming details may need to differ from one language to another to accommodate different language structure requirements (see Questionnaire Design).

Back-up personnel

Tasks

- Projects sometimes run beyond agreed times of availability of personnel. Personnel may also become unavailable for a variety of reasons. It is a good idea to have back-up personnel in place.

External assessors

Tasks

- If some parts of the translation process or translation outputs are to be subjected to external assessment, suitable assessment personnel will be required (see Translation: Assessment).

APPENDIX C

Examples of Common Causes of Mistranslation

This appendix will review the most common causes of mistranslation, with special attention to issues unique in translation of survey instruments, providing examples drawn directly from the ESS Translation Guidelines for Round 7 (European Social Survey, 2014).

Interference: False friends (lexis)

Translators can be misled by so-called 'false friends'. These do, of course, differ from one language to another: simply looking at the surface structure of language, translators may, for instance, decide that "intimate" will be translated as "intim" in German or as "intiem" in Dutch. While this may sometimes be true, in other cases, this may not work as both German and Dutch have a number of additional translations for the English word "intimate", depending on the exact intended meaning. Words that sound similar across languages may (a) cover the same scope of meaning, there may (b) be overlap in meaning or these words may (c) have different meanings. Therefore, translators should be aware that a similar sounding word may not be what is appropriate in a given context (although in some cases it certainly can be appropriate).

Example 1 (item E27, ESS Round2):

How often, if ever, have you…...misused or altered a card or document to pretend you were eligible for something you were not?

In Example 1, some countries produced a similar sounding translation for "card" ("Karte", "caart"). Independent assessors of these translations were unsure about the meaning of "card" in this context in the source text (In this context, 'card' refers, for example, to 'Identity Card'), and were even more uncertain about the translated versions ("Karte", "caart"), which did not make sense in the context.

Interference: Grammar and syntax

Being concerned about a comparable translation, translators may sometimes stick too closely to source text structures, thereby neglecting the usual target language requirements and the usual way of forming sentences in the target language. Look out for fluency and clarity in the target language *while at the same time* taking into account comparability requirements, i.e., faithfulness. A noun is not always rendered by a noun in the target language, a singular noun not always by a singular noun and an adverb not always by an adverb. Syntactical structures may equally change. For example, "information" is a typical English singular word that often gets translated by a plural noun in other languages.

One-to-one equivalencies and their fallacies
Translation documentation from previous rounds has shown that translators occasionally use the words that typically or automatically come to their mind as one-to-one equivalencies. It is an erroneous belief, however, to think that word 'x' in the source language always leads to word 'y' in the target language (Hönig, 1997). "Government" can have different translations, "work" can have different translations, "job" can have different translations and "reasonable" can have different translations, depending in each case on which of the meaning dimensions of the source language – in this example, English language – words get activated in the given context.

Inexperienced translators are especially prone to using one-to-one equivalencies without further questioning the deeper meaning of the source text (Krings, 1986). For this reason, it is of utmost importance to assemble in a team people with excellent translation and language skills.

Careless reading
There have been cases in the past where careless reading has led to mistranslations. Rather than translating "wealthy" one country translated "healthy" and then others copied this through the shared languages consultations. Rather than translating "wanting a job" countries have translated "waiting for a job". Parallel translation and review (and adjudication) are meant in particular to pick up issues such as these. These oversights can easily happen but one can expect that they are spotted in a carefully implemented team approach.

Imbalance between cognitive processes
To put it in psycholinguistic terms, understanding involves *bottom-up* and *top-down processes*. Bottom-up processes take the textual material as a basis, top-down processes activate world knowledge, experiences, etc. (Kussmaul, 1995; Kussmaul, 2007). Make sure that those processes are kept in balance. Too heavy use of top-down processes may lead to translations that divert too much from the actual source text and which, consequently, may compromise data comparability.

<u>Example 2 (ESS 7 core item A3):</u>
"Using this card, generally speaking, would you say that most people can be trusted, or that you can't be too careful in dealing with people?"

is translated as:

"Using this card, generally speaking, would you say that most people can be trusted, or that you can't be mistrustful with strangers?"

Following experience or stereotypes, the translator might have thought of strangers in connection with "can't be too careful" and thus rendered the abstract term “people” by “strangers” (= top-down processes). He or she did not adequately take into account that “people” in the English source is not specified and so covers both people you know well and strangers, so the textual material itself (= bottom-up processes) was probably not adequately taken into account.

Omission of words or phrases
Translations that are fine from the translation point of view may not be strong enough from a measurement perspective. It is crucial for translators to refrain from omitting (or changing) any words or phrases that provide temporal, spatial or any other type of framework within which the respondent is requested to position their answer (e.g. *last week; in general; on average; all things considered; mainly*; *very*, as in *‘very old’* or *‘very weak sense’*; *about* as in *‘about how many’*, etc.). Omitting words or phrases of this kind would mean that the mental calculations from respondents in your country are not comparable to those elsewhere which in turn might compromise data quality.

<u>Example 3 (based on ESS 7 core item A1):</u>
"On an average weekday, how much time, in total, do you spend watching television? "
In Example 3, a translator might be inclined to reduce the numerous adverbial references, assuming that any one of "average" or "in total" could usefully be omitted to make the sentence clearer. However, for example, if “average” was omitted, an important part of measurement would be lost; respondents might think of their most recent experience rather than taking into account their usual TV watching habits. Rightly or wrongly, the question designer presumably felt it important to include each of these phrases to "guide" the respondent in what to consider.

<u>Example 4 (ESS 7 core item F31):</u>
“What does/did the firm/organisation you work/worked for mainly make or do?”

In Example 4, if a translator in one country omitted “mainly”, that would mean that a respondent’s answer in target culture X would not be as focused on the primary tasks or functions of the firm as in countries where this was included. The respondent may say: “Well, there are many things to say. Which one should I list?” Or they might end up mentioning only one of the rarer functions and miss the main ones entirely.

Errors can also occur if translators inadvertently change the form of conjunction. Conjunctions join together elements of thought, such as words, phrases or sentences. It is important that coordinating junctions such as “and” or “or” or “because of” are adequately rendered in the target language.

<u>Example 5 (ESS 7 core item F36):</u>
“Have you <u>ever</u> been unemployed and seeking work for a period of more than three months?”

The conjunction “and” suggests that “seeking work” is to be undertaken *while* being unemployed. Translating the question along the lines of “being unemployed OR seeking work” does not tap the same concept as in the source text. It is crucial to maintain the original idea in translation.

It is also important not to omit interviewer / respondent instructions or any definitions provided to the respondent. For example, an interviewer instruction such as “CODE ALL THAT APPLY” indicates that several answers are possible. Without such an instruction, interviewers in some countries may believe that only one answer is possible and prevent the respondent from volunteering several answers. This would then compromise comparability between countries with different rules being applied. Being unsure of the meaning of certain words or phrases should never result in omitting them, i.e. in not translating them. As things currently stand, there is little basic research into how respondents specifically process questions with such multiple "signposts" .

Sequence
Translators should preserve the order of enumeration elements, listing multiple components in the target item in the same order as in the source item.

<u>Example 6 (ESS 7 core item B12):</u>
“During the last 12 months, have you done any of the following? Have you worked in a political party or action group?”

The translation in Example 6 should thus read “worked in a political party or action group” and not “worked in an action group or political party.” Intentional deviations should be documented.

Pronominal Systems/Frame of Reference
In contrast to English, many languages have complex pronoun systems that indicate number, gender, age, kinship or in-group/out-group relationships, and social status. A system is often abbreviated to a tu/vous distinction in French, distinguishing between “you, familiar” and “you, nonfamiliar.” Language-specific differences apart, adult users of languages with a tu/vous distinction address young children with the familiar tu form, and address all others with the more formal vous form. When one and the same questionnaire is to be used for different age groups, this can become a translation or version administrative logistics problem.

It is essential to consider who a survey question is asking about in instances of otherwise ambiguous pronouns. Is it the respondent themselves, the respondent’s partner, people in general, people like me, etc.? If the reference person differs between the source text and translation, this may lead to artifacts in the data that make comparison impossible.

Example 7 (ESS 7 core item C6):
"How safe do you – or would you – feel walking alone in this area* after dark? Do – or would – you feel…"
*respondent's local area or neighborhood

Example 7 addresses the respondent personally ("you"). The item is thus about the respondent's own feelings and not about others' feelings. If one country translated this item in a very general way, that is, 'How safe is this neighborhood after dark, walking alone?', the data may not be comparable if general vs. individual perceptions differ.

Some languages need to be more explicit than other languages such as English: in many languages 'you' can be translated in three senses: (1) the respondent personally (singular); (2) the respondent and any other people (plural); (3) 'you' in the sense of general statements, without referring to specific individuals.

Connotations
Be aware that words carry connotations, i.e., associations implied by a word in addition to its referential meaning. These connotations may then lead to unintended reactions on the part of the respondents, resulting in biased data. This may apply, for instance, to translations of "race."

Another example comes from the European Value Survey (EVS): the Spanish scores for an item which measured loyalty deviated from the overall pattern of results for Spain. Upon examination it appeared that, unlike in other languages, the Spanish word for loyalty that was used in the translation had the connotation of "sexual faithfulness" (Van de Vijver & Poortinga, 2005). Take care that the translations used do not convey any ambiguous / unintended connotations that would distort the results.

Unintended Ambiguity
Be careful to not introduce *unintended ambiguity* during the translation process. If, for example, the source text asks how often the respondent 'attends sporting events as a spectator' and the translation provides a formulation that can equally well be understood as directly participating in sport activities themselves, then this translation option should be discarded. Clarity on the concept required from the item will be useful in making final decisions. Ambiguity can also result from *syntactical* ambiguity. Syntactical ambiguity can arise when respondents do not know which part of the question goes with which part. These links should always be made explicit to the respondent. For example, should the item "I really dislike answering machines" be understood as "I dislike answering" or as "I dislike the machines"? (Harkness, Pennell, & Schoua-Glusberg, 2004, p. 456).

Gender

Gender is an aspect that differs between many languages and therefore often causes problems in translation, also in questionnaire translation.
Gender issues can have different forms.

a) A language may require masculine and feminine versions of certain adjectives, nouns, etc. (Harkness, 2003; Harkness et al., 2004) where the English language is gender-neutral.

   Example 8 (ESS 7 core item B20):
   "All things considered, how satisfied are you with your life as a whole nowadays?"

   In Example 8, some languages may require both masculine and feminine versions for "satisfied", e.g. in French *satisfait* and *satisfaite*. It would be good to clarify in advance how this gender issue should be dealt with in each country so that translators can accommodate the specified requirements when doing the first draft translations.

b) Gender can also become an issue in other cases, as Example 9 demonstrates:

   Example 9 (item D32, ESS Round 4):
   "Using this card, please tell me whether you think doctors and nurses in [country] give special advantages to certain people or deal with everyone equally?"

   In Example 9, "doctors" covers all doctors regardless of their sex and "nurses" covers all nurses who care for the sick or the infirm, regardless of their sex. In some languages and translations, the masculine form of "doctors" and "nurses" can be used to refer to both men and women because it can be used in a generic way. In other languages, one may need to find paraphrases in order to avoid making this item a gender-specific item: for example, "nursing staff members" could be used as a translation for "nurses." However, care should always be taken to cover the intended meaning as succinctly as possible so that questions do not become too long.

c) Similar issues also need to be taken into account when asking questions about the respondent's partner. For example, in British English, the word 'partner' could refer to a partner of the opposite or the same sex. However, in some languages both feminine and masculine partners may need to be explicitly referred to in order to allow for all possibilities, e.g. in German "Partner oder Partnerin".

Example 10 (SHARE)
Generic English Questionnaire: "Now I would like to ask you about any partners you may have had who you have not lived with. Have you ever been in a long-term relationship that was important to you, where your partner lived at a different address from you for most of the time?"

In the verification step of the translation process, a professional verifier commented: "National version excludes (from the point of view of grammar) possibility of man having a male partner or a woman having a female partner."

Depending on the target language, some countries will need to decide whether to mention both masculine and feminine forms in order to be politically correct or to only use one of these forms. In this regard, the national teams should follow the line that is best accepted in the respective country. However, the aim is not to exclude one of the genders while at the same time avoiding making a question too complicated or too difficult to ask by continuously repeating both genders.

Response Scales
Translation of scales is among the greatest challenges in questionnaire translation, as response scales represent the data that is analyzed (Behr, 2009; Harkness, 2003; Harkness et al., 2004, Harkness, et al. (2010b). Several dimensions of response categories are addressed below.

Intervals:
Make sure that the intervals in the target text are comparable to the source text. If the source has no *overlap* or *gaps*, then the translated question should not have them either.

Example 11 (ESS 7 core item A1):
- No time at all
- Less than ½ hour
- ½ hour to 1 hour
- More than 1 hour, up to1 ½ hours
- More than 1 ½ hours, up to 2 hours
- More than 2 hours, up to 2 ½ hours
- More than 2 ½ hours, up to 3 hours
- More than 3 hours
- (Don't know)

For example, if, in the translation, the third category ("1/2 hour to 1 hour") and the fourth category ("More than 1 hour, up to 1 1/2 hours") both include "1 hour," unambiguous assignment to a response category is not assured any more. If, in the translation, neither of those categories includes “1 hour”, then the respondent would be at a loss as to which category to assign his or her answer of 1 hour.

Labels of categories
Try to produce labels which are as equivalent as possible to the source text and which work at the same time in the target language context.

a) In this case try to mirror the intensity of scale points as expressed in the source language. For example, the translation of “quite interested” (cf. Example 12) should have a lower intensity than that of “very interested”, whilst “hardly interested” should be less in intensity than “quite interested” and so on. Make sure that the qualifiers (very, quite, etc.) chosen for the labels adequately convey the graduation required.

   <u>Example 12 (core item B1 in ESS 7):</u>
   “How interested would you say you are in politics – are you
   very interested,
   quite interested,
   hardly interested,
   or, not at all interested?”

b) In Example 13 below, target country translators should produce labels that convey the intensity of “extremely”. “Extremely” is a *fixed reference point*, i.e., an extreme end point on the scale where nothing can go beyond it. The same extremity should apply to corresponding labels in the translations. A literal translation of “extremely” is not required, but rather the same “extremeness” – this might be represented in target languages also by ‘completely’, ‘fully’, ‘absolutely’, ‘totally’, etc. It is important to take into account that “extremely” should not be translated using a word equivalent to “very” because they do not have the same graduation i.e. ‘very’ has less intensity.

   <u>Example 13 (“extremely” scale):</u>
   Extremely bad
   Extremely good

   We would secondly also expect countries to produce a linguistically *symmetrical scale* in cases where the source language scale is linguistically symmetrical. By a linguistically symmetrical scale we mean: “extremely” on both ends of the scale.
   However, experience – and literature – dictates caution: In some languages there may not be a close equivalent to “extremely” that collocates, that is, typically occurs in conjunction with the corresponding adjective ‘good’, ‘satisfied’, ‘happy’, etc. In addition, while “extremely” works with both positive and negative adjectives in

the English language, in other languages there may not be an adverb available that can work at both ends of a scale. In these cases it may not be possible to employ linguistic symmetry.

However, what should normally be avoided is swapping between bipolar and unipolar scales (e.g. *bad* <-> *good* becomes *not good* <-> *good*). This decision should only be taken as a last resort and must be documented accordingly.

To get a better impression of the linguistic forces at work when translating response scales and to see where research is urgently needed and to support interpretation of results, if needed, countries should document their scale translation *in case of an unavoidable deviation* making use of an English rendering or explanation so that everybody in the project can understand the nature of the deviation.

Experience tells us also that where an English source language can use the same scale unchanged for a number of items (e.g., “extremely bad” – “extremely good”), this may not be the case in other languages; other languages may, for instance, need to adapt the adjective in gender and number to the corresponding noun, or different translations of the words ‘good’ and ‘bad’ may be required in different contexts. Also in this case, countries should document *any deviations* such as additional show cards added for such reasons.

c) Experience has also told us that for some countries the translation of “not at all often” is problematic. Some countries may solve this problem by using an adverb in the form of “never” in a given context. In this case, countries should document any deviation such as this one.

Length of labels
Try to keep the length of labels as equivalent to the source as possible. This means: If the English label only contains individual words / phrases (extremely good, not at all, to some extent, etc.), do not produce entire sentences such as ‘I am *not at all* happy with the government’s work’. Contrary to that, if the English source questionnaire contains entire sentences as response category (e.g., “I plan for my future as much as possible” or “I never plan my future”), the translation should contain entire sentences as well rather than simply saying “as much as possible” or “not at all.”
In case this is, for linguistic reasons, not possible in the target language, documentation is essential.

Question beginnings
This paragraph refers to introductory phrases such as “To what extent”, “How difficult or easy …” or “To what extent do you agree or disagree …”

“To what extent do you agree or disagree …” or “How difficult or easy …” is a deliberate wording technique in order to introduce the range of answer categories. Simply asking “Do you find it difficult or easy to …” or “Do you agree or disagree …” would not match the answer categories if those range from “Very difficult” over “difficult” and “easy” to “very easy”. So please try to match this open phrasing, if possible, in your language. However, in some languages this becomes very long and clumsy and may mean a too high burden for the respondents. In these cases, the reason for deviating from the English structure should be documented and a ‘lighter’ translation used.

In addition, try to the extent possible to mirror the deliberate balancing in your language (“agree”/“disagree”; “difficult”/“easy”). This balancing suggests to the respondent that all answers are equally valid.

If the question begins with an interrogative word (i.e., *what*, *why*, *where*, *which*, *who*, or *how* if English is the source language), try to reflect the meaning in the translation.

<u>Example 14 (item B2 in ESS 4):</u>
“How often does politics seem so complicated that you can’t really understand what is going on? Please use this card.
Never
Seldom
Occasionally
Regularly
Frequently
(Don’t know)”

Regarding example 14: A translation along the lines of “It is sometimes said that politics is so complicated that one doesn’t really understand what is going on” with the response categories translated as “I never have this impression”, “I seldom have this impression”, … would deviate without reason from the formal characteristics (WH-question) of the source text and should not be implemented.

Document any cases where this is, for linguistic reasons, not possible in a particular language.

Omission and addition of answer categories
Do not add or omit answer categories. This also applies to different types of item non-response categories: E.g., when the English source text only uses the “Don’t know” category, do not add “refuse” or “no answer” categories to your questionnaire. In fact, in the past, different approaches from countries on the number of item-non-response categories added have made research into item nonresponse quite difficult. Also please do

not add answer categories. For example, you may feel that adding ‘farmer’ to an occupational answer list is necessary. But if this is only added in one country but not elsewhere this would be problematic.

Consistency between question and response scale
Question and corresponding answer categories should match linguistically.

Example 15 (item B39, ESS Round 4):
“And, using this card, would you say that [country]’s cultural life is generally undermined or enriched by people coming to live here from other countries? Response categories: Cultural life undermined vs. cultural life enriched”

In example 15, the translation chosen for “cultural life is […] undermined or enriched” in the question itself should also be used in the response categories. Be careful not to use different translations for “cultural life”, “enriched” or “undermined” in the question stem and response categories.

Scale layout
Do not change the layout of the scale, e.g. a horizontal scale should never be changed into a vertical scale. Equally, do not reverse the order of the response categories, e.g. “extremely happy” – “extremely unhappy” should not become “extremely unhappy” – “extremely happy”. If form changes like these are made they would always be seen as a deviation.

APPENDIX D

Annotations / Footnotes

Annotations (which are also called 'footnotes') help to clarify the intended meaning of a source text concept, phrase or term and provide information which allows translators, reviewers and adjudicators to focus on what is meant in survey measurement terms in order to do a better job. They are not meant to be translated verbatim or be added as footnotes to the questionnaire in the target language, in contrast to question-by-question objectives (QxQ's). In this appendix, we draw directly from the ESS Round 7 Translation Guidelines, which provide examples of the use of annotations for the translation teams (European Social Survey, 2014).

The example question below and the two corresponding annotations help to explain how annotations are to be used.

Example 1 (item B3, ESS Round 4):

"How difficult or easy do you find it to make up your mind[1] about political issues[2]? Please use this card.

Annotation [1]: Forming an opinion
Annotation [2]: 'Political issues' in this context refer to political debates, policies, controversies etc."

In Example 1, the annotation for "make up your mind" reads "forming an opinion", and the annotation for "political issues" refers to "political debates, policies, controversies, etc."

The first annotation thus explains an English idiom. Countries may end up using a translation that is a literal translation of "forming an opinion", since this is what is common in their language. Saying "Do not translate the footnote!" does refer to not adding a footnote in the translated questionnaires. However, in this case, the explanation given in the footnote ('forming an opinion') may be an appropriate solution for some target languages: Whether the explanations given in a footnote can be translated and directly be used in the translated questionnaire is a case-by-case decision. In the majority of cases, however, direct translation of footnotes cannot be used in the translated questionnaire (such as with annotation 2):

The annotation for "political issues" reads "Political issues in this context refer to political debates, policies, controversies, etc." To the extent possible, countries should not translate "debates, policies or controversies" but rather use these examples in order to find a generic expression covering all these and other examples. Countries have ended up saying things like "political topics", "political issues" or "as regards the field of politics". On annotations, see also Behr and Scholz (2011).

Annotations on source questionnaires are not intended as crutches for translators to explain what English words of phrases mean *in ordinary terms.* Instead, the goal of annotations is to provide information which allows translators, reviewers, and adjudicators to focus on what is meant in survey measurement terms in order to do a better job.

> NB: In no case the survey instruments as used in target languages should contain footnotes, as only the proper question and answer text should be used for the interview. Footnotes are only intended to facilitate the translation process!

Example 2:
"How many people, including children, live in this household?

Considering Example 2, a question commonly used in many surveys, in some cultures 'household' might be automatically associated with 'home' and hence 'family.' If the annotation notes point out that the focus is on a dwelling unit (however variously defined via 'shared cooking pot' or 'shared finances,' etc.), the intended and necessary focus becomes clear to the translator. At the same time, survey questions often use idiomatic expressions. Adding annotations for translators to help clarify the intended sense here is often necessary, and study countries should be explicitly invited to point out in advance where they would like clarification notes in the form of annotations.

APPENDIX E

Changes in Existing Translations

Survey instruments often contain items that were used in previous rounds or in other studies. In this appendix, we draw directly from the ESS Round 7 Translation Guidelines, which provide suggestions to translators on managing translations across multiple waves of a survey (European Social Survey, 2014).

The policy adapted by researchers is generally to maintain continuity, which is essential for measuring differences across countries and/or change within countries and change over time. However, it is also critical that the translations used in each country are equivalent to the source language and indeed measure what is intended by the researcher.

Both the ESS and the International Social Survey Programme (ISSP) apply the following rule: Where translations used in a target country are not equivalent to the source language and indeed don't measure what is intended by the researchers and changes are therefore considered to be absolutely necessary, changes in existing translations should be made. This is mainly the case where (a) clear mistakes have been made in previous survey rounds or waves or, for instance, (b) the language use in a target language has changed in the meantime and a previously used translation would not be used or even correctly understood anymore.

Due to the unknown impact of even minor changes to the questionnaire, it may be unwise to make desirable but inessential changes (even if they are thought to improve equivalence with the source questionnaire) in the middle of the time series. Translators are explicitly advised against amending a translation simply to improve it with small changes or enhance consistency across the questionnaire *posthoc*. Only real mistakes, that is, justified concerns, should be corrected and subsequently documented in a documentation template such as the TVFF in Appendix A. The ESS has started compiling changes applied to existing translations in a specific report in Round 5 (cf. Dorer (2014)); such reports will be made available for all subsequent rounds of the ESS and may be consulted by data users.

Spelling mistakes and typos can be adjusted at any time but should be documented appropriately. Below are recommendations for consideration of changes in existing questionnaire translations.
A country should be able to make a case for any change, noted by the translator, that they want to implement. If possible, countries that wish to change existing translations should provide some evidence of the benefit of this change (in the case of obvious translation mistakes, however, no evidence would be required). This evidence could take the form of a result from a pre-test or some other assessment. The evidence provided will facilitate the decision-making process for the project coordinators on the acceptability of a change. By discussing any

desired changes with the project coordinators, tinkering with the translation can be avoided.

As Weisberg (2005) notes: “Sometimes a researcher realizes that a question has not been worded perfectly but it is still useful, such as when it is decided that it is better to maintain old question wording so that time trends can be analyzed, even if some alternations could lead to a better question” (2005, p. 112).

However, words and their use may change over time. This change may, for instance, be triggered by altered social conditions or the introduction of politically correct language. Example from the German General Social Survey (ALLBUS) – “Gastarbeiter” (Porst/Jers, 2007): A word that in the past was used for “immigrants” can now not be used any more since the immigrant flow has changed in its composition, qualifications and countries of origins; in addition, current language use also plays a role here. Similarly one can observe dramatic language use in naming of ethnic groups in many countries over time (e.g., Smith, 1992). Awareness of needed change over time should lead to regularly reviewing any core translations and adapting them where necessary.

A translation should be changed if a real deviation between the source and the target text can be corrected. This may, for instance, be the case when:

(a) adding an interviewer instruction that was mistakenly left out of the translation previously;
(b) adding a word or phrase that was left out of the translation previously (e.g. source question asked about full-time work but translated version left out reference to ‘full-time’);
(c) deleting a word or phrase that had previously been included in the translated questionnaire but was not present in the source questionnaire (example: adding examples of different sources of income to the household income question when no such examples were in the source questionnaire);
(d) changing a word that is no longer in common usage in a country e.g. because it is no longer politically correct.
(e) changing a word or phrase in the target language so that its translation more closely matches the intended meaning in the source language

In some cases, the decision will *depend on the evidence provided* by a country. An example may be changing a word that is thought to cause serious comprehension problems where countries will need to demonstrate that the wording has caused serious problems.

Changes for the sole purpose of improving the translation in the absence of a mistake, even if it does not change the meaning in the target language, should be considered closely before implementation. It is rather advised against making such changes.

Examples of changes to the target questionnaire that should not take place between rounds include:

(a) making small amendments to tidy up the question wording e.g. using a more parsimonious phrase rather than a lengthy description;
(b) adding more words or phrases in order to match the source questionnaire more precisely;
(c) trying to harmonize response scales across all parts of the core questionnaire e.g. ensure agree / disagree scales are always translated consistently – if the translations had not been erroneous before (here the time series is more important than consistency within the questionnaire);
(d) trying to harmonize translations with other countries sharing the same language – if the translations had not been erroneous before (here the time series is more important than consistency within the shared languages).

Translation: Management and Budgeting

Janet Harkness, Dorothée Behr, Brita Dorer, and An Lui, and Peter Ph. Mohler, 2016

Introduction

The section describes models of budgeting resources as well as budget items that may need to be included for translation (see Tenders, Bids, and Contracts and Study Management for overall survey budgeting and management).

There is no one costing "recipe" for all projects. The organization and scope of the translation project will determine the structure and complexity of the budget planning. For example, in a centrally organized and centrally financed project, management may be asked to specify what funding resources are needed for top-down pre-specified procedures. Alternatively, a project at local level may be asked to organize, conduct, and budget for one or multiple translations. Depending on how various levels of the project are organized, their local level costing may be needed to estimate required costs for just one translation or be used by a central national team organizing and budgeting for a number of translations for within-country fielding. Alternatively, such costs may be needed by an international team coordinating and budgeting for a multi-country project.

In order to be of relevance for projects of various sizes and budgets, these guidelines do not assume sophisticated project management tools for translation development. They do, however, refer to the potential of such and other options (see Translation: Tools). Large-scale projects on very tight timelines are likely to have such tools.

Guidelines

Goal: To ensure that participating research teams follow widely accepted standards for ethical and scientific conduct from the design of the study through implementation and reporting.

1. **Determine the project management form and the required personnel.**

   Rationale

   Project management may vary according to the organization and scope of the translation project. In large translation efforts, centrally organized studies, and in translation projects conducted by a large organization, a coordinator may be appointed to manage the translation effort of all the languages. Additional coordinators may manage individual languages. When translation is organized at the national level and only involves the language(s) of the country, preexisting staff may take on the function of project manager.

Procedural steps

1.1 Identify the management required or specified.

1.2 Identify or appoint one or more project manager(s) as appropriate.
- 1.2.1 If several people are involved in managing the project, ensure, if possible, that one person has ultimate responsibility for signing-off on decisions, meeting deadlines, delivering products, etc.
- 1.2.2 Keep clear records so someone else can take over if this proves necessary.
- 1.2.3 If several people share the work and responsibility, set up clear sharing, delivery, and checking procedures. This reduces the likelihood of omissions and oversights when work is passed back and forth.

1.3 Identify costs for such personnel as well as management components, such as communication, offices, and meetings.

1.4 Determine whether any external verification personnel and/or system will be used, such as described in Translation: Assessment.

1.5 Identify any overhead costs not already covered.

1.6 Explore the potential and limitations of management systems, such as described in Translation: Tools, and determine whether any such system will be used.

1.7 Budget for organizing and undertaking all relevant steps above.

Lessons learned

1.1 The level of detail involved in translation project management can be easily underestimated. Good management tools are important; they need not necessarily be sophisticated technology.

1.2 Large-scale projects will benefit from content management tools, as described in Translation: Tools.

1.3 Large-scale projects will benefit if the development of translations can be integrated into a system also managing the development of any source questionnaire, as described in Translation: Tools.

2. Identify the material for translation and the language(s) required.

Rationale

The nature and the scope of the material determine which translation procedure to adopt, the number and kind of key players involved, and the schedules and budgets required.

Procedural steps

2.1 Identify the material that must be translated. Apart from the questionnaire itself, translations may be needed of interviewer manuals, contact forms, information leaflets, and programming instructions. Some may call for a combination of local adaptation and translation.

2.2 Establish how many languages are involved and identify any special requirements, such as interpreters for unwritten languages and word lists for interviewers working in regional dialects.

2.3 Identify any material already translated which will be considered for re-use; assess the quality of this material and its suitability for re-use in some form.

2.4 Select translation procedures on the basis of the material required and other relevant project considerations (see Translation: Overview and Guideline 3 below).

2.5 Determine whether special tools or software are to be used in the translation development process and whether these involve costs for the project (see Guideline 6 below, as well as Translation: Tools).

2.6 Decide how translation costs are to be calculated (see Appendix A).

2.7 Budget for preparing materials for the translation process and any preparatory steps, such as creating templates or inputting source text to software.

Lessons learned

2.1 Some materials requiring translation can be easily forgotten. For example, if each country programs its own computer application, the programming instructions will require translation. Underestimation results in underbudgeting, not just of costs but of personnel and time.

2.2 Questionnaires often have repetitive elements. If these can be identified ahead of time, consistency can be improved and, often, costs reduced. Payment for handling repetitive elements should also be determined (see Appendix A).

2.3 It is important to identify clearly any sections which are not to be translated for both the budget staff and the translators.

2.4 Shared languages which are to be harmonized will call for different budgeting. Initial / First translations in such instances may be cheaper but additional procedures for harmonization may increase costs again, depending on the procedures followed (see Translation: Shared Language Harmonization).

2.5 Good planning and preparation of material to be translated and good briefing and training are investments which can reduce later costs and improve the quality of the translation. However, such preparation must also be included in the budget.

3. **Identify the translation procedures to be followed and the human resources needed, and budget accordingly.**

Rationale

The translation protocol chosen impacts the number and kind of people involved and time allocations required, as well as management, meeting, and communication costs. Translation procedures may be prescribed or selected according to the nature of the material to be translated. Low priority material might be produced by just one translator.

Procedural steps

3.1 Determine what procedures will be followed for translating the identified materials.

3.2 Determine what people need to be involved. Plan for translation, review, and adjudication, assessment, copyediting, formatting and, if appropriate, the programming of computer applications (see Translation: Overview).

3.3 Identify personnel already available and any that need to be recruited for the translation project.

Lessons learned

3.1 Different procedures may be required by different organizations and project specifications. Large educational testing projects, such as the Trends in International Mathematics and Science Study (TIMSS), typically include a review and revision component undertaken by a commercial company. The World Health Mental Health Survey Initiative required a harmonization meeting for Spanish versions. For some of its instruments, the Gallup Organization hires a commercial company to organize translators and translations, while Gallup personnel closely monitor the output. The Survey on Health, Ageing, and Retirement in Europe requires participating countries to use a common translation tool (Braun & Harkness, 2005). Each of these factors can affect meetings, training, the preparation required, and the degree of external versus internal outlay called for, as well as the number and kind of people involved in activities.

3.2 The intensive, and possibly more costly, procedures chosen for one set of materials may not be needed for all the materials.

4. Determine the scope of selection and briefing meetings.

Rationale

Careful translator team selection and briefing is essential. Meetings for these purposes should be included in the budget (see Translation: Building a Team).

Procedural steps

4.1 Unless you are working within a framework that provides both the materials for selection and briefing and the protocols for these steps, budget for planning and developing these materials and protocols.

4.2 Include outlay for selection and briefing meetings in the budget.

4.3 Include any advertising and networking costs involved in this.

4.4 Decide whether or not in-house training is required.

4.4.1 This will depend upon the study needs and the qualifications of the translators and any other personnel involved.

Lessons learned

4.1 There are few selection and briefing resources publicly available for survey research translation. These can be developed from existing surveys.

4.2 Physical meetings may be costly; training-the-trainer meetings may be of questionable suitability. Webcasting and webinars require advance preparation and time zone scheduling but may be one viable option for a worldwide project.

4.3 Regional meetings (in whatever form) may prove more effective than too-large meetings across a project. In this case, it would be useful if at least one experienced person were able to be involved in all of the regional meetings.

5. Determine the nature and scope of review/adjudication meetings.

Rationale

Review and adjudication discussions are central to the quality of the final translation product and should be included in the budget.

Procedural steps

5.1 Identify the number of meetings required, the form of the meetings, and the people who must be involved.

5.2 Consider any catering, travel, or accommodation costs related to physical meetings and any other costs related to virtual meetings.

5.3 Develop a time schedule and plan for the meetings.

5.4 Determine the time and resources required to plan, conduct, and report on the meetings.

5.5 Reserve funds for planned meetings after the main translation phases (e.g., after pretesting), as well as for unexpected meetings to resolve last-minute problems.

Lessons learned

5.1 If personnel charges different rates at different times, meetings that need to take place during evenings or weekends may be more costly.

5.2 Time-zoning may also need to be considered.

5.3 Working days, public holidays, and "office hours" differ across countries.

5.4 See Translation: Shared Language Harmonization and Translation: Assessment for details on this and an indication of what it could mean for budgeting.

6. Budget for materials that may need to be acquired for the project.

Rationale

Any special resources, such as software, language aids, or digital recorders should be budgeted for.

Procedural steps

6.1 Determine whether or not materials such as the following are needed and already available:
- 6.1.1 Dictionaries.
- 6.1.2 Manuals for translator training and briefing.
- 6.1.3 Software or licenses (translation tools, project management tools, webcasting).
- 6.1.4 Notebooks or computers.
- 6.1.5 Projectors.
- 6.1.6 Digital recorders (audio and/or video recording for documentation and possibly later research purposes).

6.2 If they (or other materials) are not available but will be needed, budget accordingly.

Lessons learned

6.1 It may be difficult for a coordinator to identify or acquire materials with which he or she is not familiar and is uncertain how to locate.

6.2 It is a good idea to check that technical components and equipment are compatible with existing equipment at intended locations before purchase. It is also useful to check that any equipment purchased has a reasonable shelf-life.

Appendix A

Estimating translation costs

It is important to plan in sufficient funding for translation purposes. Translation costs can be estimated in a number of ways. Please note that in general, the cost structure greatly depends, for instance on the language pair and on regional habits. Therefore, we recommend to always check within your local or regional network for usual translators' rates.

Translators may be paid by hour / time spent, by standard page, standard line or by word.
Elements that may add to the final translator's cost are:

- translator's experience and training
- deadline (translators may add a supplement for short-term delivery or weekend work)
- use of technical software / translation tools
- payment decided for any repeated text segments
- need to accommodate regional variants of a language
- difficulty of specific project, for instance due to highly technical terminology used
- additional services required beyond translation
- training and briefing
- other elements, such as the number of questions in an instrument.

Apart from the extent of work to be translated, numerous factors affect what a translation will cost. Table 1 outlines additional factors relevant for estimating costs for survey translations.

Table 1: Factors affecting translation costing

Factor	***Comment***
Availability of translators for the languages involved	It is easier in given locations to find good translators for some language pairs than for others. The more difficult it is in a location to find someone for the language pair, the more expensive the payment expectations may be. The costs for translations for English into Korean or Vietnamese, for example, are likely to vary depending on where translators are sought. Some language pairs may be expensive in almost every location. It could always be difficult to find translators for a translation from Hungarian into Maltese, for example, or certainly more difficult than a translation from English into Spanish. Hungarian and Maltese are spoken by relatively small numbers and the likelihood of finding good translators diminishes accordingly.
Local standards of pay	These can vary greatly around the world. Some organizations aim for the same going rate (however decided) for all locations; the promise of a steady flow of work to translators might help an organization implement this model. Other organizations and projects try to optimize across locations, paying more in one location than in another and adjusting their decided going rate (however determined) on the basis of local rates of payment and local expectations.
A need to accommodate regional variants of a language	If a project needs to capture suitability for multiple regional variants of a language (Spanish, French, or German, for example), this will require more translators or language advisors to be involved than would otherwise be the case. Shared language harmonization meetings and their outputs (see Translation: Shared Language Harmonization) may need such additional translator input, even if not always in person.
Difficulty of text type	Conventionally some text types (specialized fields with special jargon) can command a higher rate of pay than do more everyday or accessible text types. Even if the rate were

	the same, more difficult texts could take longer and increase costs in that way (if paid by hours spent). Benchmarks of difficulty are usually related to specialized vocabulary, complex or difficult content and possible specialized terminologies. In surveys, the quality of source questions, target population needs, cultural distance from that assumed by the source questionnaire, or variation in questionnaire complexity are examples of factors which can add to difficulty. However, in terms of vocabulary and sentence structure, many questionnaires would not be considered to be difficult texts. What makes questionnaires difficult to translate is less the complexity of language used than the measurement goals pursued and the absolute need to consider these, especially with regard to obtaining the highest possible level of comparability in the final 3MC surveys.
Translation mode	Oral forms of translation (on sight oral and interpreted) may command higher rates of pay than do written texts. Here prices will probably in all cases be hour-based. Due to the difficulty to standardize the interviewing process, oral translation is not generally recommended.
Experience of translators and others involved	Experience may impact speed of translation and deliberations, as well as the quality of decisions. This will affect total time needed. On the other hand, more experienced translators would normally calculate a higher hourly rate than novices or inexperienced translators (as the quality of their work is normally priced in their standard rates).
Payment decided for any repeated text segments	If a survey instrument has many repeated sections (e.g., question introductions always framed similarly or identically, frequent repetition of response scales), this should be calculated in to reduce costs. On the other hand, as stated above: repetitions in the source text do not automatically have to be repeated in translations too (see e.g. consistency issues). Therefore, each repeated text bit needs careful consideration; so in the end, only very few repetitions in the source text merit cost reduction – and this needs to be carefully decided at project-level with experienced staff having in-depth linguistic skills in both the source and the target languages.

Time available for the translation	Express delivery or "rush jobs" normally cost more than does work submitted so as to allow the translator to fit it into his/her normal schedule.
Additional services required beyond translation	Translators can serve multiple functions beyond producing translations, either subsequent or parallel to translation. Apart from involvement in a team translation procedure (see Translation: Overview), for instance, proofreading, copyediting, and questionnaire formatting in the translated language are all tasks translators are sometimes asked to undertake. These would add to the payments made to translators, possibly also booked as "translation costs". Also commenting or providing information on cultural issues may involve additional research work by the translator, which may have to be paid for in addition to translation cost.
Time and budget for translation assessment	Assessment of the translation itself and its quality can be implemented in a variety of ways which have associated effects on both scheduling and budgets.
Training and briefing on special features of the translation	Time needed for training and briefing translators will be added to the final costs but improve quality and perhaps speed of the translation process.
Any software expenses	Software or license purchases may also be booked as part of the translation budget. This may either be paid indirectly via the translators or directly to the software providing firms.

Translation: Team

Janet Harkness with (alphabetically) Dorothée Behr, Ipek Bilgen, AnaLucía Córdova Cazar, Brita Dorer, Lei Huang, An Lui, Mathew Stange, Peter Ph. Mohler, and Ana Villar, 2016

Introduction

The following guidelines describe how to find and select suitable people for a team translation effort; they also outline a briefing for members of the team. The strategies used to select translators and others members of the translation team can also be used to train them, as relevant, in the unique aspects of survey translation. The term "source language" used below refers to the language out of which a translation is made. The term "target language" is used to refer to the language into which a translation is made.

Guidelines

Goal: To locate potential candidates for a team translation effort and to select the most suitable from among these; to brief selected translators on general features of relevance for survey translation and on specific features of the study; and to engage and brief relevant other members of the team.

1. **Search for translators in contexts in which they are likely to work, advertise, or acquire their translation training.**

   Rationale
   At the selection stage it is important, whenever possible, to have multiple candidates from whom to choose. A team effort also requires more than one translator. Organizations that employ or train translators and associations with which translators register or advertise are likely places to begin locating translators for the language(s) required.

   Procedural steps

   1.1 Identify likely organizations, associations, and places where translators register and advertise. Local options may vary greatly; search the internet and telephone directories, places of instruction (e.g., translating colleges), newspapers, and trade journals, and contact any local chambers of commerce, publishers, medical institutions, international firms, advertising companies, places of higher education, and your own network channels and institutions, as available, for help in making contact.

   1.2 Compose and write a job description. Post this at any place identified as potentially relevant. Send the description to any contacts made in organizations. If appropriate, include in the advertisement a request for help in locating suitable people.

1.3 Utilize your own organizational and personal networks. Post the advertisement or job description within your own institution, and ask people you know to suggest contacts.

Lessons learned

1.1 In some locations it may be difficult to find trained translators, either in general or for a language you require. In this case, proficient bilinguals may be the only personnel available. Follow through with them as many of the selection and briefing steps as possible.

2. Require candidates to submit application materials prior to the job interview.

Rationale

Information about a candidate's experience and training and examples of previous translation work may help decide whether a candidate merits consideration. If there are numerous applicants, these materials can be the basis for selecting people to interview.

Procedural steps

2.1 Identify the application materials required in the advertisement. If contact is not made through an advertisement, provide candidates with the job description and request delivery of the application materials before arranging an interview.

2.2 Ask applicants to provide the following:
- 2.2.1 An outline of their training and experience in translation for the languages involved (source and target). This should include the kind of translations the applicant has worked on.
- 2.2.2 Examples of any recent work if possible.
- 2.2.3 Recent references relevant to the job application.
- 2.2.4 Details of their computer skills and access to computer technology.
- 2.2.5 Details of their work experience.
- 2.2.6 Details of their education in general.
- 2.2.7 Details of how, when, and where they acquired competence in the source and target languages.
- 2.2.8 Details of whether they have knowledge about surveys and questionnaires in general and experience in questionnaire translation in particular.

Lessons learned

2.1 Application materials only tell part of the story; avoid hiring on the basis of these alone. Translations delivered for inspection are, for example, not produced under team translation conditions, nor can you know precisely who contributed to their production.

2.2 It is important to identify whether candidates are currently working in the source and target languages, or whether their exposure and use of one or the other lies in the past. Translators should ideally be embedded in the target culture and language, as well as fully conversant with the source language and, as relevant, the culture from which it springs. It is also important to ensure that applicants are competent in both speaking and writing the target and source languages.

2.3 Although language competence in the source and target languages does not guarantee that someone can translate, it is a prerequisite. If bilingual individuals without translation training represent the highest level of expertise available in a given context, select from these, using the materials described in Guidelines 4 and 5 below, and train them intensively.

2.4 Avoid engaging someone simply on the basis of recommendations whenever possible. If there are people with whom, for whatever reasons, the project team is expected to work, evaluate these people to ascertain their skills and possible language expertise. In looking for translators, you may also find suitable candidates for back-up personnel.

3. If working with translation agencies, require reference materials and specifications for both the agency and the translators.

Rationale

The professionalism of the agency needs to be verified, as well as the suitability of translators employed for the survey project. Team translation requires the translators to be available for meetings. Make sure that any agency involved understands and accepts the requirements (see Translation: Overview).

Procedural steps

3.1 Ask agencies to provide the following information about themselves:
   3.1.1 A list of clients and contact options.
   3.1.2 A list of projects (the agency experience record).

3.1.3 Experience in translating questionnaires, if available.
3.1.4 References from recent representative clients.
3.1.5 Years of operation.
3.1.6 Information about the business focus and personnel in the agency (for example, whether the owner or manager has a translation background and whether translation is a central part of the agency's activities).
3.1.7 Any agency sub-contracting procedures relevant for your project.
3.1.8 The agency's procedures for hiring and training translators.
- How they find and select translators.
- How they train, if they do so.
- How they monitor translation performance (who monitors, and how).
- How they ensure quality (4-eyes principle, systematic expert or peer reviews?)

3.1.9 How they intend to accommodate the team translation requirements of your project (meetings, repeated access to the same translators, etc.).

3.2 Ask agencies to provide the translator materials outlined in Guideline 2 above in preparation for the selection interview(s).

Lessons learned

3.1 The cost differential between translators working as self-employed professionals and those provided by agencies greatly depends on the individual context. The same holds with regard to quality. In general, agencies pay translators less than what independent translators working full time earn. Competent translators may nonetheless work with agencies. Agencies, for example, can provide a steady flow of work.

3.2 Agencies initially reluctant to cooperate on requirements for team translation may later develop into valuable and reliable partners.

3.3 If working with a translation bureau or agency, it is important to ensure that you have direct contact with the translators, that you can work with the same translators if possible over rounds (if that is what you wish) and that other requirements for your translation effort can be met. Using translation bureaus will in some cases not be a viable option, since, for example, translators may work long distance and will be unable to attend review meetings. They may also not be allowed by their employers to interact directly with you as 'clients' or, indeed, with each other. It is not common for translation bureaus to accommodate the selection procedures outlined below and they may be more expensive than individual translators are. Many fielding agencies may not be able to provide translators to fit the TRAPD model either (European Social Survey, 2014).

4. **Select translators on the basis of submitted materials and their performance in the interview.**

Rationale

The interview is the opportunity to explore and verify information provided in the application and to test performance in tasks needed for a team translation effort.

Procedural steps

4.1 Appoint one or more people with expertise in survey translation and the languages in question to conduct the interview (typically, senior translation reviewers).

4.2 Organize the interview in such a way that candidates actually demonstrate their competence on the spot, including their ability to produce translations, review existing translations, and accept critiquing of their translations, as well as indicate their knowledge of relevant tools, etc. It can be also helpful to let new translators work as interns before hiring them definitely.

4.3 Use the following indicators as the basis of evaluation criteria for selecting any given translator:
- 4.3.1 Current knowledge of and competence in the source and target languages and cultures.
- 4.3.2 Generally, the mother tongue of the translator is the target language.
- 4.3.3 Translation and review performance on test materials.
- 4.3.4 Experience and expertise in translation.
- 4.3.5 Knowledge of translation tools.
- 4.3.6 Team suitability.
- 4.3.7 Computer skills and access to computer technology. This may be a requirement in many projects.
- 4.3.8 Knowledge of and experience with translating surveys.
- 4.3.9 Availability and salary/payment requirements.

Lessons learned

4.1 Extensive translation experience in one very specialized field may be a drawback for working on survey translations. Someone with years of experience in legal translation may be unused to the everyday language and tone often aimed for in survey translation. But the opposite may also be true—a translator successful in a specialized field may be a competent and versatile translator in general and apt to adapt to the 'survey speak' very quickly.

4.2 Experience in producing survey translations should not be taken as proof of suitability, as many survey translations are poor.

4.3 Given the scarcity of training opportunities for survey translation, not many translators will have been trained to translate questionnaires adequately and may not recognize key measurement features. Thus, in many cases, proven translating skills will be more important than survey translation experience. At the interview, assessment should focus on the demonstrated ability to understand the source text and render it fluently in the target language, as well as the ability to identify problems for translation or adaptation and to ask relevant questions. Translators who have had experience in translating questionnaires but were never actually trained how to handle this kind of text may, indeed, prove difficult to (re-)train. Training on survey measurement features can follow, if a candidate is hired.

4.4 It is important to try to assess whether a candidate seems likely to work successfully as a member of a team. Inform translators at the application stage about the way the work will be organized and make the team discussion component clear. It is not uncommon that translators might be a little wary at first about the idea of discussing/critiquing versions. Take the time to explain that teamwork benefits the end product and that people involved in such teams actually enjoy sharing responsibility and can learn from one another (European Social Survey, 2014).

4.5 It is useful to have a number of applicant translators. Even if you feel you have suitable candidates used in past projects, it is suggested that these people be 'put to the test' along with new recruits. In this way, for example, it is easier to decide who might be better suited as reviewer and who as translator or which of two translators is stronger for the task at hand.

4.6 Where several different translated questionnaires are to be produced by one country, for each target language questionnaire, translation begins from the source questionnaire, not from a translated questionnaire (e.g., for a questionnaire with a source language of English and planned translations into both Catalan and Spanish, both the Catalan and Spanish translations should originate from the English version, rather than the Catalan originating from the Spanish translation). Thus, in every case, translators are needed who habitually work from the source language into the target language (this being their 'strongest' language or mother tongue) (European Social Survey, 2014).

4.7 The people most likely to be good questionnaire translators are people who are already good translators and who learn/are trained to become questionnaire translators. The procedures suggested for training include procedures which can be used to assess the suitability of applicants. Training materials can readily be developed from available questionnaire translations; old questionnaires can be used for training and practice (European Social Survey, 2014).

4.8 Applicants can be asked to identify problems in question formulations in the target language, to provide translations, with common pitfalls such as a symmetrical source scale that is difficult to match in the target language, or a skewed or difficult target scale, to comment on translations already available (old questionnaires or questionnaires specifically prepared for this purpose), to correct translations, to compare their versions with other versions, to make questions more suitable for a specified target population, to comment on questions that are culturally inappropriate or end up biased in translation, to explain what questions are actually asking, and so forth (European Social Survey, 2014).

4.9 These tasks will raise some issues that relate to the source language and source text and others that relate more to translation. In this way you should gain a sense of their target language proficiency and their skill in translation. You will also gain some impression of their ability to work with the specific materials as well as their 'ear' for suitable language for different modes and target audiences. By asking them to translate items and then engaging with them in comparison and discussion of their version against one already available, you can gain a general sense of their commenting skills, an indication of how well they can operate impromptu, and a good first impression of how they react when their translations are discussed – as will happen in the review process. Their flexibility in impromptu generation of versions (alongside the quality of these versions) is a good indicator of likely suitability (European Social Survey, 2014).

4.10 Ideally, team members should both show initiative and be able to recognize and follow good suggestions made by others. Good translators, aware of the constraints, tend to recognize good translation solutions when they see them (European Social Survey, 2014).

4.11 Interviewer training will equally require familiarization with the annotated questionnaire and with the documentation required for the translation–review process (European Social Survey, 2014).

4.12 Even once translators have been appointed, decisions sometimes need to be reversed. The first 10 percent of the first assignment should be delivered – and assessed by a project coordinator or the reviewer – for monitoring as soon as it is completed. It is unlikely that serious deficiencies can be remedied by pointing out or discussing problems. If the translation quality is not already reasonable, it is probably better to start again with a new translator. Reviewing output early also allows you to tell translators about aspects you wish them to treat differently (European Social Survey, 2014).

5. Brief translators on general features of surveys relevant for survey translation, as well as on specific features of the given study.

Rationale

Briefing translators helps them to read, understand, and translate questionnaires as instruments of measurement. Translators need to be able to recognize the design features and various components of surveys in order to handle them appropriately. For example, survey questions have special vocabulary and syntactical features that may run counter to normal written language; instruments have sections addressed to different audiences (interviewer, respondent, programmer, etc.); and questions and response scales reflect measurement goals that an untrained reader might not perceive for what they are. Translators also need to understand the function of target and source texts to see the best possible translation options. What they produce as a translation depends not only on their ability and training but on the quality of the material they are asked to translate and on the task specifications they receive.

Procedural steps

5.1 Use specially developed materials or real questionnaires in source and target languages to brief translators on the following:

5.1.1 Different components of a questionnaire.

- Questions, instructions, explanations, response scales, filters, fills, annotations, sections for official use, programmer instructions, formatting conventions, house-style requirements, etc.
- Vocabulary requirements for the target population.
- Level of vocabulary, as well as regional vocabulary considerations (see Translation: Shared Language Harmonization).
- Segments of the text which are for interviewers and which for respondents and indicate the mode intended for

different materials. Countries using computer-assisted applications should explain fills and provide, as appropriate, the hidden CAPI instructions to be translated.

5.1.2 Explain the notion of questionnaire modes and details of the mode for the project at hand (e.g., oral or written presentation, branching presentation of answer options, web-based response features, etc.).

5.1.3 Response scale designs and their purposes.

5.1.4 Adaptation and any feedback procedures to be followed. The most common causes of mistranslations in survey research. (See Translation: Overview, Appendix C on Causes of Mistranslation for a review of mistranslation causes and examples from past survey research.)

5.1.5 Translation documentation and the procedures to be followed.

5.1.6 The notions of response styles and social desirability, as well as any feedback required from translators in these situations.

5.1.7 The purpose and procedures of any pretesting planned.

5.1.8 Any translated components (e.g., instructions, response scales, replicated questions) used in earlier rounds of a survey that are to be repeated in an upcoming round should be clearly marked in what is given to the translators. Giving translators the entire document lets them see the context for which the material to be translated is intended. This is a better idea than deleting bits you do not require to have translated. If appropriate, translators can also harmonize new translations with existing translations, that is, keep new translations consistent with existing translations covering related material (European Social Survey, 2014). (See also Translation: Overview, Appendix A on Changing Material in Existing Questionnaires.)

5.2 In the briefing process, translators can be asked to identify problems in question formulations in English or the target language, to provide translations, for instance with a symmetrical source scale that is difficult to match in the target language, or a skewed or difficult target scale, to comment on translations already available (old, prepared questionnaires), to correct translations, to compare their versions with other versions, to make questions more suitable for a specified target population, to comment on questions that are culturally inappropriate or end up biased in translation, to explain what questions are actually asking, and so forth, in order to improve survey instrument translation capacity (European Social Survey, 2014).

Lessons learned

5.1 Careful briefing is important to guide translators' perception of questionnaires and ensure consideration of both respondent needs and questionnaire designers' needs in translations.

5.2 Without briefing, translators will translate according to the text models and text types with which they are already familiar. They should be reminded that survey instruments are a very specific text type intended for specific target populations. (a) to (c) are examples of issues that may be particularly relevant when briefing your translation teams. For obvious reasons, this will always have to be a case-to-case decision:

(a) For instance, unless they are reminded that an instrument is intended for *oral* presentation, they may produce a translation more suited as a *written* questionnaire. These guidelines do, in general, assume that questionnaires are administered in an oral way, for instance, in face-to-face or telephone interviews. Written administration, such as in self-completion situations or for web-surveys, sometimes requires a different way of writing. But this is more to be understood as a questionnaire design issue for the source instrument than a proper translation problem.

(b) Questionnaire translators should also be informed that questionnaire elements such as visual presentations may be modified between source and target instruments: for instance, local conventions in terms of vertical vs. horizontal or ladder versus triangular presentations of response scales have to be considered and changed, if necessary. Or right-to left vs. left-to-right response scales in Arabic or Hebrew questionnaires (See Adaptation chapter for more information).

(c) Also regional language use, social class, or accents may be an important point to brief your translating teams in. Should the target instrument be drafted for a specific target population in this regard? The translating teams will need some guidance on how to decide in this regard. (See the example of German used in Switzerland or to diglossia in Adaptation.)

5.3 Briefings should include motivating information to encourage translator commitment and care. Survey translation may call on translators to work repeatedly on the same questions; this iterative process may run counter to their expectations. If they are informed about the high-stakes nature of a survey and the survey costs involved should questions go wrong, they understand repetitive aspects of team procedures better.

5.4 If not given job specifications, translators mentally decide their own, since they cannot translate in a vacuum. Task specifications must thus indicate the intended audience, level of literacy and tone of text (e.g., official or more casual tone), the function of the text (e.g., a questionnaire for fielding or notes to describe the contents of a questionnaire), and the degree of freedom permitted in translation. Translators need to be informed of how close or free the translation is required to be (European Social Survey, 2014).

5.5 Translators should be encouraged to produce questions that do not sound like translations and to use vocabulary that can be understood by less well-educated respondents as well as the better educated. Translators must take into account that questions are intended to be offered (said) once and to require only a normal degree of textual processing (European Social Survey, 2014).

5.6 Translators who are informed about the measurement components of questions and are trained to be sensitive to design requirements as well as target audience requirements are in an optimal position to produce good versions. They are also more likely to be able to point out when a requirement cannot be met and to recognize problems (Hulin, 1987; Hambleton, 1993). It is thus strongly recommended that translators are given support materials, example texts, and the information relevant for their part in producing instruments. For example, the format of an annotated questionnaire and the documentation required are likely to be new to many translators and this should be covered in the briefing session (European Social Survey, 2014).

5.7 Monolingual source language dictionaries listing the different meanings of a word may help finding out what words can mean in various contexts. Sometimes one only thinks of the most typical meaning of a given word and then ignores all others, or one even is not aware of the fact that a word can also have different meanings than those that are usually known. Monolingual dictionaries can help in deciding which meaning of a word is activated and, in addition, they may help in finding the appropriate translation by offering paraphrases and near synonyms which could be used as a basis for translation (European Social Survey, 2014).

6. Identify and engage suitable other personnel required for the translation effort: the senior reviewer —who may also coordinate the project—the adjudicator, and substantive experts. (Translation: Overview and Translation: Overview, Appendix B outline the tasks and procedures involved.)

Rationale

Finding good translators is only one requirement to produce suitable target language instruments. The other personnel should be chosen with care so as to bring together the skills and knowledge required for the project, as outlined in Translation: Overview.

Procedural steps

6.1 Identify the procedures to be undertaken and the skills required for this as described in Translation: Overview and Translation: Overview, Appendix B and seek suitable personnel.

6.2 Require these personnel, as appropriate, to demonstrate their abilities for the tasks in which they will be engaged, possibly along the model outlined above for translators.

6.3 Tailor their briefing and training to the contributions they will make. Ensure this includes a general overview of the planned translation project phases, procedures, and responsibilities.

6.4 If there are people with whom, for whatever reasons, the project team is required to work, meet with and evaluate these people to ascertain their skills and possible language expertise.

6.5 Increase the size of the team as necessary to ensure the right mix of skills is available. Not everyone will be required at all times throughout the project (see Translation: Overview).

Lessons learned

6.1 The senior reviewer and the translators are likely to be the people most important for translation quality; it makes sense to select the best people available.

6.2 Training and briefing can greatly improve the performance of individuals and the team. Be sure to factor in adequate time for training and briefing when scheduling the translation process.

7. Use documentation as a deliberate quality assurance and control tool to enhance selection, training, and briefing and to record performance.

Rationale

Selection is partly based on reviewing documentation submitted on team members' performance and experience. It is also partly based on candidates' performance on materials and documents presented at selection and training meetings. Thus selection materials serve multiple functions. First, they allow selection committee members to prepare for the selection process, permit comparisons of candidate experience and performance, and are the basis of benchmarking. Later, selection materials used to test ability and understanding can function as training and briefing documents.

Procedural steps

7.1 Previous guidelines indicated the kinds of material to request of candidates and what to prepare for selection, testing, and briefing.

Lessons learned

7.1 Over time, an array of materials can be assembled. Documents produced for one round of selection and briefing can be used again for other projects. Materials from surveys can be good resources.

7.2 For some translation performance testing or briefing, it may be easier to create examples and tests.

Translation: Scheduling

Janet Harkness, Dorothée Behr, Brita Dorer, and Peter Ph. Mohler, 2016

Introduction

This section discusses scheduling the translation effort. Scheduling in a multinational project very much depends on how the translations figure in the project as a whole. Translations might, for example, be anticipated in features of the questionnaire design (for carrying out 'advance translations', see Dorer (2011)). There may be centrally approved specifications for how they should be conducted and when; and there might be centrally organized quality monitoring procedures. When translations are produced centrally for a multinational, multicultural, or multiregional survey ("3MC" survey) project, it is likely that a document management system is used in the production and scheduling of source questionnaires and translations.

The following guidelines focus on translation efforts managed at the local or national level. This will be the normal procedure for many projects. However, many of the points considered would also need to be addressed in projects using centralized development and management systems. When translation is carried out at the local level, differences and deviances across local schedules will affect timing and milestones for the general project.

No units of time per task are identified here since time required depends upon the length, the repetitive nature, and the difficulty of the material to be translated, as well as on the number and experience of those involved.

Guidelines

1. **If possible, schedule translation after the source questionnaire has been finalized.**

   Rationale

   The exact length, nature, and coverage of the source instrument cannot be known until the instrument is finalized. All of these affect planning, scheduling, and quality procedures. Consistency lists, glossaries, and checklists, for example, are harder to implement without a finalized version of the source instrument. Material still to be determined may affect existing parts of the questionnaire and implementing adjustments later is complex and error-prone. Organizing translation procedures is also more complicated with regard to split options, language harmonization, and iterative review. These challenges are greatly increased if the instrument in question is long and has many submodules.

Procedural steps

1.1 Make the importance of having a finished source version clear to those involved in procedures that impact its completion and aim to schedule accordingly.

1.2 Optimize scheduling of the source questionnaire to accommodate translation as relevant and possible.

1.3 Optimize scheduling of all steps related to translation.

Lessons learned

1.1 Many steps can be completed before translation begins. Provided the nature and scope of the material is clear and the languages required can be specified, translation team members can be selected and briefed and some tools prepared.

1.2 Time constraints may require translation to start with only a pre-finalized source text or with parts of the source text still missing. In such cases, mechanisms should be in place to efficiently and consistently update the source text and to inform all team members of the changes (see Translation: Tools). In this case, a first round of translation can be followed later with a second TRAPD round. This increases costs but can resolve problems arising from working on partially finished instruments.

2. If possible make intensive use of advance translation in multiple languages and thus schedule translation when the source questionnaire, although seen as complete and "finalized," can still be adjusted if problems are encountered.

Rationale

Careful question design and pretesting can help identify problems in the source questionnaire. Nonetheless, some problems often become apparent only when translating into multiple languages. If adjustment can still be made to the source questionnaire and integrated in the translated questionnaires, quality and comparability can be enhanced. This does not contradict the recommendation to use final source questionnaire as the basis for translation – these adjustments should rather be of a minor degree so as to enable they can still be incorporated before fielding the translated questionnaire. For a formalized use of translation in the questionnaire design process, see the advance translation (Dorer, 2011).

Procedural steps

2.1 Schedule sufficient time between finalizing the source questionnaire and fielding in any location to permit feedback on the source questionnaires resulting from translation. Even when the source questionnaire is finalized, there may be corrections required to be made afterwards, and these often arise through the translation activities.

2.2 Optimize scheduling of the source questionnaire.

2.3 Optimize scheduling of all steps related to translation.

2.4 Identify how and to whom feedback (i.e., information about perceived difficulties) is to be conveyed.

2.5 Establish and schedule deadlines for feedback.

2.6 Emphasize that timely feedback is essential.

Lessons learned

2.1 Since problems related to the source instrument may only become apparent when translation begins, researchers sometimes recommend advance translation (Dorer, 2011) before beginning the formal TRAPD team approach to translation.

3. Schedule time to find, select, and brief translation team members, including any external assessment and verification personnel.

Rationale

Source text quality and client specifications impact the potential quality of translations. Apart from these, however, translation quality depends to a large extent on the competence of the staff involved. It is important to allow sufficient time to recruit and select the best possible people for the job.

Procedural steps

3.1 Consult the guidelines in Translation: Building a Team and Translation: Assessment and set the time-frame appropriately.

3.2 Include time for material preparation for these procedures (see Translation: Building a Team).

Lessons learned

3.1 Finding, selecting, and briefing the translation team can be done before the source text is finalized, provided the language(s) and the nature of the instruments to be translated are sufficiently known.

3.2 Engaging people already familiar with translation team procedures may reduce time for some of these steps.

3.3 Contacting translators who worked well on other kinds of survey projects might reduce the time involved in locating potential staff.

3.4 It may be necessary to retrain long-established translators or other team members if the current project has different needs than those of previous projects.

4. Schedule time to prepare the translation documents.

Rationale

Essential preparation steps for the translation effort must be included in scheduling.

Procedural steps

4.1 Prepare translation and documentation tools for translators as soon as the source text is finalized (see Translation: Tools). Easy-to-use translation and documentation tools speed up the translation process and make subsequent steps more efficient.

4.2 Prepare instructions on how to comply with and use the documentation tools.

Lessons learned

4.1 Allow sufficient time if the tools have to be produced manually. If mistakes are made in producing templates to be used in every location, for example, later attempts to correct these across locations may be unsuccessful.

4.2 Some preparatory work can begin before the source material is finished even if its completion has to wait on the source material.

4.3 If tools required for the project are provided by a central coordinating center, the delivery date of these tools often determines when the translation project can start at the national or local level.

4.4 Local teams may wish to begin translation as soon as they have the source instrument. If tools are not available when that happens, they may translate without the tools. Intended quality assurance and control steps related to tools may then not be in place.

5. Schedule time to prepare the translation instructions and assemble reference materials.

Rationale

Clear project instructions and comprehensive reference materials help translation teams to produce translations that meet the needs of the project. Preparation time and delivery dates for these need to be scheduled.

Procedural steps

5.1 Include time to compile documentation for the team on such relevant aspects of the survey as:

5.1.1 The target population (different language variants might be applicable according to the educational background, age, or region of the targeted population): for instance, a questionnaire targeted for teenagers will have to use a different language style and different words than a survey for the elderly; unsuitable language use in this regard may later have an impact on the response behavior. On regional language use, see also the example of Swiss-German.

5.1.2 The mode or modes planned and how these impact the formulation and structure of the instrument.

5.1.3 How to “read” the source materials. For example, how to recognize in the source material the intended recipient for text segments (respondent, interviewer, programmer, etc.) and how to understand specific measurement features (e.g., such multiple specifications as: “Generally speaking, on an average weekday, how many times in total do you usually ”…").

5.1.4 The purpose and character of source materials (e.g., interviewer manual, showcards, computer-assisted applications, explanations).

5.1.5 As applicable, style guides, quality checklists, and glossaries.

5.1.6 As applicable, reference materials, such as parallel texts, previous source text versions, available translations of the same study, and relevant background information on the research goals.

Lessons learned

5.1 If translation team members are poorly informed about the needs of the project, quality suffers and review and adjudication become longer and more burdensome.

5.2 Release all materials at one time rather than sending bits and pieces to the translator teams. This makes it less likely that important materials are overlooked or forgotten.

5.3 If some or all instructions are provided by a central coordinating center, local coordinators only need to write or assemble the materials needed at their level.

6. Schedule time to produce the initial parallel translations.

Rationale

Quality concerns require that a reasonable time frame be determined for the initial parallel translations.

Procedural steps

6.1 Agree on deadlines for delivery with the translators; these include the deadline for quality control (see Translation: Overview, Guideline 2) and the review deadline.

6.2 Instruct translators to report well in advance if a timeframe or deadline cannot be met, so that project management can respond accordingly.

Lessons learned

6.1 The timeframe available for production of the initial parallel translations may be very short. Translators often work on multiple projects simultaneously. The sooner they are informed about the time schedule, the easier it is for them to organize their workloads accordingly.

7. Schedule time to prepare for and hold review meetings.

Rationale

Quality concerns require a reasonable time frame for review.

Procedural steps

7.1 (See Guideline 5 in Translation: Overview).

7.2 Include time to
 7.2.1 Prepare documents for review (e.g., merge documents).
 7.2.2 Send translations to all team members involved in the review.
 7.2.3 Prepare for the review meeting(s).
 7.2.4 Hold the review meeting(s) and refine the translation(s).

Lessons learned

7.1 The earlier team members are informed about the time frame (i.e., the time available between receiving review documents and the review itself), the better they can prepare. This is true even if there is little time between these steps.

7.2 The time needed for the review meeting depends on the length and difficulty of the texts to be discussed, on the experience of teams, and on successful management of time during the review (see Guidelines 5 and 8 in Translation: Overview).

8. Schedule time for copyediting in the target language and checking against the source language. Copyediting takes place several times.

Rationale

Copyediting text produced is an essential step in quality assurance and control.

Procedural steps

8.1 Establish the stages at which copyediting will be undertaken and schedule accordingly.

8.2 See Guidelines 5 and 8 in Translation: Overview.

Lessons learned

8.1 Equipping copyeditors with a list of the most important features to check can streamline the process and reduce time and costs (see Translation: Tools).

8.2 The last rounds of copyediting should particularly focus on anything recently changed (following review or pretesting, for example); any programming specifications; and checking against the source questionnaire or other relevant materials.

9. **Schedule any necessary harmonization between countries with shared languages before any assessments, including pretesting.**

Rationale

In 3MC surveys, multiple countries or communities may field surveys in the same language. However, the regional standard variety of a language used in one country usually differs to varying degrees in vocabulary and structure from regional standard varieties of the same language used in other countries. As a result, translations produced in different locations may differ considerably. Harmonization should take place before pretesting to avoid unnecessary differences across their questionnaires.

Procedural Steps

See Translation: Shared Language Harmonization.

10. **Schedule assessment and verification of translations using some combination of procedures discussed in Translation: Assessment, potentially independent of formal pretesting.**

Rationale

Assessment of translation prior to pretesting can identify certain types of errors that are difficult to detect through pretesting alone, and also allow for a more accurate questionnaire for evaluation in the pretest.

Procedural Steps

See Translation: Assessment.

11. **Include time for adjudication and its documentation.**

Rationale

In the course of developing the translation, multiple versions of the instrument or given questions can be generated. In order to implement quality assurance and control steps, a decision must be made and recorded about which instrument or question version is taken as the final version for a given phase.

Procedural steps

11.1 See Translation: Overview, Guideline 6 on adjudication. Adjudication is recommended at different steps during the TRAPD process: it is likely to be carried out before pretesting and also after discussing pretesting findings (see “Figure 2. European Social Survey Translation Process” in Translation: Overview). Schedule time accordingly.

Lessons learned

11.1 The resolution of some problems from the review may take more time than expected, especially when external informants or the source text designers themselves need to be contacted.

12. Schedule time for pretesting and discussion of pretest findings.

Rationale

Pretesting is an essential component of quality assurance and quality monitoring.

Procedural steps

12.1 Schedule time for producing a version of the instrument and any other relevant materials adequate for pretesting and for the pretesting itself (see Pretesting).

Lessons learned

12.1 When multiple steps are involved in translation development (e.g., multiple languages for one location or multiple varieties of one language calling for shared language harmonization), the timetable for pretesting and revision can become very tight.

13. Schedule time for producing the final translated questionnaire or application.

Rationale

Completion of the translation is not synonymous with completing a questionnaire or application ready for either pretesting or final fielding and time should be scheduled for this. Final checks may again need to be made.

Procedural steps

13.1 This step includes formatting and producing any paper-and-pencil instruments and programming any computer-assisted instruments. If provided with adequate specifications, those with experience in these areas can provide estimates of the time needed.

13.2 Include time for any final testing required.

Lessons learned

13.1 Mistakes can be introduced at this phase too. Incorrect photocopying or scanning of a source questionnaire page used in preparing a translated version can result in a question being inadvertently omitted, for example. Programming errors and oversights at a late stage can also negatively affect quality.

14. Schedule time for consistency checks across documents.

Rationale

If some documents are related to other documents, it may be necessary to check for consistency across them. For example, if show cards repeat questions or response scales from the questionnaire, consistency needs to be checked across these. The same holds for documents such as interviewer manuals.

Procedural steps

14.1 Identify which documents are involved and which sections of these documents need to be checked.

14.2 Schedule time accordingly.

Lessons learned

14.1 It is important to check not only for the presence of various components in the documents which need to be consistent but to check the consistency of order and fashion in which they are presented. The order of response scale response categories could be inadvertently reversed, for example.

15. Schedule time to translate, check, and produce any other materials needed.

Rationale

If other materials are needed, then they will need to be included in the time schedule and budget.

Procedural steps

15.1 Schedule time to:
- 15.1.1 Determine the nature of the other materials and for which stage of the study they are required.
- 15.1.2 Organize and realize their translation.

Lessons learned

15.1 If the other material is not dependent on formulation and content in the questionnaire, translation can be scheduled whenever it is expedient to meet production requirements for this material.

15.2 If the other material repeats or depends on many questionnaire components, it is better to wait until the questionnaire translation is finalized.

15.3 If time constraints dictate simultaneous production of such other materials and the instrument, it is wise to schedule time for later consistency checks.

Translation: Shared Language Harmonization

Janet Harkness, Brita Dorer, and Peter Ph. Mohler, 2016

Introduction

Shared language harmonization is developing a common version (vocabulary or structure) across questionnaires for different regional varieties of a "shared" language. The guidelines in this chapter address the fact that it is important for countries or locations that share a language to take steps to avoid unnecessary differences across their questionnaires (Harkness, 2000/2008; Harkness, 2007; Harkness et al., 2008a).

Why harmonize language?

In multinational, multicultural, or multiregional surveys, which we refer to as “3MC” surveys, multiple countries or communities may field surveys in the same language. Languages such as Russian, French, German, Spanish, and Chinese, for example, are spoken as a first language by populations in a number of countries. However, the regional standard variety of a language used in one country usually differs to varying degrees in vocabulary and structure from regional standard varieties of the same language used in other countries. For example, American English, British English, and Indian English differ systematically in many ways. Often differences relate to vocabulary and pronunciation, but differences in syntax and other grammatical features of the language are also found.

As a result, translations produced in different locations may differ considerably—not only because there is usually more than one way to translate a question (see Translation: Overview) but because of regional differences in language, social reality, and culture. Thus differences in translation may reflect the given regional standard (e.g., Mexican Spanish versus Castilian Spanish), may simply reflect the fact that there is more than one way to say and to translate the same source text, may actually reflect different interpretations of what the source text intends to convey, or may stem from different social and cultural realities.

Which differences are ‘necessary’ – and should thus be kept – or ‘unnecessary’ – and should therefore be avoided – needs to be defined within each study. In general, the following rule of thumb may be useful: any differences due to (a) factual differences (e.g. referring to different political, educational, or social security systems) or (b) different language use (boot/trunk, grill/broil or storm in a teacup/tempest in a teapot) should be kept. However, where representatives of each country sharing one language agree that a common version can be found, this common version should be used: often, this is the case where the different national teams had – in their initial translations – synonyms or expressions that may equally be used in several countries using one language.

A further complicating factor is that the written regional standard variety of a language may differ systematically and markedly from the spoken form of that language the same community uses. Spoken Swiss German, for example, differs notably from region to region. There is no standard written Swiss German. What is normally used in survey instruments is written Standard-German with some vocabulary and grammatical adaptations to get closer to a kind of “least common denominator oral Swiss-German” so that oral adaptations are less complicated for the interviewer. The interviewer has then to convert written Standard-German to oral Swiss-German, and additionally to the regional needs.

When there are shared languages across one or more countries, each country sharing a language with another is asked to produce and discuss its own initial translation (that is, to carry out the TRA steps of the TRAPD model at the national level) and then consult with other countries fielding in this language. Consultation may provide a fresh perspective on questions a given country may have ‘struggled’ with. In addition, it provides the opportunity for country A to benefit from a neater or better translation made by country B but also suitable for country A. Most importantly, unnecessary and potentially harmful differences in wording can be avoided. Comparing versions may sometimes lead both country A and country B to modify what they have and arrive at a better (perhaps) common version (European Social Survey, 2014).

Guidelines

1. **Harmonize the wording of questionnaires in one language whenever possible.**

   Rationale

   All else being equal, it is preferable to keep the wording constant within a language across locations. If no policy of harmonization is followed, unnecessary differences may proliferate. Some of these, such as differences in translating response scales, may negatively affect measurement (Villar, 2009).

   Procedural steps

   1.1 Decide upon the policy and procedures to be adopted on harmonization (obligatory or optional, full or optimized, top-down or bottom-up; see Appendix A), as well as whether a full or split translation procedure will be used in the case of a bottom-up procedure.

   1.2 Decide on the tools to be used; these should include a documentation component.

1.3 Inform all locations sharing a language of the harmonization policy and procedures and related requirements.

1.4 Schedule and organize any translations so that harmonization is possible given the project's overall schedule and constraints. If working from a single translated questionnaire towards localized versions, prepare and distribute the single translation. If such a top-down approach is used, the single translated version should be produced in a team translation approach that includes input for the different regional varieties of the languages that are to be accommodated.

Lessons learned

1.1 The increased effort, time, and outlay to undertake harmonization may be an obstacle to implementing it.

1.2 Without advance planning, the short time often available for translation may make harmonization preparation and meetings to discuss versions difficult and makes pretesting of alternatives unlikely.

1.3 Without clearly defined protocols and some training, the local teams asked to harmonize may have difficulty making informed decisions about harmonization. They may also not properly record their decisions and their motivations.

1.4 When new locations join an ongoing study, new harmonization needs may arise in previously harmonized versions of questions. No research could be identified on whether it is better for the older harmonization decisions to be kept and the new country to deviate or for all to change. There is "received wisdom" about changing as little as possible but this is always over-ruled when change becomes necessary. These will be case-by-case decisions, depending on the study and also on the translation issue at hand.

1.5 Content management system and localization software can aid identification of text requiring harmonization and provide a documentation option for differences retained (see Translation: Tools).

1.5.1 Keeping the words the same across questionnaires in different locations does not automatically mean that perceived meaning and intended measurement are retained across populations. Pragmatic meaning also needs to be considered (see Braun & Harkness, 2005, Harkness, 2003; Harkness, Villar, & Edwards, 2010; Schwarz, Oyserman, & Peytcheva, 2010). At the same time, there is little research available that clarifies how to keep both semantic meaning and pragmatic meaning stable across surveys in different languages. Pragmatic considerations might also stand in conflict with retaining semantic meaning. It remains to be established how "sameness" and comparability are best ascertained at the textual level (see Braun & Harkness, 2005; Harkness, 2003; Harkness, 2010a).

1.5.2 Localized versions based on a single common translation may have fewer differences across versions in a shared language. This does not mean that the instruments are necessarily better than those with more differences. Careful testing should be carried out to make sure that each population does understand the questions as intended (Harkness et al., 2010b; McKay et al., 1996).

1.5.3 In instances where there is a language shared across more than two countries – Russian, for example – the following procedure may be applied: Two of the affected countries in a 3MC project (e.g., Russia and Ukraine) agree on a common, de-centered 'master' translation. A de-centered translation is one that does not use terms that have precise linguistic equivalence, but rather phrases that are more general and do not rely on a specific linguistic context (e.g., rather than using the English-specific phrase "every cloud has a silver lining", using instead "something good comes from any misfortune") (Smith, Bond, & Kagitcibasi, 2006). This master version would then be used by all countries sharing this language as the 2nd initial translation in their 'national' TRAPD process, i.e. it should be used as one of the two translations in the review session. Also with this option, care must be taken to keep up a communication between all countries involved in order to discuss any criticisms or questions arising during the different review meetings and reconciliation efforts. There must be a thorough review meeting when using the de-centered master translation as the second translation in the TRAPD process in every country. Like in all review meetings, the participation of both people with linguistic and/or translation expertise and with survey knowledge is crucial; and it would be useful if a representative from one of the countries producing the master version could participate in the review meetings (European Social Survey, 2014).

1.5.4 A 'lighter' approach along the line of a 'de-centered master translation' is acceptable in case it is not at all possible to create such a 'de-centered master translation' in instances of shared language across more than two countries. The reason for this may be that schedules of the translation processes in the countries sharing one language vary so much that it is not even possible to organize any reconciliation efforts between two countries. In this exceptional case, countries should be allowed to use the final translation from another country using the same 'shared language' as the 2nd initial translation in the national TRAPD process even if this translation has not been agreed upon with a second country. Again, in the Russian language, for example, this would preferably be the final translation from Russia or Ukraine.

- However, some points need to be considered: (a) this option should only be applied in exceptional cases, that is, if the translation schedules are so distant from each other, that no other reconciliation methods detailed above are possible; in any case, reconciliation methods where all participating countries make a more active contribution to the final translation(s) will be more rewarding for all those participating; (b) there must be a thorough review meeting when using the final translation from another country as the second translation in the TRAPD process in every country; if possible, there should be a communication with the country producing this first translation, giving feedback and also asking questions or providing comments in cases of criticism of this translation; like in all review meetings, the participation of both people with linguistic and/or translation expertise and with survey knowledge is crucial; (c) the disadvantage of this option is that the country finalizing their translation first would normally not benefit from the opportunity of discussing their translation with experienced native speakers from other countries (European Social Survey, 2014).

1.5.5 Splitting a questionnaire between translators can save time and effort, particularly if a questionnaire is long (Martinez, Marin, & Schoua-Glusberg, 2006; Schoua-Glusberg, 1992). At least one translator from each country plus a reviewer and adjudicator (or reviewer-cum-adjudicator) is needed. The translation is divided up between translators in the alternating fashion used to deal cards in card games. The questionnaire should not, for example, be divided into a first half and a second half, nor divided by thematic module. By giving each translator material from the same sections, possible translator bias is avoided and translator input is maximized evenly across the material. Each translator translates his/her own questions (European Social Survey, 2014).

1.5.6 Care is needed to ensure that consistency is maintained across the translation, and 'split' questionnaires require particular care. Steps should be taken to ensure that material or terms which re-occur across the questionnaires are translated consistently. At the same time, it is important to remember that although the source language may use one and the same expression in different contexts, target languages may need to choose different terms for different contexts (e.g., the term "government") (European Social Survey, 2014).

2. **Only keep necessary differences.**

Rationale

There are often several ways to formulate a survey question, an explanation, or even instructions. Teams cooperating in a harmonizing effort must try to lay aside personal preferences. Differences that are maintained across questionnaires should be considered genuinely necessary—and, preferably, demonstrated through testing to be so. It is also possible that countries decide they need different versions. However, countries should try and follow the 'as close as possible to each other but as different as necessary' principle. In all cases, the emphasis must be on 'better' versions, not on 'word level sameness' for the sake of 'word level sameness'. In such cases, countries should document changes made as a result of consultation with each other as well as any differences across sharing countries which are necessary to keep in a form such as the TVFF (see Translation: Overview, Appendix A and European Social Survey, 2014).

Procedural steps

If harmonization takes place on the basis of individual translations made by each national or regional group (bottom-up approach):

2.1 Organize templates to enable easy comparison of the initial translations to be reviewed for harmonization.

2.2 Organize the harmonization meeting(s).
- 2.2.1 These can be face-to-face, perhaps piggy-backing on another meeting. However, webcasting, webinars, or "skyping" may be the only affordable modes of meeting.
- 2.2.2 Share versions prior to the meeting and produce a central document aligning them side by side; use a format that also allows each user to see the source and target questions easily (see Appendix B).
- 2.2.3 If possible, appoint someone to identify types of difference (or just differences) ahead of the meeting, both on the basis of any past experience and by checking the translations to be

harmonized. If this person is someone who also attends the meeting, he or she might usefully introduce each question, summarizing points noticed.

2.2.4 Appoint a meeting chair and determine how group/location decisions will be made, ensuring fair representation of each group/location.

2.2.5 At a reconciliation meeting, translators and the translation reviewer(s) go through the questionnaire question by question discussing versions and agreeing on a common version. The adjudicator(s) may attend the review process or already be involved as reviewer(s). Alternatively, the reviewed version moves on to adjudication.

If common wording in the form of a single translated version is the starting point (top-down approach):

2.3 Organize templates to enable easy comparison of the suggested localizations.

2.4 Have each local team propose modifications it considers necessary to the common version.

2.5 Organize the reconciliation meeting(s).

2.5.1 These can be face-to-face if possible, perhaps piggy-backing on another meeting. However, webcasting, webinars, or "skyping" may be the only affordable modes of meeting.

2.5.2 Define the goals of this meeting (e.g., to review suggested changes, to try to find new shared alternatives, to share questions about the single translation).

2.5.3 Share localization suggestions prior to the meeting and produce a central document aligning them side by side; use a format that also allows the users to see the source questions easily.

2.5.4 If possible, appoint someone to identify the types of localization proposed ahead of the meeting, both on the basis of any past experience and by checking the localizations proposed. If this person is someone who also attends the meeting, he or she might usefully introduce each question, summarizing the suggestions made and questions raised.

2.5.5 Appoint a meeting chair and determine how decisions will be made, ensuring a fair representation of each group/location.

2.5.6 At a reconciliation meeting, translators and the translation reviewer(s) go through the questionnaire question by question discussing versions and agreeing on a common version. The adjudicator(s) may attend the review process or already be involved as a reviewer. Alternatively, the reviewed version moves on to adjudication.

Lessons learned

2.1 Personal language perception and usage can be mistaken for generic language usage. It would be mistaken to assume that because one or more speakers make a distinction that these are then distinctions made by all speakers of a given speech community.

2.2 It may not serve the study's purpose to make decisions on the principle of a "majority" vote. The aim is ultimately to allow necessary difference in any given version.

2.3 Harmonization is not limited to the choice of words or phrases; it can include decisions, for example, about how sentences are structured and response scales organized.

2.4 Sometimes harmonization takes the form of adding a term or an example to whatever is common with other shared language versions. Thus if a question about tobacco use does not cover a special form that is only relevant (but important) for one population, mention of it could be added for that population alongside the other forms of tobacco use mentioned in the other versions of the question. This strategy of keeping what is common but adding a local requirement is frequently found in adaptations (see Adaptation).

2.5 If the top-down localization model is used, teams may spend more time discussing the single translation than any of their localizations. This has advantages and disadvantages. One benefit in discussing the available translation is that the group may have new ideas about a possible common version or a common version with occasional "add-ons" as just described. One possible disadvantage is that consideration of the range of localized suggestions is reduced, with each team member ultimately focusing more on resolving what to choose for his or her own version.

3. Schedule harmonization at an appropriate time.

Rationale

Harmonization efforts can result in changes in one or all national questionnaires. The harmonization decisions need to be made when each questionnaire version (or the single translation) is at an advanced stage of development. Although desirable, iterative rounds of pretesting are not likely to be feasible. Thus if a team translation procedure (documented translation review, adjudication, and pretesting) is followed, harmonization should precede pretesting and thus final adjudication (see Appendix A and Translation: Overview, and in particular, Figure 2. European Social Survey Translation Process). Pretesting can be used to check harmonization decisions. It may also indicate that further changes are required in one or more versions.

Procedural steps

3.1 Identify the time at which a well-developed version of each questionnaire to be harmonized will be pretested (or the single common version is well advanced) and arrange for harmonization before that time. In cases where expert assessment, such as verification or survey quality predictor software (SQP), are part of the translation processes, shared language harmonization might intervene at different point in time: before submitting national translations to verification, and after receiving back the expert checks from verification and survey quality predictor software (SQP) coding (see also “Figure 2. European Social Survey Translation Process” in Translation: Overview). Before harmonization occurs, each country should complete the initial translation process as outlined in Translation: Overview and a preliminary review and revision of these translations.

3.2 Countries then exchange translations with the other country or countries sharing a given language; the arrangements between these countries will be decided on by the countries themselves; the procedure chosen and the different steps should be documented accordingly.

3.3 Countries consult together on the final version for each country. They “harmonize” and remove differences across countries as appropriate and comment on any difference retained, and document every decision accordingly.

3.4 Schedule in-person reconciliation meetings whenever possible. Representatives from all countries involved meet in person in order to discuss all newly translated or adapted questions. At least one person per country must participate in this meeting – ideally this would be the person acting as reviewer/adjudicator from each country; it is recommended that at least one translation expert participates in the meeting (e.g. from the host country of the meeting so that there are no further travel expenses). Of course, additional people can participate, such as translators or technical experts. The outcomes of these reconciliation meetings must also be documented (e.g., in the appropriate column in the TVFF called ‘Shared

Languages Discussion') (see Figure 3 in Translation: Overview, Appendix A).

3.5 If in-person meetings are not possible, plan to exchange translated questionnaires via email and/or telephone. In this case, it is important that the countries involved have a thorough discussion on all critical issues or discussions and also document the outcomes of their deliberations. Similarly, discussions can be held in the form of a web- or telephone-based conference, which may require higher financial and organizational efforts.

3.6 Demographic questions which are country-specific or questions that require national consultation processes between the project leaders and the national teams do not need commentary on differences between national versions (e.g. country specific education variables, alcoholic drinks and quantities common in each country).

Lessons learned

3.1 If countries are fielding at different times, a group fielding much later than others may have trouble carrying out (or funding) harmonization preparations in time for groups fielding earlier. The sooner harmonization is organized and scheduled, the greater the chances are of successful schedule coordination between countries or locations.

3.1.1 In practice, recommending harmonization rather than requiring it may not be sufficient to motivate countries or locations to engage in the extra effort. The European Social Survey (ESS) has been aiming for optimized harmonization and recommending it to participating countries. Since harmonization is not a requirement in the ESS, countries are left with considerable freedom as to whether they harmonize or not. Historically, the countries' various time schedules also did not always easily accommodate a harmonization step. Harmonization has further been complicated by countries with shared languages joining the project at different times (Andreenkova, 2008). But in the 7th round of the ESS, there were harmonization steps for almost all shared languages. Even ex-post comparison of other versions of the same language – that is, no proper 'harmonization' effort, but a mere comparison – can be rewarding in terms of enhanced harmonization and similarity of shared language versions.

3.1.2 Without harmonization, the differences that may arise across different regional versions of questionnaires in a shared language can be considerable and may often be unnecessary (Andreenkova, 2008; Harkness, 2000/2008).

3.2 The differences in regional varieties of languages, at least in terms of what needs to be captured in questionnaires, may sometimes also be overestimated.

3.3 While recognizing and emphasizing that same wording does not mean same meaning or comparable measurement, differences across questionnaires may introduce unnecessary and potentially serious measurement error. It is, therefore, important to include harmonization procedures in the study design.

4. Determine and stipulate documentation requirements and tools for the process and outcomes.

Rationale

Those undertaking documentation should have a clear understanding of what is required and should be provided with aids that enable them to maintain documentation without undue burden. Documentation templates play an essential role while deliberating on harmonization as described above. Documentation also provides the evidence examined in quality monitoring and assurance steps, for any coordination of harmonization efforts that may exist in a project and provides secondary analysts and other users of data with information about differences across instruments.

Procedural steps

4.1 Determine documentation needs and create stipulations to be followed by those involved in harmonization in order to achieve these needs.

4.2 Develop templates for the language harmonization process and the harmonization outcomes (see Appendix B).

4.3 Distribute templates and specifications to all those involved well in advance and ensure they are familiar with their purpose and how to use them.

4.4 Provide examples of what is sufficient documentation and what is not.

4.5 Differences should be documented (e.g., in the TVFF). (See Translation: Overview, Appendix A)

Lessons learned

4.1 Good and accessible documentation is essential to shared language harmonization efforts. It enables teams to compare options more easily while making decisions and also to record clearly the decisions taken. Users of data also benefit from documentation on differences across instruments.

5. Undertake shared language harmonization within a quality assurance and control framework as that relates to translation quality.

Rationale

Language harmonization is undertaken to reduce unnecessary variance across versions of a questionnaire in one language that may negatively affect measurement in any of a variety of ways. The purpose of harmonization is, thus, to enhance measurement quality.

Procedural steps

5.1 Plan and undertake harmonization in controlled procedures as described above.

5.2 Plan to follow harmonization with a pretesting phase.

5.3 Develop the relevant materials needed as described above.

5.4 Identify and engage suitable people to be involved in harmonization as described above.

5.5 Brief team members on the materials, purpose and strategies used in harmonization.

5.6 Complete the main harmonization process.

5.7 Pretest and then modify instruments as relevant.

5.8 Share findings in a well-documented and timely fashion with any coordinating center, as relevant.

Lessons learned

5.1 The more rigorous the translation procedures and the various sub-activities such as harmonization and pretesting become, the more important scheduling, budgeting, and briefing are.

5.2 Long-term, the benefits of having and being able to share well-developed, well-translated and tested instruments can be very considerable.

5.3 It may be more effective to require locations to engage in harmonization than to recommend that they do.

Appendix A

Ways to organize and implement language harmonization

There are several ways to organize and implement harmonization between countries sharing one language with regard to whether it is obligatory or not and in terms of how the procedure is organized. These are outlined in Table 1 below.

Table 1: Shared language harmonization options

Term	Explanation	Advantages	Disadvantages
Obligatory shared language harmonization	The project stipulates that shared language harmonization (in whatever form) must be undertaken.	• Participating locations will be more likely to engage in harmonization procedures. • Unnecessary differences have a better chance of being avoided.	• Obligatory participation might be a real burden on some participants or difficult to realize for scheduling reasons.
Optional shared language harmonization	The project recommends shared language harmonization but does not make it an obligatory requirement.	• Recommending rather than requiring shared language harmonization might be a more realistic requirement in some contexts.	• A recommendation may not be enough to ensure countries engage in the additional effort required. • Unnecessary differences across versions and negative effects on measurement may result.
Full shared language harmonization	The project aims to produce a single language version to be used for all the locations using that language.	• The wording of the questions is the same in each location.	• The "same" wording may be systematically understood differently in different locations, not understood in one or more locations, or even not be correct in some locations.
Optimized shared language	The project aims to harmonize as much as possible, but to permit local divergence from the	• As much as possible is kept common but needed differences are permitted.	• Teams may have difficulty distinguishing between their preferences and what are really required differences. This holds for bottom-up and top-down approaches.

harmonization	shared wording as necessary. Harmonization is pursued only to the degree to which it optimizes comparability.		Teams may lack experience in harmonization decision-making. This holds for bottom up and top-down approaches. • Therefore, it is of utmost importance to have (a) native speakers living in the respective countries and experienced in dealing with linguistic issues, and (b) people experienced in shared language harmonization in all teams.
Top-down approach (localization from single version)	A single target language version is first produced (this may also be called 'master version'). This is then adjusted as necessary for the different varieties of the target language. Production of the single version should take into consideration the needs of the different language varieties to be accommodated. The team translation procedures described in Translation: Overview would be useful for this.	• By beginning with a shared common version or 'master' version, locations may end up with more shared common (or more similar) wording than by using a bottom-up approach. • Teams may lack experience in harmonization decision-making, especially if the teams are new; however, in long-standing and long-running projects, the translating teams may be quite experienced in shared language harmonization. This holds for bottom-up and top-down approaches	• The success of the single translation in anticipating and accommodating needs of different locations can determine how much of the translation is left intact. If the single translation meets with opposition from many groups/locations involved with respect to many components, this will greatly complicate the harmonization effort. • The fact that one translation (and only one) is on the table may make it harder to spot where differences are needed. • People might not propose alternatives they would have seen if each location had made an independent translation. • Shared wording might not mean shared understanding or comparable measurement.
Bottom-up approach (shared language harmonization of different	Each location produces an initial translation (ideally the TR, or, if possible, TRA steps from the TRAPD model). A good version produced on the basis of team translation prior to pretesting should suffice (see	• Every location has already worked in-depth on the source questionnaire and considered an optimal version for their location. • The initial translation coming from each location has	• Locations may be unwilling to produce a draft translation that is ultimately changed again. • Locations might over-perceive the need to retain their versions. • Teams may have difficulty distinguishing between their preferences and what are

versions)	Translation: Overview). These translations form the basis of the harmonization review.	already been worked upon by a team (typically the T-R-A steps have been carried out at national level before going into the harmonization step). • The harmonization review has all the alternatives at its disposal to decide commonalities, possibly find new solutions in the shared language and determine and document needed differences.	really required differences. This holds for bottom-up and top-down approaches. • Depending on the project and the team composition, teams may lack experience in harmonization decision-making. This holds for bottom-up and top-down approaches.

Appendix B

Documentation templates for shared language harmonization steps

Clear instructions and documentation templates help researchers conduct and document shared language harmonization products. Below are a few examples of templates used in recent cross-national surveys in connection with shared language harmonization.

The WHO Mental Health Survey Initiative aimed for an optimized and maximally harmonized questionnaire. The output of harmonization procedures for Spanish in Latin America and Spain is presented in Table 1 below.

Table 1: Examples of harmonization carried out in Spanish-speaking countries in MHSI.

A	B	C	D
English Term Término en inglés	**Terms proposed for Spanish Términos propuestos en español**	**Terms actually chosen Términos seleccionados**	**Terms used in individual locations when harmonization not possible Términos alternativos según país****
Free base, (cocaine-based drug)	Free base		Basuco(1, 3, 8), pasta base (6)
Herbalists	Herbolarios, Naturistas		Naturistas (1,2), homeópatas (1,2), herbolarios (1), herbalistas (2) yerberos/ yerbateros (3, 8)
Hot flashes	Sofocos		Sofocos(1), sofocones (2), bochornos (5,6), calores (8)
Ulcer in your stomach or intestine	Úlcera estomacal o intestinal	Úlcera de estómago o intestinal	

Unhappy	Desdichado(a) Desgraciado(a)	Infeliz o desgraciado(a)	
Upset	Molesto	Alterado	
Using a 0 — 10 scale	Utilizando una escala de 0 a 10	En una escala de 0 — 10	
Usual, usually	Habitual, Habitualmente		Habitual/habitualmente (1), usual/usualmente (2)
Normally	Normalmente	Generalmente	
Was it before you were a teenager?	¿Fue antes de la adolescencia?	¿Fue antes de los trece años?	
What is the day of the week?	¿A qué día de la semana estamos?	¿En qué día de la semana estamos?	
What is the longest period of days, weeks, months, or years you were...?	¿Cuánto duró el periodo más largo de días, semanas...?	¿Cuántos días, semanas, meses o años duró el periodo más largo durante el que...?	
What number describes...?	¿Qué cifra describe...?	¿Qué número describe mejor...?	
What season of the year is it?	¿En qué estación...?		¿En qué estación (1), época (3,8), del año estamos?

Note: The numbers in Column D indicate the countries using the term, as follows: (1) Spain, (2) Latin America, (3) Colombia, (4) Puerto Rico, (5) Mexico, (6) Chile, (7) Argentina, (8) Panama. Table 1 is adapted from Harkness et al., 2008b.

The translation team of the ESS investigated differences across shared language versions in the survey using templates similar to Template 1 below. This template brings together German translations made for different countries and comments on any documentation made in various countries on differences. It was not intended for public use. The people using it understood German and therefore did not explain everything noted to each other. A document intended for public use would need to be more explicit, but this document was satisfactory for the purpose of translation harmonization into German within the context of this project.

Template 1: German translations across participating countries

Code	**Source**	**German Austria**	**German Germany**	**German Lux**	**German Switzerland**	**Comment**
Q. A1	On an average weekday, how much time, in total, do you spend watching television?	Wie viel Zeit verbringen Sie an einem normalen Wochentag insgesamt mit Fernsehen?	Identical to Lux. Wie viel Zeit verbringen Sie an einem gewöhnlichen Werktag insgesamt damit, fernzusehen?	Identical to Germany	Karte 1. Wie viel Zeit verbringen Sie an einem gewöhnlichen Werktag insgesamt mit Fernsehen?	weekday versus work day: not mentioned in notes Watching TV explicit in D/L (verb formulation) nominalized in A and CH; not commented on
I	Please use this card to answer.	Bitte verwenden Sie diese Karte zur Beantwortung.	Bitte sagen Sie es mir anhand von Liste 1.		Bitte verwenden Sie für Ihre Antwort Karte 1.	House styles not commented on
RC	No time at all	See GER/Lux gar keine Zeit	See Austria/Lux Gar keine Zeit		Überhaupt keine Zeit	no comments on differences between CH and the others
	Less than 1/2 hour	See CH weniger als 1/2 Stunde	Weniger als eine 1/2 Stunde		See Austria Weniger als 1/2 Stunde	Differences not commented upon
	1/2 hour to 1 hour	mehr als 1/2 Stunde, bis zu 1 Stunde	1/2 bis zu 1 Stunde		1/2 Stunde, bis zu 1 Stunde	"More than 1/2 an hour up to 1 hour "versus "1/2 to 1 hour" or "1/2 an hour to 1 hour" CH comma possibly disruptive for reading.

Note: The header "Code" in the first column on the left refers to the abbreviations in that column; QA1 = the question code, I = Instructions, RC = response categories.

Also excel templates used for documenting questionnaire translation processes – such as the Translation and Verification Follow-up Form TVFF used in the ESS (see Translation: Overview, Appendix A) can be used to compare translations into one shared language. The columns showing the translations from the different countries can easily be copied next to each other.

Figure 1: TVFF – section on shared language harmonization (example ESS Round 6 German: Germany-Switzerland)

VERSION AFTER REVIEW Swissgerman	COMMENTS AFTER REVIEW Swissgerman	VERSION AFTER REVIEW Germany	COMMENTS AFTER REVIEW Germany	SHARED LANGUAGES - DISCUSSION	V A FO
Ich finde, es gibt vieles, was ich gut kann.	das Wort "Dinge" vermeiden wegen CH-D	Ich habe das Gefühl, in vielen Dingen gut zu sein.		CH unkomplizierter	

Translation: Assessment

Janet Harkness, Brita Dorer, and Peter Ph. Mohler, 2016

Introduction

This chapter on translation assessment will consider different forms of qualitative and quantitative assessment related to translation and present the current state of research and relevant literature as available. It is useful to distinguish between procedures that assess the quality of translations as translations and those that assess how translated questions perform on questionnaire instruments. Survey instrument assessment must address both translation and performance quality (Harkness, Pennell, & Schoua-Glusberg, 2004).

Evaluations of the translations focus on issues such as whether the substantive content of a source question is captured in the translation, where there are changes in pragmatic meaning (what respondents perceive as the meaning), and whether technical aspects are translated and presented appropriately (e.g., linguistic and survey appropriateness of response scales). Approaches combining translation, review, and adjudication, as part of the TRAPD model of translation, are seen to be the most useful ways to evaluate and improve translation quality and implicitly underscore the relationship between design and translation.

Assessments of performance can focus on how well translated questions work for the target population, how they perform in comparison to the source questionnaire, and on how data collected with a translated instrument compares with data collected with the source questionnaire. In the first case, assessment may indicate whether the level of diction is appropriate for the sample population, in the second, whether design issues favor one population over another, and in the third, whether response patterns for what is nominally “the same question” differ (or do not differ) in unexpected ways across instruments and populations.

Translation quality and performance quality are obviously linked, but good translation does not suffice to ensure that questions will function as desired in performance. Thus, well-translated questions may work better for an educated population than for less well-educated population of the same linguistic group, either because the vocabulary is too difficult for the less-well educated or because the questions are less salient or meaningful for this group. Problems of question design, such as asking questions not salient to the target population, should be addressed at the questionnaire design level; they are difficult to resolve in terms of translation. As testing literature points out, question formats also affect responses if the chosen format is culturally biased and more readily processed by respondents in one culture than in another (Geisinger, 1994; Solano-Flores & Nelson-Barber, 2001; Tanzer, 2005).

Assessment and evaluation of translation and performance quality assume that criteria of evaluation are available with which to assess the quality of given translation products and benchmarks and that standards exist against which translation products can be "measured". In the survey research field there is only limited consensus on what these criteria and benchmarks might be and what translations that meet these criteria might then look like.

However, items are measurement instruments in comparative survey research. From this follows that in the end the measurement properties of items must be comparable within well-defined limits in comparative research across countries, cultures or regions. There are a number of statistical methods available that allow the researcher to test for statistical comparability (aka equivalence) ranging from Cronbach's Alpha to Structural Equation Models (See Statistical Analysis) (Braun & Johnson, 2010, van de Vijver & Leung, 1997). Within the Total Survey Error framework other quality issues must also be dealt with (see below).

The guidelines below include several different qualitative and quantitative approaches for translation assessment, identifying criteria of obvious relevance for survey translations and specifying which may or may not be of relevance in a given context. It is unlikely that any one project would employ all the techniques discussed; it is most appropriate for the topic and target population to guide researchers in choosing the most efficient methods of assessment.

Guidelines

Goal: To assess whether the translation of the survey instrument in the target language accurately reflects all aspects of the source language instrument. The material will be divided into subsections as follows:

Assessment and survey translation quality

Assessment and evaluation assume that criteria of evaluation are available with which to assess the quality of given translation products and benchmarks and that standards exist against which translation products can be "measured". In the survey research field there is only limited consensus on what these criteria and benchmarks might be and what translations that meet these criteria might then look like.

This section will deal with these issues. It will identify criteria of obvious relevance for survey translations and will identify others which may or may not be of relevance in a given context.

1. **Assessment as part of team translation.**

Rationale

Qualitative assessment of initial translations as they are being developed is an integral and essential component of team translation procedures (see Translation: Overview).

Procedural steps

(See Translation: Overview.)

Lessons learned

1.1 The TRAPD model is one effective method of detecting translation errors. See Willis et al. (2010) for a discussion of the kinds of mistakes discovered at different stages of translation review in projects based on the TRAPD model.

2. **Translation assessment using external translation assessors and verification procedures in a quality control framework paradigm.**

Rationale

Various models use external reviewers and external verification procedures in survey translation efforts. Some projects currently rely on external review teams to provide most of their assessment; others combine internal assessment procedures with outside quality monitoring.

The word “verification” in this context refers to a combination of checking the linguistic correctness of the target version and checking the “equivalence” of that target version against the source version. And, “equivalence” refers to linguistic equivalence, including equivalence in quality and quantity of information contained in a stimulus or test item, as well as equivalence in register or legibility for a given target audience (Dept, Ferrari, & Wäyrynen, 2010). See Johnson (1998) for more information.

The role of verifiers is to: (a) ensure linguistic correctness and cross-country equivalence of the different language versions of the source instrument; (b) check compliance with the translation annotations provided in the source questionnaire; (c) achieve the best possible balance between faithfulness and fluency; and (d) document all changes for all collaborating countries and any overall project or study coordinators. Verifiers should ideally have prior experience in verifying (or producing) questionnaire translations for other cross-cultural social surveys.

Procedural steps

2.1 An external translation verification firm (e.g., cApStAn) uses a monitoring tool - such as the Translation and Verification Follow-up Form (TVFF) used in the European Social Survey (ESS) - to assess translation and adaptation decisions and to ensure appropriate documentation (see Appendix A; see also Translation: Overview, Appendix A for a discussion of the TVFF independent of its utility in assessment).

2.2 The verifier uses the TVFF (or a similar tool) to label each “intervention” (i.e., recommendation for change or other notation) as necessary for each survey item in question.
   2.2.1 Examples of intervention categories are “minor linguistic defect”, “inconsistency in translation of repeated term”, “untranslated text”, “added information”, “annotation not reflected”, etc. See Appendix B for complete list of intervention categories used in verification of translations of Round 6 of the ESS. See also complete ESS Round 7 Translation Guidelines (European Social Survey, 2014)

2.3 The verifiers may prioritize their interventions using the TVFF (or a similar tool):
   2.3.1 Interventions are categorized as “key” (an intervention that could potentially have an impact on how the questionnaire item works) or “minor” (a less serious intervention that could improve the translation).
   2.3.2 This categorization can help translation adjudicators and other team members to identify which errors are more/less serious.

2.4 Or the verifiers may be asked to require follow-up on all interventions by the national teams, as is the case in ESS Round 7. The idea behind this decision is that no intervention should stay without follow-up by the national teams, otherwise it may be that important corrections are not made if the national teams don’t feel the necessity (European Social Survey, 2014).

2.5 The TVFF (or other documentation form used) is returned to the national team. Each notation by the verifier should be reviewed and any comments/changes/rejections of suggested changes should be marked accordingly. It may be advisable to require the national teams to get back to the verifiers in order to either confirm acceptance of the verification intervention or, in case these interventions are not incorporated, to justify this decision.

Lessons learned

2.1 The purpose of documenting adaptations and other issues in the TVFF is not only to record such issues but also to provide the external verifier with all the relevant background information s/he will need for the verification assignment, to avoid unnecessary comments and changes, and to be as time-efficient as possible.

2.2 The requirement that national teams provide feedback on whether they incorporate verification interventions or not [in the TVFF] provides better control of how verifiers' suggestions are implemented. In addition, the different loops between the verifiers, national teams and translation experts within the survey may trigger interesting discussions about translation and verification issues.

2.3 Recent use of the verification system by cApStAn in ESS translation assessments has found that verification:
   2.3.1 Enhances understanding of translation issues for:
   - The ESS translation team for languages they do not understand;
   - National teams when choosing a translation by encouraging reflection on choices made;
   - Source question designers, enabling them to have a better understanding of different country contexts and issues in translation.

   2.3.2 Enhances equivalence with source questionnaire and across all language versions, especially for problematic items.
   2.3.3 Gives the ESS translation team a better idea of translation quality/efforts/problems in participating countries.
   2.3.4 Prevents obvious mistakes, which otherwise would lead to non-equivalence between countries, from being fielded.

2.4 Systematic external verification streamlines overall translation quality

3. Translation assessment using Survey Quality Predictor Software (SQP) coding

Rationale

SQP can be used to prevent deviations between the source questionnaire and the translated versions by checking the formal characteristics of the items. SQP coding is meant to improve translations by making target country collaborators more aware of the choices that are made in creating a translation, and the impact these choices have on comparability and reliability of the question. The ESS has been using SQP Coding as an additional step of translation assessment since Round 5 (European Social Survey, 2012).

Procedural steps

3.1 Provide each study country team with access to the SQP (Saris et al., 2011) coding system.

3.2 A team member from each study country uses the SQP program to provide codes for each item in the target country's translated questionnaire.
  3.2.1 SQP codes refer to formal characteristics of items including:
  - Characteristics of the survey question, including the domain in which the variable is operating, (e.g., work, health, politics, etc.), the concept it is measuring (e.g., feeling, expectation, etc.), whether social desirability bias is present, the reference period of the question (past, present, future), etc.
  - The basic response or response scale choices (e.g., categories, yes/no scale, frequencies, level of extremeness, etc.).
  - The presence of optional components; instructions of interviewers, of respondents, definitions, additional information and motivation.
  - The presence of an introduction in terms of linguistic characteristics such as number of sentences, words, nouns, adjectives, subordinate clauses, etc.
  - Linguistic characteristics of the survey question.
  - Linguistic characteristics of the response scale.
  - The characteristics of the show card, if used.

3.3 SQP coding can also be used in the process of designing the source questionnaire.

3.4 The team dealing with SQP coding will then compare the SQP codes in the target language(s) and the source language.
  3.4.1 Differences in SQP coding resulting from mistakes should be corrected.
  3.4.2 No action is needed for true differences that are unavoidable (e.g. number of words in the introduction).
  3.4.3 True differences that may or may not be justified necessitate discussion between the central team and the national team, with possible change in translation necessary.
  3.4.4 True differences that are not warranted (e.g., a different number of response categories between the source and target

language versions) require an amendment to the translation as submitted.

Lessons Learned

3.1 In Round 5 of the ESS, SQP coding produced valuable information that allowed to detect deviations in translations that – had they been undetected – would have affected the quality of the items as well as the design of experiments (European Social Survey, 2012).

3.2 See ESS Round 6 SQP Guidelines (European Social Survey, 2012, November 6a) and Codebook (European Social Survey, 2012, November 6b) for further detail.

4. Translation assessment using focus groups and cognitive interviews with the target population.

Rationale

Various pretesting methods using both focus groups and cognitive interviews can be used to gain insight into the appropriateness of language used in survey translations.

Procedural steps

4.1 Focus groups can be used to gain target population feedback on item formulation and how questions are perceived (Schoua-Glusberg, 1988). They are generally not suitable for assessment of entire (lengthy) questionnaires. To optimize their efficiency, materials pertinent for many items can be prepared (fill-in-the blanks, multiple choice, etc.) and participants asked to explain terms and rate questions on clarity. At the same time, oral and aural tasks are more suitable than written when target population literacy levels are low or when oral/aural mode effects are of interest.

4.2 Cognitive interviews allow for problematic issues to be probed in depth, and can identify terms not well understood across all sub-groups of the target population.

4.3 Protocols should be developed and documented for all types of pretests, with particular care toward designs to investigate potentially concerning survey items (see Pretesting).

4.4 Interviewer and respondent debriefings can be used after all types of pretests, with full documentation of debriefing, to collect feedback and probe comprehension of items or formulations.

Lessons learned

4.1 Focus groups and cognitive interviews are useful for assessing questions in subsections of the target population. For example, focus groups conducted to validate the Spanish translation of the U.S. National Health and Sexual Behavior Study (NHSB) revealed that participants did not know terms related to sexual organs and sexual behaviors considered unproblematic up to that point (Schoua-Glusberg, 1988).

4.2 Interviewer and respondent debriefing sessions are valuable opportunities for uncovering problematic areas in translations. Debriefing sessions for the 1995 ISSP National Identity module in Germany revealed comprehension problems with terms covering ethnicity and confirmed cultural perception problems with questions about “taking pride” in being German (Harkness et al., 2004).

4.3 Tape recording of any pretesting allows for behavioral coding for particular questions of interest.

4.4 If computer-assisted pretesting is used, paradata, such as time stamps and keystroke data, can be used to identify items that are disrupting the flow of the interview, and may be due to translation issues (Kreuter, Couper, & Lyberg, 2010).

5. Translation assessment using quantitative analyses.

Rationale

Textual assessment of translation quality does not suffice to indicate whether questions will actually function as required across cultures; statistical, quantitative analyses are required to investigate the measurement characteristics of items and to assess whether translated instruments perform as expected. The central aim is to detect bias of different types that distort measurement systematically. Statistical tests can vary depending on the characteristics of an instrument, the sample sizes available, and the focus of assessment (for general discussion, see Geisinger (1994), Hambleton (1993), Hambleton, Merenda, & Spielberger (2005), Hambleton & Patsula (1998), van de Vijver (2003), van de Vijver & Hambleton (1996); van de Vijver & Leung (1997)).

Procedural steps

5.1 Variance analysis and item response theory can be used to explore measurement invariance and reveal differential item functioning, identifying residual translation issues or ambiguities overlooked by reviewers (Allalouf, Hambleton, & Sireci, 1999; Budgell, Raju, & Quartetti, 1995; Hulin, 1987; Hulin, Drasgow, & Komocar, 1982).

5.2 Factor analysis (adapted for comparative analyses: exploratory factor analysis or, confirmatory factor analysis), and multidimensional scaling can be used to undertake dimensionality analyses (Fontaine, 2003; Reise, Widaman, & Pugh, 1993; van de Vijver & Leung, 1997). See Statistical Analysis Chapter for more information

5.3 For the evaluation of individual items, item bias can be estimated using multitrait, multimethod procedures (MTMM), as described in Saris (2003) and Scherpenzeel and Saris (1997).

Lessons learned

5.1 Some procedures like SQP used in the ESS (Saris et al., 2011) rely on intensive analyses of questions collected (like a corpus in linguistics). However, the questions accepted as input in the corpus were not systematically evaluated using standard quality inspection such as checking for double barreled or double negation or response scales that do not fit the question etc. Thus the scores obtained might be biased and researchers should carefully use such systems.

5.2 Where scores are relevant (e.g., in credentialing tests), a design is needed to link scores on the source and target versions (Geisinger, 1994).

5.3 The emphasis placed on quantitatively assessing translated instruments and the strategies employed differ across disciplines.
   5.3.1 Instruments that are copyrighted and distributed commercially (as in health, psychology, and education) are also often evaluated extensively in pretests and after fielding.
   5.3.2 Some quantitative evaluation strategies call for a large number of items (e.g., item response theory) and are thus unsuitable for studies that tap a given construct or dimension with only one or two questions.
   5.3.3 Small pretest sample sizes may rule out strategies such as multidimensional scaling and factor analysis.
   5.3.4 Some assessment techniques are relatively unfamiliar in the social sciences (e.g., multitrait multimethod (MTMM)).

5.4 Post hoc analyses that examine translations on the basis of unexpected response distributions across languages are usually intended to help guide interpretation of results, not translation refinement. Caution is required in using such procedures for assessment because bias may also be present when differences in univariate statistics are not.

5.5 For multi-wave studies, document any post-hoc analyses for consideration when carrying out future translations.

Appendix A

Use of TVFF in Assessment through Verification

Figure 5 displays an additional component of the TVFF discussed in Translation: Overview, Appendix A, which permits documentation of the external verification process by an external reviewer. In the ESS, since Round 5, the firm cApStAn has been performing verification of each target language's translation and documented the intervention category and any commentary in the TVFF below (European Social Survey, 2014b).

[For clarification: the abbreviation is sometimes using brackets "(T)VFF" and sometimes not "TVFF": in the case of the ESS, the national teams have the choice to use this excel file for their translations ("T") too – but it is used for verification ("VFF") in all cases; this optional use for translation is mirrored by retaining the T in these guidelines.]

Figure 5 – Verification Area of the TVFF, reserved for the verifiers

MENTS TER ICATION	VERIFIER'S VERSION	VERIFIER INTERVENTION CATEGORY	VERIFIER'S RATIONALE	FOLLOW-UP REQUIRED?	COUNT
	PADUOTI RESPONDENTUI 46 KORTELĘ	OK			
ve Prašau cite, kada karta lkoholio enį, nį, enį ar ienį	Prašau prisiminkite, kada pastarąjį kartą gėrėte alkoholio pirmadienį, antradienį, trečiadienį ar ketvirtadienį.	Register/Wording	Ver suggests using the alternative translation as per comments after adjudication (col. K), as it sounds more natural and the meaning remains the same. BT of suggestion: Please remember/recall the last time.... Ver suggests to translate "think" as "remember" as in the comment in col. K, because it sounds more natural in this context.	Requires follow-up	We would 'recall/rem context, a

After the external verification is complete, the TVFF is returned to the national teams. These then use the blue columns of the TVFF in Figure 6 ("country comment") below to review the verifier interventions, and, for suggested changes, either accept the change or reject the change with justification.

Figure 6 – Post-Verification Area of the TVFF, reserved for national teams

FOLLOW-UP REQUIRED?	COUNTRY COMMENT	CST
Requires follow-up	We would also go for 'recall/remember' in this context, as verifier suggests.	ok

Appendix B

Definitions of Verifier Intervention Categories in Assessment through Verification *(verification by cApStAn for the ESS)*

OK	**No intervention** is needed. The verifier has checked and confirms that the text element or segment is equivalent to source, linguistically correct, and – if applicable – that it conforms to an explicit translation/adaptation guideline.
ADDED INFORMATION	Information is present in the target version but not in the source version, e.g. an explanation between brackets of a preceding word.
MISSING INFORMATION	Information is present in the source version but omitted in the target version.
CONSISTENCY	1. **Within-item consistency**: repetitions or literal matches and/or synonymous matches that occur in the source version of an item should reflect the same pattern in the target version. If a word or expression is used consistently across the source questionnaire, the same level of consistency should be reflected in the translations, unless fluency is affected. 2. **Across-item consistency**: unless fluency is affected, recurring elements such as response categories or prompts that occur in a number of items should always be translated the same way, measurement units should be written the same way, etc.
ADAPTATION OR CULTURAL ISSUE	An adaptation is an intentional deviation from the source version made for cultural reasons or to conform to local usage. They should be agreed by the ESS translation team at GESIS and the ESS-ERIC HQ at City University London. An adaptation or cultural issue occurs when an adaptation would be needed but was not made, or when an inappropriate or unnecessary adaptation was made.
MISTRANSLATION	A wrong translation, which seriously alters the meaning. A <u>mistranslation should always be reported with an explanatory back-translation and/or accompanied by an English rendition of what the incorrect target version says</u>. Note: a vague or inaccurate translation should rather be classified as a Register/Wording issue (or sometimes a Grammar/Syntax issue). This category may cover cases where the source has been misunderstood, but also copy/paste errors that unintentionally result in a wrong text element or segment.
REGISTER / WORDING ISSUE	1. *Register*: difference in level of terminology (scientific term >< familiar term) or level of language (formal >< casual, standard >< idiomatic) in target versus source. 2. *Wording*: inappropriate or less than optimal choice of vocabulary or wording in target to fluently convey the same information as in the source. This category is used typically for vague or inaccurate or not quite fluent translations.
GRAMMAR / SYNTAX ISSUE	1. *Grammar*: grammar mistake in the target language, e.g. wrong subject-verb agreement, wrong case (inflected languages), wrong verb form. 2. *Syntax*: syntax-related deviation from the source that affects fluency, or other syntactic problems due e.g. to overly literal translation of the source; any syntax error in the target language.
MINOR LINGUISTIC DEFECT	Typo or other linguistic defect (spelling, grammar, capitalization, punctuation, etc.) that does not significantly affect comprehension or equivalence.
LEFT IN SOURCE LANGUAGE	A text element or segment that should have been translated was left in source language.
ANNOTATION NOT FOLLOWED	An explicit translation/adaptation guideline for a given text element or segment given in an annotation was overlooked or was not addressed in a satisfactory way.
ALERT NOT REFLECTED	A late change made to the source questionnaire – released as an 'Alert' – has not been reflected in the target version.
LAYOUT / VISUAL ISSUE	A deviation or defect in layout or formatting: disposition of text and graphics, item labels, numbering/lettering of questions and, response categories, styles (**boldface**, <u>underlining</u>, *italics*, UPPERCASE), legibility, tables, number formatting (decimal separators, "five" versus "5"), etc. This category will only be used if submitted translations are already formatted.

Source: ESS Round 7 Verification Instructions (European Social Survey, 2014b).

Translation: Tools

Janet Harkness, Dorothée Behr, and An Lui, 2010
Updated by Brita Dorer and Peter Ph. Mohler, 2016

Introduction

This section discusses tools that support survey translation, including:

- Standard reference sources
  - Dictionaries, thesauri, and other hardcopy reference materials
  - Internet and Web-based reference materials

- Standard aids
  - Checklists
  - Listservers and newsgroups
  - Standard translator procedures, such as consistency procedures

- Templates for the translation process and translation output

- Technological support, such as translator software
  - Translation Memory (TM)
  - Terminology and Alignment tools
  - Concordances
  - Tools supporting the entire translation workflow

Appendix A provides a description of various translation tools.

Increasingly, large-scale international survey translation efforts for multinational, multicultural, or multiregional surveys, which we refer to as 3MC surveys, combine source document production with that of translated versions. The source text is then entered into a content management system which anticipates the needs and documentation of later production steps in other languages (Bowker, 2002). In order to be more inclusive, the guidelines following do not assume such a system; they do, however, include consideration of the technological components that would be available in an integrated document production and management system (Harkness, Dinkelmann, Pennell, & Mohler, 2007).

Tools and aids for translation can be provided by the translation project coordinator or can be a normal part of a translator's own toolkit. Who provides what may vary by project. A project might, for example, require translators to use project-specific software to produce translations, as is the case with the Survey on Health, Ageing and Retirement in Europe (SHARE) (Amin & Balster, 2010). Translation aids can also be developed using Translation: Overview, Appendix A to help translators identify common missteps.

Guidelines

1. **Identify relevant materials and tools, provide them to translators, and instruct, as necessary, translators and other translation team members on their use.**

Rationale

The more relevant the information and support that competent translators receive, the better they can meet the needs of a project. Other translation team members should also know about the tools and materials used in developing the translation. Depending on project organization, they will also need to use some of the tools (e.g., templates).

Procedural steps

1.1 Consider the following materials:

1.1.1 The website (intranet and/or internet) of the survey project, if it provides background information and documentation of the project.

1.1.2 The entire questionnaire, even if only parts of it require translation. This enables translators to:
- See the context in which the parts to be translated belong.
- Plan for consistency.

1.1.3 Any available sections already translated that have been vetted for quality.
- This contributes to consistency.
- Material not yet vetted for quality may also be provided but must be considered for re-use with great caution.

1.1.4 A bilingual glossary for any terms or phrases whose translation has already been established.
- This helps to ensure compliance with required translations and promotes consistency.
- It supports the review and copy-editing phases.

1.1.5 A style sheet guide, if relevant, detailing how to treat standard components of the source text (e.g., formats, use of bolding and italics).

1.1.6 Tracking documents that list major recurring elements and their location.
- These can be produced automatically as part of a content management system and can be created during development of the source questionnaire. Project coordinators would set the parameters for what should be included.
- They may also be part of translation software.

- In modestly funded projects, tracking documents can be developed manually.

1.1.7 Quality checklists, created for each country's final copy-editing effort. Include frequent or likely oversights in the checklist (e.g., "Check the order of answer categories"). As an example, see the European Social Survey (ESS) Translation Quality Checklist (European Social Survey, 2014c).

1.2. Consider translation tools. A distinction should be made between translation software readily available on the market – that is, not specifically designed for questionnaire translation – and tools that are specifically developed for survey translation needs. Appendix A describes in detail both types of translation tools.

Lessons learned

1.1 If existing translated material that has not been vetted for quality is made available to translators, coordinators must decide whether the translators will be able to assess its quality accurately. These issues may also arise when translators access "parallel texts" (e.g., texts from other surveys) in the target language. These parallel texts might include very similar questions or include translations for standard components such as response scales. Researchers need to be aware that existing translations may not be appropriate for their new purposes.

1.2 The purpose of various tools and procedures called for in survey research may not be self-evident to those involved in translation production; the translation staff may need to be briefed regarding their purpose and use.

2. Provide translators and others involved in the translation with documentation tools and specifications and require them to use them.

Rationale

Documentation is part of the translation quality assurance and control framework at local and general project levels. Providing thorough documentation of decisions, problems, and adaptations at each step of the translation process guides and enhances subsequent steps. Documentation tools and specifications can ensure that each participating unit provides systematic and comparable documentation.

If the project uses a text content management system, translation documentation may be part of the development of the source document.

Procedural steps

2.1 Clearly identify what requires translation and what does not.
   2.1.1 Some work platforms allow the user to freeze sections that should not be translated.

2.2 Produce translation templates that align source text segments, target text fields, and comments fields (see Translation: Overview, Appendix A).
   2.2.1 Questions, instructions, and response scales are examples of obvious source text segments.
   2.2.2 Subdivisions in the template, at least to sentence level, are often useful.
   2.2.3 A simple MS Word or Excel table, produced manually, may suffice (an example of an Excel-based template, the Translation and Verification Follow-up Form (TVFF), is presented in Translation: Overview, Appendix A).
   2.2.4 Translation software and content management systems may produce templates automatically.

2.3 Develop translation aids using Translation: Overview, Appendix C (Causes of Mistranslation) to help translators identify common missteps.

2.4 Provide instructions for translators and any other users on how to use the templates and how to document. For example, clearly explain the kinds of information expected in any comments field (see the example of the 'ESS Verification Instructions' that also contain a section explaining the use of the TVFF (European Social Survey, 2014b).

2.5 Hold meetings to merge template inputs. Since individual team members fill their templates, this allows them to compare options, notes, or comments (see Translation: Overview).

2.6 Pass final output from one phase on in a modified template for the next phase of work.

Lessons learned

2.1 The following issues apply in particular to the manual production of templates:
   2.1.1 The manual production of templates, including the source text, is labor-intensive and calls for care. In many cases, it may be the only option. As relevant, budget for the time and effort to produce translation templates manually. Involve at least two

suitable people with adequate bilingual proficiency and proofreading skills for the final proofreading effort (one reading out, the other checking).

2.1.2 Remember to check layout and format issues, not just wording.

2.1.3 Working between different source versions of a question and different translated versions within or across languages can be complicated. Any version control requires a tracking system to identify which elements should or do differ across versions.

2.1.4 Although, ideally, template production should begin after the source text is finalized, this may not always be feasible. If production of the templates starts prior to source text finalization, a tracking system for version control of templates is essential to check modifications at either the source or target text levels.

2.1.5 A procedure and protocol for alerting locations or teams to changes in either source documents or translation requirements is needed. For example, in a centrally organized project, the source text may be modified after templates have been sent out to translating locations (countries). Locations need to be able to recognize unambiguously what needs to be changed and then incorporate these changes into their templates (or at least into their translations). In the ESS 'alert' system, for example, both the source questionnaire and the translation template (that is, the TVFF), get updated and sent to all participating national teams as soon as an alert (that is, the announcement of a change in the already finalized source questionnaire) has been emitted.

2.1.6 Remember that copy-and-paste mistakes occur frequently. Technology (e.g., use of translation memory) may or may not make such errors more likely.

3. Provide translators with appropriate task instructions and briefing (see Translation: Building a Team).

Rationale

Provision of appropriate briefing and instructions helps translators and other team members understand what is required of them.

Procedural steps

3.1 See Translation: Building a Team.

Lessons learned

3.1 Provide for adequate training not only on the translation procedures to be followed but also on the translation templates and especially translation tools to be used. The more complex and demanding these are the more elaborate training activities need to be. These can, for instance, consist in webtraining, in-personal training or presentations or easy to use written training material. As an example, the ESS lays out its translation strategies in its Translation Guidelines (see e.g. European Social Survey, 2014), and the translation template, the TVFF, and its use are described in detail in a separate Verification Instructions document (see ESS Round 7 Verification Instructions (European Social Survey, 2014b)).

4. Consider networking translation teams within the project.

Rationale

Consultation within a language family can be helpful for all. Consultation across language families can also be of benefit, since some generic issues are shared by rather diverse languages and cultures. Although research on this is sparse, recent work suggests that a reasonably wide range of languages and cultures face similar translation challenges (Harkness et al., 2007).

Procedural steps

4.1 Decide how collaboration between teams sharing one language or translating into similar language groups is organized.
  4.1.1 If it is to be documented, decide on the template and detail required.
  4.1.2 Official collaboration and official documentation help to unify practices across and within projects.

4.2 Set up a protocol and schedule for sharing experiences or solutions and documenting these. Procedures described in Translation: Shared Language Harmonization may be useful.

Lessons learned

4.1 The publication of collaborative benefits, procedures and successful outputs experienced within one translation group (that is, the teams translating into one 'shared language') may inspire other groups that have not considered such collaboration. This argues strongly for documentation of work undertaken, even if it is not an official project requirement.

4.2 Even if the languages for which they produce translations differ considerably from one another, researchers may find numerous common difficulties in translating out of the source language (Harkness et al., 2007). In general, to the extent possible, any collaboration between national teams / different locations may be useful.

4.3 If researchers fielding in different regional forms of a "shared" language do not collaborate, many differences across versions may result that could otherwise have been avoided (see Translation: Shared Language Harmonization).

5. Make tools a deliberate part of the quality assurance and control framework for developing and checking the translated questionnaire. If possible, integrate this development with that of the source questionnaire.

Rationale

Tools make it easier to check that procedures are implemented and facilitate checking the quality of outputs at various stages of translation production.

Procedural steps

5.1 Determine the translation production budget and the budget available for tools of various kinds.

5.2 Identify tools of value for the procedures to be undertaken and identify outlay for each of these. A number of these are identified in the present section; more are discussed in Appendix A.

5.3 Obtain or create tools to be used for the translation procedures.

5.4 Train those using the tools on their use well in advance; monitor performance as appropriate, and refresh training as needed from time to time.

Lessons learned

5.1 Tools need not be expensive and technologically sophisticated in order to work.
5.2 Some tools will be familiar and seen as standard aids by the translating team, while others may be unfamiliar. Good briefing and instructions will foster proper and more extensive use of tools.

5.3 It is useful to point out the risks associated with tools as well as their advantages (e.g., "copy and paste" can be useful and can go wrong).

5.4 Multilingual projects should investigate management systems which manage both source questionnaire development and translation development. An example of an integrated tool for questionnaire translation and workflow is the Translation Management Tool (see Appendix A).

Appendix A

A list and description of translation tools

Dictionaries: There are many kinds of dictionaries and related textbooks. Good use of dictionaries requires knowledge of their strengths and weaknesses, familiarity with the way in which dictionary entries are structured, and familiarity with the abbreviations and descriptive labels used in entries. In all instances experienced translators ought to be familiar with the key relevant dictionaries for their language pairs and their area of work and know how to read and use dictionary entries.

- Monolingual dictionaries
  - Source language dictionaries
    Monolingual dictionaries list and explain the different typical meanings a source language word may have in different contexts. They may help translators check what a word or term meant in a particular context.
  - Monolingual target language dictionaries
    (Monolingual) Target language dictionaries may help clarify possible meanings in the target language and provide collocations (usual word combinations). They may also offer synonyms.
- Bilingual dictionaries
  - General bilingual dictionaries
    These dictionaries list under one entry the associated terms in another language which correspond to the various meanings possible for that term. Experienced translators may use these dictionaries as checking tools or to remind themselves of definitions they may have forgotten. Inexperienced translators may mistakenly think such dictionaries can provide them with a correct word to use which they do not already know. However, if a translator does not know a word, it is dangerous for her or him to use it on the basis of having found it in a dictionary.
  - Terminological or specialized dictionaries
    Bilingual dictionaries can be especially useful when it comes to subject-specific terminology (e.g., medical terminology). However, languages differ in the extent to which they use technically correct terminology for subjects or prefer more everyday terms (compare "He has athlete's foot" to "He has tinea pedis"). Translators should not use terms with which they are not familiar unless they have solid evidence that these are the right terms for their needs. They may need to consult experts on a final choice. The more information a dictionary offers on the context in which suggested equivalents are embedded, the better for the translator.
- Spelling dictionaries
  Spelling dictionaries are useful at copyediting and proofreading stages undertaken by translators. Incorrect spelling (and punctuation, layout, etc.) can trip up both interviewers and respondents when reading questions. Incorrect spelling may also create a poor impression of the project in

general. Spellcheckers included in word processors are useful but manual proofreading remains a necessary final step to recognize errors a machine cannot (e.g., form/from, on/in, healthy/wealthy)

- Online dictionaries
  There are numerous online dictionaries and thesauri, both monolingual and bilingual. See, for example:
  http://www.yourdictionary.com/ or http://www.lexicool.com/ or http://www.wordreference.com/ .

Thesauri: Thesauri group together words of similar or related meaning. They can be helpful for finding the most appropriate word after looking up a related word known not to be quite right. The user may know the word passively and recognize it among those offered. Since a thesaurus only offers synonyms and does not define words, extensive knowledge of the language is required to identify the starting place for a search and to decide whether a term found is appropriate.

Word processors such as MS Word also offer modestly comparable functions as "Synonyms" and "Thesaurus" in at least some languages.

Internet: The Internet makes it possible to see multiple examples of words in context and to check how frequently they seem to be used (e.g. through Google Research). However, the Internet offers usage without quality assurance. A particular word might only appear on translated websites or on websites from countries that do not use the language in question as a first language. The word or phrase then found may not be correct for the target language or for the level of diction required for the survey. So, sites such as Google Research should always be used with caution and not without double-checking the nature of the site from which one intends to extract information.

The Internet can be used to check:

- The frequency of occurrence of particular phrases or words.
  But again, this does not necessarily have to tell a lot about the real use of a term or expression because, for instance: (1) sometimes certain websites are linked to each other and appear more often than others, (2) the context in which a term or expression is found does not always correspond to the context you are interested in – but is nevertheless counted as a hit, (3) the websites using a certain term or expression may be translated, so no guarantee of correct language use at native-speaker level.
- The contexts in which words appear.
- Official terminology versus everyday terminology as evidenced by the contexts in which occurrences are found.

Listservers and newsgroups: Translators often use translation-related listservers and/or newsgroups to post questions and enquiries. Survey translation needs might not be well addressed but questions about general usage (e.g., regional terms or levels of vocabulary) could be answered. Some languages are likely to be better served than others. A list of translation-related newsgroups can be found at: http://www.translationjournal.net/journal/00disc.htm .

Translation software: We distinguish below between general translation software readily available on the market – that is, not specifically designed for questionnaire translation – and tools that are specifically developed for survey translation needs.

1. General translation software, not specifically designed for survey translations

Demonstration versions of general translation tools are usually available on software producer websites. Companies also usually offer to consult on prospective customers' needs. The usefulness of any of these tools for a given project depends on many factors, including the repetitive nature of the project, the scope or complexity of the project, the suitability of the tools for the specifics of a project, the budget available, and the ability of staff to work with such tools.

(a) *Computer Assisted Translation* tools (http://www.translationzone.com/products/cat-tools/) help to produce consistent translations across languages and time by relying on Translation Memories. For instance, they provide translators with standard phraseology such as response scales used over and over in a survey. Depending on the product, they can also provide systematic documentation of the translation process including document and project management. Survey agencies and international projects often use proprietary translation tools. There are, however, also tools on the market such as SDL Trados or Deja Vue that can be adapted to comparative survey translation. Some examples of Computer Assisted Translation tools are:

- Across: http://www.across.net/en/
- Déjá Vu: http://www.atril.com/
- MetaTexis: http://www.metatexis.com/
- RR Donnelley: http://www.rrdonnelley.com/languagesolutions/
- SDL Trados: http://www.sdl.com/en/
- Transit: http://www.star-group.net/ENU/group-transit-nxt/transit.html
- Wordfast: http://www.wordfast.net/

(b) *Fully automated translation systems / Machine translation* such as Google Translate are explicitly not recommended here as they do not provide procedures for consistent translation (translation memory) and process quality control via systematic documentation. Also, these systems are not able to consider the context, which is a crucial element for finding optimal translation solutions. Nor do they allow to optimize translation systematically as it is done via the TRAPD process.

(c) Different types or elements of translation-related software available on the market:

- Translation memory: A translation memory is a database that stores translations, as they are produced, for future use. "Future use" can be within the same translation, only a few minutes after first being produced or could be for an entirely new translation task months later. The source text segment and the corresponding target text segment produced as a translation are saved as a "translation unit". A segment may consist of a few words, whole sentences, or, depending on the material involved, extended stretches of text. Translation memories display source and target text segments alongside each other and thus facilitate review. In addition, they facilitate making sure that all segments up for translation have been translated because the system runs through the entire text automatically without leaving any gaps.

  When translation memory is used, it offers "100% matches" for completely identical and previously translated source text segments and "fuzzy matches" for similar, but not identical source text segments previously translated. Depending on the software used, the degree of match required in order for it to be presented to the translator can be defined. Translators accept or reject matches offered. Whatever a translator may produce as a new translation or revise by modifying an existing translation also becomes part of the dynamically created and expanding translation memory. Translations produced using translation memory can thus benefit from technology but must be driven by translator decisions. The translation memory software simply presents (offers) pre-existing translation choices for consideration. There is no quality component with regard to how appropriate the translation offered is for a specific new context. It is therefore essential that the memory has been created through submitting good translations – and that the staff translating and using the software is highly qualified and experienced (see Translation: Team).

Properly vetted translation memories can be useful for texts that are highly repetitive and where consistency of repetitive elements is a crucial issue. They can also be of value with texts that are used repeatedly but with slight modifications.

- Terminology tool: A terminology tool stores multilingual terms alongside additional information on these terms, such as a definition, synonyms, and context examples. Often, a terminology tool is used alongside a translation memory as a source of richer information.
- Alignment tools: Alignment tools can be used to compare a source text and its translation and match the corresponding segments. With alignment tools it is possible to align translations produced post-hoc, that is, after a translation has been finalized, and these can then be imported into a translation memory and be available for future translations. Alignment

tools are typically used when a Translation memory could not be used until finalization of a translation, thus allowing to have the final version and not only draft version of a translation in the database.

- Translation memory versus machine translation:
Translation memories do not 'translate' but just offer similar translations (if these do exist) from a database, but that need to be worked upon by a competent and experienced translator.
Translation memories are built upon the basis of human translation. Machine translation, per se, is a fully automatized process.
Quality translations never rely on machine translation alone. Survey questions are a complex text type with multiple functions and components; as complete and easy understanding by the average population is of utmost importance, they need to respond to communication requirements also in the target languages. As a result, any reduction of human involvement in the decision-making process of survey translation is ill advised.
- Concordance function: This software feature (existing in Translation memory software) allows the translator to search for terms within the translation memory: the contextual usage of a given word is then displayed, much as in a concordance.
- Corpora: A *corpus* is "a large collection of authentic texts that have been gathered in electronic form according to a specific set of criteria" (Bowker and Pearson 2002:9). The relevance and usability of corpora for research stems from three essential characteristics. Firstly, corpora present language 'as is', i.e. they empirically show how language is actually used. Secondly, corpora typically comprise very large collections of texts, which enables statistical analysis and inferencing about frequencies of various phenomena in language use. Thirdly, corpora in electronic formats are searchable and often equipped with various tools (such as concordances, frequency lists, key words in context etc.) and, as such, can be a useful source of insights about language in use.

Corpora may be based on various *design criteria*. For instance, they may comprise texts of specific genres, or texts from specific authors, fields of knowledge or historical periods. Other corpora aim to provide a broad cross-section of various genres, styles and authors. Many of the latter are termed 'national corpora' (e.g. the British National Corpus) and are usually compiled by academics with public support in an effort to represent the 'general language' of a particular country, area or group.

Corpora may be *monolingual* (such as most national corpora) or *multilingual.* Multilingual corpora usually contain parallel texts and, as such, are known as *parallel corpora*. Texts in a parallel corpus may represent original writing on similar topics in multiple languages (e.g. news collections in various languages) or the different language versions may be interrelated (e.g. texts in the original language *aligned* with their translations into various languages).

The latter are called *translational corpora* and provide insights into the characteristics of translated texts and the so-called *'translatese'* in various language pairs or groups. One of the largest such searchable collections is EUR-Lex, the collection of European Union law in EU official languages.

Corpora may contain texts produced by native speakers or those generated non-native speakers, such as language learners. Learners' corpora help researchers to identify typical errors and enhance language teaching materials or curricula on this basis.

Moreover, while corpora started off with written texts, there has been an increasing effort to compile spoken language corpora (including corpora of interpreted speech, such as EPIC, the parallel corpus of European Parliament speeches and their simultaneous interpretations).

Corpora have found multiple uses in areas such as linguistics (language features such as lexical density, semantic prosody etc.), language learning, discourse analysis (incl. critical discourse analysis), translation studies etc..

There is a number of *corpus analysis tools* (known as *concordancers*), which can interrogate corpora in various ways. They can be applied to existing public and non-public corpora or to specific corpus-based research projects. Queries are facilitated if corpus elements have been previously *tagged*, i.e. marked for various characteristics, such as parts of speech, grammatical tense or other relevant characteristics.

3MC surveys can be informed by *corpora of survey questionnaires* with translations from various research projects, particularly if the translated versions are official and have undergone a rigorous procedure, such as some version of '*committee approach' or TRAPD* (see above and Harkness 2003). At present (early 2016), no such corpora are available. However, with such corpora in place, researchers could reuse survey questions and their existing approved translations (to enhance comparability within and across surveys), and avoid translating the same questions again (to reduce costs and eliminate errors in new translations). Such corpora could also be a useful learning resource for item designers, questionnaire translators and researchers studying 'survey.

Another idea is to compile question banks from various surveys, in a specific language or regardless of language. Such an attempt has been undertaken by GESIS-Leibnitz Institute for the Social Sciences (Germany) which is running a databank of survey items and scales in social sciences (http://zis.gesis.org). Such question banks could also provide a useful starting point for creating a translational corpus of survey questions.

- Translation management: In addition to facilitating translation, tools are available that facilitate project management of the entire translation workflow. Most of the commercial packages listed in Further Reading offer such management tools. Also the Translation Management Tool offers support for managing the whole translation workflow (see below).

Translation software specifically designed for survey translations
To our knowledge, there have been some tools to facilitate questionnaire translation, but rather for internal use within some institutes or projects.

As these are not publicly searchable and not open to public use, we would like to concentrate on one particular tool in these Guidelines which has been developed specifically for questionnaire translation and is currently adapted in order to be useable for the team approach or TRAPD translation scheme. The so-called "*Translation Management Tool"*, as the name indicates, will not only be useable for the whole questionnaire translation process, including the TRAPD model plus quality assurance steps, but will also facilitate managing the whole translation workflow. CentERdata has been developing it for the Survey on Health, Ageing and Retirement in Europe (SHARE), which has been using this tool since its first wave (however, its predecessor, the "Language Management Tool", is a different product with some common feature).

CentERdata is now collaborating with the translating team of the ESS (European Social Survey, 2014) to make it useable for the rigorous ESS questionnaire translation scheme, consisting in the team approach following the TRAPD model. Once it has been developed, it will be useable online and references will be added here when it is available.

References

Acquadro, C., Jambon, B., Ellis, D., & Marquis, P. (1996). Language and translation issues. In B. Spilker (Ed.), *Quality of life and pharmacoeconomics in clinical trial* (2nd ed.) (pp. 575-585). Philadelphia, PA: Lippincott-Raven.

Allalouf, A., Hambleton, R. K., & Sireci, S. G. (1999). Identifying the causes of DIF in translated verbal items. *Journal of Educational Measurement, 36*(3), 185-198.

Amin, A., & Balster, E. (2010). *Managing translations in a single CAPI implementation across SHARE countries*. Paper presented at the International Workshop on Comparative Survey Design and Implementation, Lausanne, Switzerland.

Andreenkova, A. V. (2008). *Translating the instrument for a comparative survey in Russian*. Paper presented at the International Conference on Survey Methods in Multinational, Multiregional, and Multicultural Contexts (3MC), Berlin, Germany.

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., & Scholz, E. (2011). Questionnaire translation in cross-national survey research: On the types and value of annotations. *Methoden, Daten, Analysen (mda) 5*(2), pp. 157-179. Retrieved July 14, 2016, from http://nbn-resolving.de/urn:nbn:de:0168-ssoar-282120

Behr, D., & Shishido, K. (2016). The translation of measurement instruments for cross-cultural surveys. In: C. Wolf, D. Joye, T. Smith, Y-Ch Fu, (Eds.): *The SAGE Handbook of Survey Methodology*, London, UK: SAGE.

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

Bowker, L. (2002). *Computer-aided translation technology: A practical introduction*. Ottawa: University of Ottawa Press.

Bowker, L., & Pearson, J. (2002). *Working with specialized language: A practical guide to using Corpora*. London & New York: Routledge.

Braun, M., & Harkness, J. A. (2005). Text and context: Challenges to comparability in survey questions. In J. H. P. Hoffmeyer-Zlotnik & J. A. Harkness (Eds.), *Methodological aspects in cross-national research* (pp. 95-107). Mannheim, Germany: ZUMA. Retrieved April 18, 2010, from http://www.gesis.org/fileadmin/upload/forschung/publikationen/zeitschriften/zuma_nachrichten_spezial/znspezial11.pdf

Budgell, G. R., Raju, N. S., & Quartetti, D. A. (1995). Analysis of differential item functioning in translated assessment instruments. *Applied Psychological Measurement*, *19*(4), 309-321.

Dept, S., Ferrari, A., & Wäyrynen, L. (2008). *Comparative overview of test adaptation and translation verification procedures in three international surveys.* Paper presented at the International Conference on Survey Methods in Multinational, Multiregional, and Multicultural Contexts (3MC), Berlin, Germany.

Dept, S., A. Ferrari, & Wäyrynen, L. (2010). Developments in translation verification procedures in three multilingual assessments: A plea for an integrated translation and adaption monitoring tool. 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. 157-173). Hoboken, NJ: John Wiley & Sons.

Dorer, B. (2011*). Advance translation in the 5th round of the European Social Survey (ESS)*. *FORS Working Paper Series,* paper 2011-4. Lausanne: FORS.

Dorer, B. (2014). *European Social Survey, Round 5 - Report on changes to existing translations applied by national teams in the ESS Round 5 questionnaire*. Mannheim, ESS Translation Team. Retrieved October 25, 2014, from http://www.europeansocialsurvey.org/docs/round5/methods/ESS5_changes_to_existing_translations.pdf

European Social Survey (2012). *ESS Round 6 SQP Guidelines*. Barcelona, European Social Survey, Research and Expertise Centre for Survey Methodology Universitat Pompeu Fabra. Retrieved July 2, 2014, from http://www.europeansocialsurvey.org/docs/round6/methods/ESS6_sqp_guidelines.pdf

European Social Survey. (2012a). ESS6 SQP coding guidelines. Retrieved from http://www.europeansocialsurvey.org/docs/round6/methods/ESS6_sqp_guidelines.pdf

European Social Survey. (2012b). SQP codebook for all characteristics. Retrieved from http://www.europeansocialsurvey.org/docs/round6/methods/ESS6_sqp_codebook.pdf

European Social Survey (2014). *ESS Round 7 Translation Guidelines*. London: ESS ERIC Headquarters, Centre for Comparative Social Surveys, City University London. http://www.europeansocialsurvey.org/docs/methodology/ESS_R7_Translation_Guidelines_FINAL.pdf

European Social Survey (2014b). *ESS Round 7 Verification Instructions for National Coordinators*. London: ESS ERIC Headquarters, Centre for Comparative Social Surveys. http://www.europeansocialsurvey.org/docs/round7/methods/ESS7_verification_instructions.pdf

European Social Survey (2014c). *ESS Round 7 Translation Quality Check List: Proof-reading and copy-editing on your questionnaire*. London: ESS ERIC Headquarters, Centre for Comparative Social Surveys. http://www.europeansocialsurvey.org/docs/round7/questionnaire/ESS7_translation_quality_checklist.pdf

Fishman, J. A. (2003). Bilingualism with and without diglossia; diglossia with and without bilingualism. C. B. Paulston & G. R.Tucker (Eds.), *Sociolinguistics: The essential readings* (pp.359-366). Oxford: Blackwell.

Fontaine, J. (2003). Multidimensional scaling. J. A. Harkness, F. J. Van de Vijver, & P. P. Mohler (Eds.) *Cross-cultural survey methods*. Hoboken, NJ: Wiley-Interscience, 235-246.

Geisinger, K. F. (1994). Cross-cultural normative assessment: Translation and adaptation issues influencing the normative interpretation of assessment instruments. *Psychological Assessment*, *6*(4), 304.

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

Guillemin, F., Bombardier, C., & Beaton, D. (1993). Cross-cultural adaptation of health-related quality of life measures: Literature review and proposed guidelines. *Journal of Clinical Epidemiology*, *46(*12), 1417-1432.

Hambleton, R. K. (1993). Translating achievement tests for use in cross-national studies. *European Journal of Psychological Assessment*, *9*, 57-68.

Hambleton, R. K., & Patsula, L. (1998). Adapting tests for use in multiple languages and cultures. *Social Indicators Research: An International and Interdisciplinary Journal for Quality-of-Life Measurement*, *45*, 153-171.

Hambleton, R. K., Merenda, P. F., & Spielberger, C. D. (Eds.). (2004). *Adapting educational and psychological tests for cross-cultural assessment*. Mahwah, NJ: Erlbaum [Global refs has date of 2005, chapter has 2004]

Harkness, J. A. (2003). Questionnaire translation. In J. A. Harkness, F. van de Vijver & P. Ph. Mohler (Eds.), *Cross-cultural survey methods* (pp. 35-56). Hoboken, NJ: John Wiley & Sons.

Harkness, J. A. (2007). Improving the comparability of translation. In R. Jowell, C. Roberts, R. Fitzgerald & G. Eva (Eds.), *Measuring attitudes cross-nationally: Lessons from the European Social Survey* (pp. 79-94). London: Sage.

Harkness, J. A. (2008a). *Round 4 ESS translation strategies and procedures*. Retrieved April 16, 2010, from http://www.europeansocialsurvey.org/docs/round4/methods/ESS4_translation_guidelines.pdf

Harkness, J. A. (2008b). Comparative survey research: Goals and challenges. In E. D. de Leeuw, J. J. Hox, & D. A. Dillman (Eds.), *International handbook of survey methodolog*y (pp. 56-77). New York, NY/London: Lawrence Erlbaum Associates.

Harkness, J. A. (2010a). *Survey translation and the question of quality*. Paper presented at the International Workshop on Comparative Survey Design and Implementation, Lausanne, Switzerland.

Harkness, J. A., & Schoua-Glusberg, A. S. (1998). Questionnaires in translation. In J. A. Harkness (Ed.), *Cross-cultural survey equivalence* (pp.87-126). ZUMA-Nachrichten Spezial, 3. Mannheim, Germany: ZUMA.

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

Harkness, J. A., Chin, T-Z., Yang, Y., Bilgen, I., Villar, A., Zeng, W., et al. (2007). *Developing response scales in multilingual studies*. Paper presented at the Fifth International Workshop on Comparative Survey Design and Implementation, Chicago, IL.

Harkness, J. A., Dinkelmann, K., Pennell, B-E., & Mohler, P. Ph. (2007). Final report and translation tool specifications (Deliverable No. 5), *European Social Survey Infrastructure (ESSi) project: Improving social measurement in Europe*. Mannheim: ZUMA.

Harkness, J.A., Pennell, B-E., Villar, A., Gebler, N., Aguilar-Gaxiola, S., & Bilgen, I. (2008a). Translation procedures and translation assessment in the World Mental Health Survey initiative. In R. Kessler & T. B. Ustun (Eds.), *The World Health Organization Mental Health Survey* (pp. 91-113). New York, NY: Cambridge University Press.

Harkness, J. A., Schoebi, N., Joye, D., Mohler, P., Faass, T., & Behr, D. (2008b). Oral translation in telephone surveys. In J. M. Lepkowski, C. Tucker, J. M. Brick, E. de Leeuw, L. Japec, & P. J. Lavrakas, (Eds.), *Advances in telephone survey methodology* (pp. 231-249). Hoboken, NJ: John Wiley&Sons.

Harkness, J., Villar, A., Kephart, K., Behr, D., & Schoua-Glusberg, A. (2009). *Research on translation assessment procedures: Back translation and expert review.* Paper presented at the International Workshop on Comparative Survey Design and Implementation, Ann Arbor, MI.

Harkness, J. A., Villar, A., Kephart, K., Schoua-Glusberg, A, & Behr. D. (2009). *Survey translation evaluation: Back translation versus expert* *review.* Paper presented at the annual conference of the American Association for Public Opinion Research, Hollywood, FL.

Harkness, J. A., Villar, A., Kruse, Y., Branden, L., Edwards, B., Steele, C., & Wang, Y. (2009a). *Interpreted telephone survey interviews*. Paper presented at the International Workshop on Comparative Survey Design and Implementation, Ann Arbor, MI.

Harkness, J. A, Villar, A., Kruse, Y., Steele, C., Wang, Y., Branden, L., Edwards, B., & Wilson, C. (2009b). *Using interpreters in telephone surveys.* Paper presented at the annual conference of the American Association for Public Opinion Research, Hollywood, FL.

Harkness, J. A., Villar, A., & Edwards, B. (2010a). Translation, adaptation and design. 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, multicultural and multiregional contexts* (pp.117-139). Hoboken, NJ: John Wiley & Sons.

Harkness, J. A., Edwards, B., Hansen, S. E., Miller, D. R., & Villar, A. (2010b). Designing questionnaires for multipopulation research. 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, multicultural and multiregional contexts* (pp. 33-58). Hoboken, NJ: John Wiley & Sons.

Hönig, H. G., (1997). Konstruktives Übersetzen. (2. Durchges. Aufl.). Tübingen: Stauffenburg.

Hulin, C. L. (1987). A psychometric theory of evaluations of item and scale translations fidelity across languages. *Journal of Cross-cultural Psychology*, *18*(2), 115-142.

Hulin, C. L., Drasgow, F., & Komocar, J. (1982). Applications of item response theory to analysis of attitude scale translations. *Journal of Applied Psychology*, *67*(6), 818.

Johnson, T. P., O'Rourke, D., Chavez, N., Sudman, S., Warnecke, R., Lacey, L. et al. (1997). Social cognition and response to survey questions among culturally diverse populations. In L. Lyberg, P. Biemer, M. Collins, et al. (Eds.), *Survey measurement and process quality* (pp. 87-113). New York: Wiley & Sons.

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,* 282-296.

Krings, H. P. (1986). Was in den Köpfen von Übersetzern vorgeht: Eine empirische Untersuchung zur Struktur des Übersetzungsprozesses an fortgeschrittenen Französischlernern. Tübingen: Narr.

Kurzman, C. (2014). *World values lost in translation*. The Washington Post. Retrieved form https://www.washingtonpost.com/news/monkey-cage/wp/2014/09/02/world-values-lost-in-translation/

Kussmaul (Kußmaul), P. (1995). *Training the translator*. Amsterdam: John Benjamins.

Kußmaul, P. (2007). *Verstehen und Übersetzen: Ein Lehr- und Arbeitsbuch*. Tübingen: Narr.

Martinez, G., Marín, B. V., Schoua-Glusberg, A. (2006). Translating from English to Spanish: The 2002 National Survey of Family Growth. *Hispanic Journal of Behavioral Sciences*. *28*(4), 531-545.

McKay, R. B., Breslow, M. J., Sangster, R. J., Gabbard, S. M., Reynolds, R. W., Nakamoto, J. M. et al. (1996). Translating survey questionnaires: Lessons learned. *New Directions for Evaluation*, *70*(Summer), 93-105.

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. Statistical Research Division, Washington, DC: U.S. Bureau of the Census.

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, et al. (Eds.), *Total survey error in practice*. New York, NY: John Wiley & Sons.

Porst, R., & Jers, C. (2007). Die ALLBUS-„Gastarbeiter-Frage. Zur Geschichte eines Standard- Instruments in der Allgemeinen Bevolkerungsumfrage der Sozialwissenschaften (ALLBUS). *Soziale Welt*, *58*, 145-161.

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin*, *114*(3), 552.

Saldanha, G., & O'Brien, S. (2013). *Research methodologies in translation studies*. London & New York: Routledge.

Saris, W. (2003). Response function equality. In J. A. Harkness, F. J. van de Vijver, & P. P. Mohler (Eds.) *Cross-cultural survey methods*. Hoboken, NJ: Wiley-Interscience.

Saris, W., Oberski, D., Revilla, M., Zavala-Rojas, D., Lilleoja, L., Gallhofer, I., and Gruner, T. (2011). The development of the program SQP 2.0 for the prediction of the quality of survey questions. RECSM Working Paper 24.

Scherpenzeel, A. C., & Saris, W. E. (1997). The validity and reliability of survey questions: A meta-analysis of MTMM studies. *Sociological Methods & Research*, *25*(3), 341-383.

Schoua-Glusberg, A.S. (1988*). A focus-group approach to translating questionnaire items*. Paper presented at the annual meeting of the American Association for Public Opinion Research, Toronto.

Schoua-Glusberg, A. (1992). *Report on the translation of the questionnaire for the national treatment improvement evaluation study.* Chicago, IL: National Opinion Research Center.

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, multicultural and multiregional contexts* (pp.177-190). Hoboken, NJ: John Wiley & Sons.

Smith, P. B., Bond, M. H., & Kagitcibasi, C. (2006). *Understanding social psychology across cultures: Living and working in a changing world*. London: Sage.

Smith, T. W. (1992). Changing racial labels: From “Colored” to “Negro” to “Black” to “African American.” *Public Opinion Quarterly*, 56(4), 496.

Solano-Flores, G., & Nelson-Barber, S. (2001). On the cultural validity of science assessments. *Journal of Research in Science Teaching*, *38*(5), 553-573.

Sperber, A. D., Gwee, K. A., Hungin, A. P., Corazziari, E., Fukudo, S., Gerson, C., & Whitehead, W. E. (2014). Conducting multinational, cross-cultural research in the functional gastrointestinal disorders: issues and recommendations. A Rome Foundation working team report. *Alimentary Pharmacology & Therapeutics*, 40(9), 1094–102. http://doi.org/10.1111/apt.12942

Tanzer, N. K. (2005). Developing tests for use in multiple languages and cultures: A plea for simultaneous development. In R. K. Hambleton, P.F. Merenda, & C. C. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment*. Mahwah, NJ: Lawrence Erlbaum Associates.

van de Vijver, F. J. (2003). Bias and equivalence: Cross-cultural perspectives. In J. A. Harkness, F. J. Van de Vijver, & P. P. Mohler (Eds.), *Cross-cultural survey methods* (pp. 143-155). Hoboken, NJ: Wiley-Interscience.

van de Vijver, F., J. & Leung, K. (1997). *Methods and data analysis for cross-cultural research* (Vol. 1). Thousand Oaks, CA: Sage.

van de Vijver, F. J. R., Hambleton, P. M. (1996). Translating tests: Some practical guidelines. *European Psychologist*, *1*(2), 89-99.

van de Vijver, F. J. R., & Poortinga, Y. H. (2005). Conceptual and methodological issues in adapting tests. In R. K. Hambleton, P. F. Merenda, & C. D. Spielberger (Eds.), *Adapting educational and psychological tests for* *cross-cultural assessment* (pp. 39-63). Mahwah, NJ: Lawrence Erlbaum Associates.

Villar, A. (2009). Agreement answer scale design for multilingual surveys: Effects of translation-related changes in verbal labels on response styles and response distributions. *Survey Research and Methodology Program (SRAM)-Dissertations & Theses, 3.*

Weisberg, H. F. (2005). *The total survey error approach*. University of Chicago Press: Chicago.

Willis, G. B., Kudela, M. S., Levin, K. L., Norberg, A., Stark, D. S., & Forsyth, B. H., et al. (2010). Evaluation of a multistep survey translation process. In J. A. Harkness., M. Braun, B. Edwards, T. P. Johnson, L. Lyberg, P. Ph. Mohler, et al. (Eds.), *Survey methods in multinational, multicultural and multiregional contexts* (pp. 137-152). Hoboken, NJ: John Wiley & Sons.

Further Reading

Austermühl, F. (2014). *Electronic tools for translators*. Routledge, Abingdon/New York.

Behr, B., & Kuniaki, S. (2016). Translation measurement instruments for cross-cultural surveys. In C. Wolf, D. Joye, T.W. Smith, & Y. Fu (Eds.), *The SAGE Handbook of Survey Methodology*. Thousand Oaks, CA: Sage.

Cantor, S., Byrd, T., Groff, J., Reyes, Y., Tortolero-Luna, G. & Mullen, P. (2005). The language translation process in survey research: A cost analysis. *Hispanic Journal of Behavioral Sciences*, 27(3), 364-370.

Hutchins, J. (2003). The development and use of machine translation systems and computer-based translation tools. *International Journal of Translation*, *15*(1), 5-26.

Hutchins, J. (2005). Current commercial machine translation systems and computer-based translation tools: system types and their uses. International *Journal of Translation*, *17*(1-2), 5-38.

Joye, D., Schneider, S., & Wolf, C. (2016). When translation does not help: Background variables in comparative surveys. In C. Wolf, D. Joye, T.W. Smith, & Y. Fu (Eds.), *The SAGE Handbook of Survey Methodology*. Thousand Oaks, CA: Sage.

Matis, N. (2006). Translation project management. Retrieved February 2, 2016, from http://www.translation-project-management.com/translation-project-management

Mossop, B. (2014). *Revising and editing for translators, 3rd edition*. Routledge, New York.

Pérez, C. (2002). Translation and project management. *Translation Journal*, 6(4), February 2, 2016, from http://translationjournal.net/journal/22project.htm

Ruthven-Stuart, P (2000). Concordances. Retrieved May 03, 2010, from http://www.nsknet.or.jp/~peterr-s/concordancing

SDL Trados. (2010). *Translation memory*. Retrieved July 2, 2014, from http://www.sdl.com/technology/language-technology/what-is-translation-memory.html

Translation Project Management: http://www.translation-project-management.com/home (Retrieved 2 February 2016)

Adaptation

Peter Mohler, Brita Dorer, Julie de Jong, and Mengyao Hu, 2016
(2010 Version: Janet Harkness)

Introduction

The term adaptation, as used in this chapter, refers to the deliberate modification of a question or questionnaire to create a new question or questionnaire. It is also referred as asking different questions (ADQ) in Questionnaire Design.

Adaptation needs may be considered at different stages in different multinational, multicultural, or multiregional surveys, which we refer to as "3MC" surveys, and it is likely that some adaptation needs only become apparent during translation or during pretesting of a translated questionnaire. It is therefore not possible, in terms of the survey lifecycle, to identify a single unique stage as the stage at which adaptation needs might be recognized or addressed. This may differ, for instance, for different projects or surveys.

However, the most common procedure is to consider and address adaptation needs together with the translation process – therefore, the terms are often used in combination as in 'Translation and Adaptation'. This is also mirrored in the survey lifecycle.

Overall, there are different ways to integrate adaptation in the survey lifecycle, that is, in the production of survey instruments in the target languages. Examples are creating separate teams and experts for adaptation and translation; another possibility is to deal with adaptation at the same time as translation in a one-team framework: in this case, translation experts would also need to be experienced or knowledgeable in adaptation needs. In this chapter, we will focus on the former approach, where translation and adaptation are conducted through separate teams. It is highly recommended that the two teams collaborate and work closely together for questionnaire design. Translation team will be responsible for the same questions asked in both cultures (see asking same question and translating (ASQT) in Questionnaire Design), and adaptation team will mainly be responsible for asking different questions or asking questions in different ways to adapt new cultural needs (see asking different questions (ADQ) in Questionnaire Design).

Why adapt questions?

Adaptation may be made to the content, format, response scales, or visual presentation of any part of a question, questionnaire, or instrument. The purpose of adaptation is to better fit the needs of a new population, location, language, or mode, or any combination of these (Harkness, 2010b; Harkness, 2008b;

Harkness, Villar, & Edwards, 2010a; see also Instrument Technical Design and Translation: Shared Language Harmonization).

When developing new studies, researchers frequently modify questions that have been used in other studies and then use these modified versions. The motivations for such modifications may or may not be documented. Sometimes changes are related to adapting to meet new needs. However, sometimes changes are made simply because those making the changes consider them to result in a generally "better" version or instrument. This chapter focuses only on changes made to meet new needs as described above (adaptations).

In one language (monolingual) contexts, questions and questionnaires may be deliberately adapted for a variety of reasons. In longitudinal surveys, for example, wording might be updated to stay abreast with current usage; "wireless" could be replaced by "radio" (Smith, 2005), for example. Wording might also be changed to better reflect current social realities, such as adding social media as means of communication or the Internet as an information source in media usage questions (Smith, 2005). Changes might also be made to accommodate a new population; modifying vocabulary, presentation, and instructions to suit a child population rather than an adult one, for example.

In 3MC projects, adaptation is often related to the need to translate a questionnaire into another language in order to study new populations. In the following chapters, the terms "source language" and "target language" are used to distinguish between the language translated out of (the source language) and the language translated into (the target language).

In some projects, adaptations may already be anticipated in the source questionnaire, that is, the questionnaire on which other language versions are based and derived. Thus a source questionnaire question about pride in one's nationality, "How proud are you to be [nationality]*?"* anticipates a country-specific adaptation inside the square brackets, with each participating country entering the relevant nationality (e.g., Spanish, German, Chinese) in the slot indicated by the square brackets in their version of the questionnaire.

Socio-demographic questions often require adaptations to be made in different locations and languages (see Translation: Shared Language Harmonization).

The need to make some adaptations might only become apparent in the course of translating the source questionnaire into a given target language. This could be because features of the target language itself make adaptation necessary or because a translated version of the source question, although possible, would not achieve the required measurement goals.

Response scales provide examples of adaptations occasioned by features of the target language. Agreement scale response categories developed in English frequently have a middle category "neither agree nor disagree." In languages such as Hebrew and Swahili, this phrase cannot properly be translated by simply translating the words. The closest semantic option available to translate "disagree" in Hebrew, for example, corresponds to "no agree." In addition, the words "neither" and "nor" are the same as the target language element corresponding to "no." Thus "neither agree nor disagree," if translated element for element, would produce something like "no agree, no no agree" this makes little sense in Hebrew (Harkness, 2003). The Hebrew phrase thus used in ISSP studies for the category "neither agree nor disagree" corresponds to "in the middle." On a study on adapting or translating response scales, see Villar (2009).

Frequently adaptations are motivated less by features of the target *language* than by the need to fit social, cultural, or other needs of the new linguistic group to be studied. Examples of adaptation not directly related to linguistic considerations abound. A recent international project proposed fielding the question, "Can you lift a two-liter bottle of water or soda…," in multiple countries. The source question itself was not developed cross-culturally (see Questionnaire Design). Several locations (countries) noted that (a) the normal size of bottle in their context was 1.5 liter bottles, not 2 liter bottles, (b) that they were unsure whether the bottle referred to was intended to be glass or plastic (which would affect the lifting task), (c) that "soda" was not a salient generic concept in their locations, and (d) that the formulation in English which indicates that the bottle is not empty ("bottle of water or soda") needed to become "a full bottle of water" or "a bottle full of water" in their translations. However, there was some concern that these much more explicit renderings of "bottle of water" might alter respondent perceptions of the lifting task.

Usually, as reflected also in these examples, the needs of translation and those of adaptation are entangled. Thus, the appropriate or viable translation for a given context may also be a translation that includes adaptation of content, format, or some other questionnaire feature. For example, translations of an American question referring to being able to walk "several blocks" also needed to adapt the phrase "several blocks" for Great Britain and provide the distance for European locations in terms of yards or meters (Harkness, 2008b). It is not always possible, therefore, to distinguish neatly between translation needs and the need to adapt other features of the question or questionnaire. Therefore, in several studies, both terms can be used in combination as “Translation and Adaptation.” It is also thus essential that translation team and adaptation team work closely together for questionnaire development.

There can be a delicate balance between when adaption is needed and when the changes required are so great that they indicate that the original question should be discarded because even the best adaption cannot result in a question that could be considered equivalent. It is also important to note that the scale of the adaptation work involved is likely to be very different when there are only two languages involved, and useful back and forth can optimize an item and its best translation or adaption, versus when dozens of languages are involved.

Common forms of adaptation

The categories identified below are based on distinctions found in Harkness, (2010b), Harkness (2008b), Behr & Shishido (2016) and Harkness, Villar, & Edwards (2010a).

System-driven adaptation

Units of measurement differ across countries and may require adaptation (e.g., Imperial [yards, pounds] versus Metric [meters, kilos]; Fahrenheit versus Celsius). Direct conversions may be exact and completely equivalent but can produce an odd sounding question. For example, asking about 100 yards would mean asking about 91.4 meters, which while precisely equivalent is an odd metric distance. Adaptations will need to be considered for any mention in instruments of length, area, dry volume, liquid capacity, weight or mass, and also currency.

Adaptation may also be needed to account for structural differences in government, government policies, and laws. For example, a question involving the head of state would ask about the prime minister in the United Kingdom and the President in the United States. Perhaps less straightforward is how to how to adapt questions about a law or policy that may exist in some contexts but not others, such as a minimum wage.

Questions involving currency can also raise adaptation challenges. There are different exchange rates (e.g. official vs. informal) and straight conversion is often not meaningful because of differences in purchasing power. Some economic questions try to achieve equivalence by using some standard such as the median wage as the reference point in each country.

Adaptation to improve or guide comprehension

In preparing to use the question, "Can you run 100 yards?" in Vietnam, local researchers worried that the distance would not be clear to Vietnamese respondents and adapted it to, "Can you run 100 yards or the distance of three light poles?" to help respondents envision the distance intended (Hanh et al., 2005; Harkness, 2008b). In this particular example, the distance mentioned in the source version is retained but also supplemented by a localized indication of the intended distance.

Adaptation to improve conceptual coverage

Sometimes question components are added for a given location to better tap the intended dimension or construct. For example, the symptoms shown by patients with a given disease (as well as the treatments, the attributed causes, and the places to get help) can differ across cultures. Including mention of local symptoms, as relevant, can improve the accuracy of information collected at the local level and for the combined data set.

Adaptation related to cultural discourse norms

Speech communities differ in the way in which they frame and carry out communication. Depending on the culture and language involved, indicators of politeness or deference may be required in the interview script or the self-completion questionnaire (polite imperatives, acknowledgment of relative status of interviewer and respondent, apologies for asking a question, etc.).

In some contexts, adaptations are made without the scientific community currently acknowledging these as part of questionnaire adaptation needs. For example, Korean is a language with a systematic honorifics system reflecting social status, age, interpersonal relationships between participants in a discourse, and, indeed, much more (Strauss & Eun, 2005). In interviewer-assisted applications, such discourse and etiquette requirements affect what interviewers say, depending on whom they are interviewing. In some diglossic linguistic contexts, the gap between written forms of a language and spoken forms can be quite large. This can mean that interviewers have a written script that conforms to the norms of the written standard of the language but are required**,** in "speaking the script," to conform to spoken norms of the language (see Harkness et al. (2008b) and Paulston & Tucker (2003)).

Adaptation and cultural sensibilities

Cultural sensibilities with regard to a wide range of topics differ from culture to culture. Such differences motivated adaptations for Japan in the Picture Completion section of the Wechsler Intelligence Scale for Children (WISC-III). Instead of a depiction of a person with a body part missing, the Japanese version used an inanimate object with a part of that object missing (Ueno & Nakatani, 2003).

Adapting design components or characteristics

Changes to the technical design of an instrument can be motivated by many factors. The direction languages are read or written in, a population's familiarity with certain visual representations (thermometers, ladders, scales using faces (Kunin, 1955)), and a wide range of culturally anchored conventions related to visual presentation, including color symbolism, representational preferences, and conventions of emphasis, may call for adaptation of components of the source questionnaire (see also Instrument Technical Design).

Adaptation related to lexicon and grammar

The lexicon (a language's vocabulary) and grammar of a language may also make changes in design necessary. An example already discussed is the response category "neither agree nor disagree" which has been rendered in Hebrew International Social Survey Programme questionnaires as "in the middle."

Adaptation to maintain or to reduce level of difficulty

Educational and cognitive ability tests are biased if it is easier for one population to answer correctly or perform a task required than it is for another population of equal ability on that item (van de Vijver & Leung, 1997). A wide range of question types is thus sometimes adapted to maintain the same level of difficulty across different populations. Research in educational and psychological testing discusses such issues (see, for example, Georgas, Weiss, van de Vijver, & Saklofske, 2003 and Hambleton, Merenda, & Spielberger, 2005).

In studies of opinions, behaviors and attitudes, the goal is generally more one of keeping respondent burden low. Adjustments may thus sometimes be made to simplify the vocabulary used in a translation for populations with expected low levels of education or to increase instructions and explanations for those unfamiliar with the procedures of survey research. Response scale presentation is sometimes supplemented for populations unfamiliar with the notions of rating, for example, or for those unfamiliar with conceptualizing the response scale concepts in relation to entities asked about (Bullinger, Kirchberger, & Ware, 1995; Struwig & Roberts, 2006).

Guidelines

Goal: To make a survey instrument better fit the needs of a new population, location, language, or mode.

1. **Determine the policy, people, and procedures for adaptation for the project.**

   Rationale

   Adaptation needs will arise in most comparative projects and should therefore be sufficiently prepared for. Any quality assurance and quality monitoring framework must therefore include a plan for how to deal with adaptation. This plan should propose procedures to identify and address adaptation needs for each location and how to make decisions about documentation. It should also determine how any effort to coordinate adaptations or their documentation is to be organized (see Harkness (2010b)).

Procedural steps

1.1 Plan coordination of adaptation development and the tools to be used to develop and document the process and outputs.

1.2 Identify a suitable team with the necessary skills to work on adaptation problems (see Guideline 2 below).

1.3 Decide on a procedure for approving adaptation by the persons assigned to decide and approve adaptations. In projects aiming to ask the same questions (ASQ) of each population, substantive adaptations should only be made if they are required to ensure comparable measurement or avoid some other important negative consequence.

1.4 Decide on a strategy to ensure that participating groups (locations, countries, etc.) are informed about adaptations being proposed by other members and can contribute their own proposals or reactions.

Lessons learned

1.1 By anticipating certain adaptations in an ASQ source and translate (ASQT) model, the translated versions are likely to be more consistent with the measurement intended in the source questionnaire. However, it is very likely that some adaptation needs will not be recognized until translated versions are available.

2. Recruit a team to work on adaptations.

Rationale

Adaptations are made to address modifications necessary to be able to interview multiple populations. The spread of skills and range of cultural experience required cannot be provided by one person.

The team should bring together knowledge about and an understanding of (1) adaptation needs in general, (2) the types of adaptation, (3) the strategies commonly used to adapt, (4) measurement comparability needs, (5) language proficiency in whatever languages are involved, and (6) relevant cultural information.

The team should work in close cooperation with the translation team. Depending on the project and the team composition, it may be that the same people carry out translation and adaptation tasks. In any case it is important to have people with adaptation knowledge and skills in the overall team for 'transforming' a source survey instrument / source questionnaire for use by target populations.

Procedural steps

2.1 Identify a small group of people who can, as a team, provide the skills and competencies needed for the six points mentioned above.

2.2 Identify at least two people for each given location or instrument to work as an adaptation team. This team should supplement the translation team to carry out adaptations as needed. These additional team members contribute only to the specific instrument they can provide input on. They provide the specific cultural awareness and language competence needed for a given location and language. However, issues identified for one location and population may prove relevant for others too.

2.3 Brief all team members on the goals of the adaptation steps, the procedures, any tools to be used, and the documentation required.

Lessons learned

2.1 Briefing and providing examples of what is desired and not desired is important. Members of such teams might be working consciously on adaptation for the first time. In addition, some team members with experience with adaptation might have learned practices the current team does not want to endorse. Providing examples for discussion during briefing and training reduces the likelihood of members making incorrect assumptions about what is required and how to proceed.

3. Review, as relevant, the source questionnaire for adaptation needs.

Rationale

Identifying and resolving adaptation needs in the source questionnaire may result in a better source questionnaire (that is, one that is easier to work with as a source questionnaire). By identifying and resolving elements to consider for adaptation in the source document, comparability across different questionnaire versions can also be enhanced.

Procedural steps

3.1 Assign the work to a person or persons familiar with the common forms of adaptation in surveys, knowledgeable about the questionnaire as well as the measurement goals of each question, and with a good understanding of the cultural and social realities of both source and target populations. Provide a format for indicating potential adaptation elements.

3.2 Keep a record of all elements identified and the rationale for these.

3.3 Provide examples of what is required in terms of adaptation in the record.

3.4 Check the suggestions made with a range of locations participating in the project. The members engaged for local consultation would be useful contacts for this.

3.5 Adjust the adaptation proposals for the source questionnaire as seems appropriate.

Lessons learned

3.1 It may not be easy to find people with experience in adaptation procedures. People with extensive experience in drafting questionnaires for multicultural projects and translators may be good first choices; each can provide different insights based on their different knowledge and experience.

3.2 The ability to look at a questionnaire with an awareness of other cultures' needs can be trained but it needs to be based on some background of cross-cultural experiences and awareness. Translators develop the ability to think across and between cultures in the course of their training. Their insights and their explication of motivations for suggested changes could help others in the team learn what is needed. At the same time, translators cannot be expected to understand all the measurement factors to be considered in question adaptation. In addition, translators are not necessarily in touch with the on-the-street reality of interviewing and the everyday language of the target population. This is why a team providing a spread of expertise is recommended.

4. Review the translated questionnaire or instrument for adaptation needs.

Rationale

A review with respect to adaptation can be incorporated into the translation phases. Some adaptation proposals are likely to result from the translation process in any case. However, some adaptation needs that are unrelated to translation may not be apparent to the translation team. It is, therefore, important to check for other adaptation needs once the translation is completed. In addition, the adaptation team may have access to knowledge about adaptation undertaken in other languages involved in a multi-lingual project that an individual translation team does not.

Procedural steps

4.1 The adaptation team should collaborate closely with the translation team. The persons chosen should, together, provide language and translation skills and a good understanding of the cultural contexts of target versions. The team producing the local target version of the questionnaire could help them as necessary to be aware of source version implications and cultural assumptions inherent in it. These people need not be extremely proficient in the language of the source questionnaire. If suitable local people are readily available, using two different people from those advising on adaptation for the source questionnaire could minimize repetition and transfer of topics from the source questionnaire review to the current review.

4.2 Provide a format for indicating potential adaptation elements, along with examples.

4.3 Keep a record of all elements identified and the rationale for these.

4.4 Check the suggestions made by the adaptation team with groups formed from other locations and adjust the adaptation proposals accordingly. This step might best be undertaken as a late step in deciding adaptations for the entire project.

Lessons learned

4.1 Given the meager literature on the rationale and procedures of adaptation in surveys (for an example, see Behr & Shishido (2016)), adaptation teams may end up making decisions based on common sense and best guesses. Pretesting adaptation decisions before implementation is thus essential.

5. Document adaptations and the rationale for making them

Rationale

Documentation of adaptations is important for version control across locations and adaptations in one round of a survey. It also makes it possible to check content and presentation through any longitudinal iterations of a survey or a question. Such documentation can also ultimately inform the development of a more refined understanding of adaptation practices.

Lessons learned

5.1 Ensure the documentation of changes and their rationale is made publicly available. At the moment it is not easy to find literature on adaptation that presents procedures and motivations in detail (for an example, see Behr & Shishido (2016)). The documentation taken by teams as proposed above will form an important basis for surveys in the future and help advance this area of methodology.

5.2 The motivation for adaptations may also not be evident to those not involved in the adaptation process. Secondary analysts, for example, would benefit from a record of the rationale behind adaptations.

6. Test adaptations made with the target population.

Rationale

Adaptation results in new questions. New questions should be tested with people representative of the target population.

Procedural steps

6.1 Pretest adapted instruments to find out whether the questions are understood as intended and can be answered without undue burden.

6.2 Include quantitative assessment (see Pretesting).

Lessons learned

6.1 It is important to streamline development of adapted instruments as much as possible in order to have enough time and resources to undertake the various steps and testing of these steps. Adaptation needs should be considered at each stage of development; however, in several surveys, they may be mainly dealt with during the translation stage. Development and pretesting of the source questionnaire should keep adaptation needs in mind. The question about being able to lift a 2-liter bottle of water or soda, for example, could have been evaluated in terms of the availability of bottled beverages, the saliency of the size of the bottles, and the material of which they might be made. So the need to make adaptations from a final source instrument can be reduced during the questionnaire design phase, as translation alone cannot remedy such matters.

6.2 If adaptation is left until the last moment, there may be no more time or resources to pretest.

6.3 If sharing findings and conclusions about adaptation across locations involved in a project is not organized in an efficient and timely fashion, individual locations are not able to benefit from solutions or problems found in other locations.

6.4 Extensive evaluations of various kinds are needed to establish whether adapted or translated questions result in comparable measurement. The health-related quality of life literature on translated instruments, even on just the SF-36 Health Survey, is revealing in this respect. See, for example, Bolton & Tang (2002) and references cited there.

References

Behr, D., & Shishido, K. (2016). The translation of measurement instruments for cross-cultural surveys. In C. Wolf, D. Joye, T. Smith, & Y-C Fu (Eds.): *The SAGE handbook of survey methodology*.

Bolton, P., & Tang, A. M. (2002). An alternative approach to cross-cultural function assessment. *Social Psychiatry and Psychiatric Epidemiology*, *37*(11), 537–543.

Bullinger, M., Kirchberger, I., & Ware, J. (1995). Der deutsche SF-36 Health Survey (Űbersetzung und psychometrische Testung eines krankheitstibegreifenden Instruments zur Erfassung der gesundheitsbezogenen Lebensqualität [The German SF-36 Health Survey: Translation and psychometric testing of a generic instrument for the assessment of health-related quality of life]. *Zeitschrift für Gesundheitswissenschaften*, *3*(1), 21–36.

Georgas, J., Weiss, L. G., van de Vijver, F. J. R., & Saklofske, D. H. (Eds.). (2003). *Culture and children's intelligence: Cross-cultural analysis of the WISC*-III (pp. 215–225). Amsterdam: Academic Press.

Hambleton, R. K., Merenda, P. F., & Spielberger, C. D. (Eds.) (2005). *Adapting educational and psychological tests for cross-cultural assessment*. Mahwah, NJ.: Lawrence Erlbaum Associates.

Hanh, V. T. X., Guillemin, F., Cong, D. D., Parkerson, G. R. Jr., Thu, P. B., Quynh, P. T., & Briancon, S. (2005). Health related quality of life of adolescents in Vietnam: Cross-cultural adaptation and validation of the Adolescent Duke Health Profile. *Journal of Adolescence, 28*(1)*,* 127–146.

Harkness, J. A. (2003). Questionnaire translation. In J. A. Harkness, F. van de Vijver, & P. Ph. Mohler (Eds.), *Cross-cultural survey methods* (pp. 35–56). Hoboken, NJ: John Wiley & Sons.

Harkness, J. A. (2008b). Comparative survey research: Goals and challenges. In E. D. de Leeuw, J. J. Hox, & D. A. Dillman (Eds.), *International handbook of survey methodology* (pp. 56–77)*.* New York/London: Lawrence Erlbaum Associates.

Harkness, J. A. (2010b). *Towards guidelines on survey adaptation*. Paper for presentation at the XVII ISA World Congress of Sociology, Gothenburg, Sweden.

Harkness, J. A., Schoebi, N., Joye, D., Mohler, P., Faass, T., & Behr, D. (2008b). Oral translation in telephone surveys. In J. M. Lepkowski, C. Tucker, J. M. Brick, E. D. de Leeuw, L. Japec, P. J. Lavrakas et al. (Eds.), *Advances in telephone survey* *methodology* (pp. 231–249). Hoboken, NJ: John Wiley & Sons.

Harkness, J. A., Villar, A., & Edwards, B. (2010a). Translation, adaptation, and design. In J. A. Harkness, M. Braun, B. Edwards, T. P. Johnson, L. E. Lyberg, P. Ph. Mohler, et al. (Eds.), *Survey methods in multinational, multicultural and multiregional contexts* (pp. 117-140). Hoboken, NJ: John Wiley & Sons.

Kunin, T. (1955). The construction of a new type of attitude measure. *Personnel Psychology*, 8(1), 65-77.

Paulston, C.B., & Tucker, G.R. (Eds.). (2003). *Sociolinguistics: The essential readings*. Oxford: Blackwell.

Smith, T. W. (2005). The laws of studying societal change. *Survey Research, 36*(2), 1-5. Retrieved April 28, 2016, from http://www.srl.uic.edu/Publist/Newsletter/2005/05v36n2.pdf

Strauss, S., & Eun, J. O. (2005). Indexicality and honorific speech level choice in Korean. *Linguistics*, *43*(6), 611–651.

Struwig, J., & Roberts, B. (2006, December). *Data collection methodology in South Africa*. Paper presented at the Sixth ZUMA Symposium of Cross-Cultural Survey Methodology, Mannheim, Germany.

Ueno, K., & Nakatani, I. (2003). Japan. In J. Georgas, L. G. Weiss, F. J. R. van de Vijver, & D. H. Saklofske (Eds.), *Culture and children's intelligence: Cross-cultural analysis of the WISC-III* (pp. 215–225). Amsterdam: Academic Press.

van de Vijver, F. J. R., & Leung, K. (1997). *Methods and data analysis for cross-cultural research*. Thousand Oaks, CA: Sage Publications.

Villar, A. (2009). Agreement answer scale design for multilingual surveys: Effects of translation-related changes in verbal labels on response styles and response distributions. *Survey Research and Methodology Program (SRAM)-Dissertations & Theses*, 3.

Pretesting

Rachel Caspar, Emilia Peytcheva, Ting Yan, Sunghee Lee, Mingnan Liu, and Mengyao Hu, 2016

Introduction

Pretesting plays an essential role in identifying and potentially reducing measurement error that damages statistical estimates at the population level and thus endangers comparability across populations in multinational, multiregional, and multicultural surveys, which we refer to as "3MC surveys". Pretesting involves a variety of activities designed to evaluate a survey instrument's capacity to collect the desired data, the capabilities of the selected mode of data collection, and the overall adequacy of the field procedures. Throughout this text we refer to a "pretest" as the collection of the qualitative and quantitative techniques and activities that allow researchers to evaluate survey questions and survey procedures before data collection begins. Table 1 provides a summary of the most commonly used pretesting techniques, such as pilot studies, cognitive interviewing employing concurrent or retrospective think aloud techniques, focus groups, behavior coding, and so on.

As suggested in the survey lifecycle, many pretesting activities take place once the questionnaire and other survey materials have been developed, adapted and translated. However, pretesting techniques such as focus groups and vignettes are often used in advance of the overall research and questionnaire design in order to inform question wording and other aspects of the research design (appropriate target population, data collection mode and procedures, etc.).

"Pilot studies," also referred to as "dress rehearsals," or "field tests," encompass pretesting procedures that employ all the procedures and materials involved in data collection (regardless of how small of a scale) before the actual data collection begins. They are typically used to achieve a specific goal or multiple goals – from estimating response rates under a particular recruitment protocol to identifying an optimal design characteristic (e.g., incentive amount) through experimentation. Hambleton, Yu, & Slater (1999) identify the following as reasons for conducting a pilot study:

- check the length of the instrument or interview relative to the culture of interest
- check adaptations of instruments
- check the target population's familiarity with units of measure (e.g., currency, English vs. metric system)
- check the target population's familiarity with constructs and concepts (e.g., proper names, "hamburgers")
- check the target population's familiarity with the instrument layout
- identify the customary answering process in the culture of interest (e.g., checking boxes, circling answers, etc.)

- compare item difficulty statistics (for example, see http://wwwn.cdc.gov/qbank/Home.aspx and http://sqp.upf.edu/)

Researchers often draw on a combination of qualitative and quantitative methods to test draft questionnaires and other study materials. Using qualitative methods for an overall mixed methods instrument design serves as a process of integrated (and often iterative) design and pretest.

This chapter provides examples mainly based on U.S. surveys that sample ethnic minorities and immigrants and are administered in different languages, but attempts to extrapolate experiences and lessons learned to cross-national surveys.

When multiple languages are used in the same survey, pretesting the different language versions is an essential part of ensuring measurement equivalence and cultural (Devins, Beiser, Dion, Pelletier, & Edwards, 1997) and cross-cultural equivalence (Hui & Triandis, 1985) (see Translation: Overview). In addition, it is often difficult to employ the same mode of data collection across countries participating in a cross-national project. It is important to test in advance the suitability of the selected mode for the survey topic and population (see Study Design and Organizational Structure). Pretesting techniques may have limited application in a given context and culture. Research into how pretesting strategies may need to be tailored to suit different populations is only beginning to be undertaken systematically. See Pennell, Cibelli Hibben, Lyberg, Mohler & Worku (2017) for a discussion.

Guidelines

Goal: To ensure that all versions of the survey instrument adequately convey the intended research questions, measure the intended attitudes, values, reported facts and behaviors, and that the collection of data is conducted according to specified study protocols in every country and in every language.

1. **Identify what the pretest should achieve and choose a pretest design that best fits the study goals and each population** (Song, Sandelowski, & Happ, 2010).

    Rationale

    Determining what issues have to be addressed allows for the best use of the various pretesting techniques— whether the researchers want to test all field procedures, or only the survey instrument (or parts of it), or the equivalence of the survey instrument across languages and modes of data collection. Pretesting for a study may combine several complementary pretesting techniques (Oremus, Cosby, & Wolfson, 2005) (see below) and should be done in each country participating in the research. Even if some or all of the questions have been used in other studies, pretesting for the local context is necessary to assess their performance in the mode and question order of the current study, the performance of the translation, and with the target population.

Table 1 summarizes the most commonly used pretesting techniques with a brief description, list of their strengths and weaknesses, and the context in which each is typically used.

Procedural steps

1.1 Using Table 1 as an aid, decide what pretesting technique(s) will best fit the study's purpose.

1.2 Consider the cultures within which the study will be conducted and, where possible, establish standardized pretesting protocols across countries regarding:
  1.2.1 How to best convey the objective of the task.
  1.2.2 How to standardize or harmonize the pretesting protocol.
  1.2.3 How to select staff members for the pretest.
  1.2.4 How to train staff.
  1.2.5 How to monitor quality. Audio and video recordings are often made during cognitive interviews and focus groups to help with the reporting process. However, such recordings can also be used to monitor interviewers and focus group moderators to ensure adherence to the pretesting protocol guide. Computer-Assisted Recorded Interviewing (CARI) allows for monitoring during field pretest and field data collection to detect interviewer fraud and ensure data quality (Smith, 2009; Smith & Sokolowski, 2011). For a larger discussion on the importance of quality control and how to incorporate it at various survey stages see Survey Quality and Paradata and Other Auxiliary Data.
  1.2.6 How to analyze results of the pretest (e.g., whether the analysis will be qualitative and/or quantitative).
  1.2.7 How to report and address problems.
  1.2.8 How to decide on changes to the survey instrument.

1.3 After selecting a pretesting technique,
  1.3.1 Assess whether to conduct the pretest(s) in-house or to contract the testing to an outside organization.
  1.3.2 Establish a time schedule that adequately matches the pretesting design, allowing sufficient time to implement any revisions which may be deemed necessary based on results from the pretest prior to implementing the full study.

1.3.3 Budget accordingly. Be sure to include expenses related to interviewer and staff training, respondent recruitment, and incentives, if applicable, for pretest subjects.
1.3.4 Plan how to document the procedures and findings and how to best share them with teams in other countries.

Lessons learned

1.1 In 2012, the German Data Forum established an expert group to provide minimal requirements for assessing and documenting the measurement quality of established and newly developed survey instruments. Six quality standards were derived for each stage of the measurement process. Rammstedt (2014) presented these quality standards for survey instruments and contrasted them with existing alternative standards from other countries and/or disciplines. (Rammstedt, 2014).

1.2 Available pretesting techniques may vary across countries, depending on testing traditions, resources, trained staff, and respondents' familiarity and experience with the pretesting techniques. Even when the same pretesting technique is used, if its implementation varies drastically across countries, it becomes impossible to determine whether observed differences are due to differences in the response process, translation, or the conceptual spectrum. For example, it is not safe to assume that procedures for conducting cognitive interviews will be the same across all countries. Differences may exist in the experience of the interviewers, the location of the interviewing, methods used to recruit participants, approaches to creating the interviewing protocol, and respondents' experience with cognitive interviews. Recent work in seven countries (eight languages) has focused on creating a common approach to cognitive interviewing for questions designed to measure health status (Miller et al., 2008). To ensure equivalence, all parties involved in the project agreed upon the method to be used for recruiting participants, administering the protocol, and documenting results.

1.3 Even when standardized protocols are used across countries, pretesting techniques such as cognitive interviews do not always work equally well across cultural groups without modification (Goerman, 2006; Pan, Craig, & Scollon, 2005). Pan (2004) investigated the efficacy of concurrent think aloud as a pretesting strategy with Chinese respondents. Her investigation identifies challenges and limitations of taking methods developed in one language and culture and directly applying them to another. She points to the need to include consideration of sociolinguistic conventions appropriate to different cultural groups when conducting cognitive interviews because cognitive processes in survey interviews are influenced by cultural background encompassing language. Some recent studies have examined ways of improving the cognitive interviewing experience for Spanish-speaking respondents in the United States (Goerman and King, 2014) and respondents outside of the United States and Europe (Kelley, Cibelli Hibben, Pennell, and Yan, 2015).

1.4 Culture or language specific probes may be needed to test the translation/adaptation of a survey instrument. The Census Bureau conducted cognitive tests of the translations of introductory letters and informational brochures for the American Community Survey in seven languages (Pan, Landreth, Park, Hinsdale-Shouse, & Schoua-Glusberg, 2010). The focus of the study was to examine how cognitive interviews work in non-English languages given cultural differences in communication. Remarkable differences in the way participants from different language groups provided responses were reported. Chinese and Korean respondents tended to provide limited responses and their answers were not focused on the topic; Russian respondents showed a tendency to always give ‘confident’ answers; Spanish and Chinese respondents tended to repeat questions verbatim when asked to paraphrase them (Coronado & Earle, 2002). Such differences in response patterns raise questions related to data quality and the comparability of cognitive interview results across language groups.

1.5 In addition to standard pretesting methods, which focus on question wording and format, ethnographic pretesting techniques may be used to identify shared cultural characteristics. Ethnographic techniques emphasize cultural variables, such as belief systems and everyday practices, which determine whether or not a question makes sense within the culture (Willis, 2005).

1.5.1 Consensus panels are similar to focus groups but are more structured and limit discussion among participants. A panel of people is selected for their expertise and other characteristics deemed to be relevant. They are invited to answer one or more questions about which there may be considerable doubt or disagreement in order to see if a consensual view can be reached.

1.5.2 Questerviews are standardized self-completed questionnaires administered within the context of an in-depth qualitative interview (Oremus, Cosby, & Wolfson, 2005). Respondents are encouraged to discuss their definitions of terms and responses to items while they complete the standardized questionnaire. Usually, questerviews are tape-recorded and transcribed for analysis to identify emergent themes.

1.5.3 Ethnographic pretest interviews ask broader questions than cognitive interviews (Gerber, 1999; Willis, 2005) and may be used to find additional terms regarding a domain of interest and to identify cultural schemas. They are unstructured, nondirective interviews that focus on understanding the interviewed individual's cultural background so that the questions are appropriate to that individual's life (Willis, 2005). Gerber recommends asking ethnographic questions after completing the regular cognitive interview. Willis (2005) offers the following examples of probes which may be used to study various cultural groups:

- "Tell me about the types of activities you do that take physical effort or that make you feel physically tired."
- "The question has a list of foods in it. Are these the types of foods that your family usually eats?"
- "What types of things do you think of as 'work'?"
- "Are you always paid in cash for the work you do, or are there other ways in which you get paid?"

1.6 A related practical question is whether to create cognitive protocols in English and then translate into the target languages, or to develop the protocols directly into the target languages, accounting for different cultural norms and socialization styles. Each approach has benefits and weaknesses that must be weighed against one another given the specific survey conditions (e.g., simultaneous development of the protocol guides may not be as feasible in multilingual projects as it is in bilingual studies) (Pan, 2008; Goerman, 2006; Pan. Craig, & Scollon, 2005; Pan et al., 2010; Lanham, 1974; Scollon & Scollon, 2001). Goerman and Caspar (2010) discuss approaches for creating protocol guides in multiple languages that ensure culture and language appropriateness and present strategies for respondent recruitment, interviewer selection and training that allow adequate testing of instrument translation.

1.7 While focus groups are a quick way to gain in-depth insight into participant knowledge and attitudes, Helitzer-Allen, Makhambera, and Wangel (1994) argue that studies, particularly in the health field, are relying too heavily on this technique. While previous research has shown that focus groups are generally useful in collecting information of a sensitive nature, some topics are exceptions. In a case study in Malawi, adolescent girls were interviewed using two different methods: in-depth interviews and focus group discussions. The study, conducted through the National AIDS Control Programme, utilized mixed methods through quantitative data collection of census information and highly-structured questionnaires as well as qualitative observation, less-structured interviews, and focus groups. Overall, the study found that studies cannot solely rely on focus groups because some topics are so sensitive that individuals will not discuss them in front of one another. For the female subjects in Malawi, menstruation was too sensitive to discuss in focus groups. The authors recommend that researchers use both methods, with in-depth interviews conducted before focus groups. They found that by asking females sensitive questions during their in-depth interviews, they were then able to follow up some of the interview questions by asking if the subject would be willing to discuss this topic in groups of girls.

2. **Combine pretesting techniques to create a comprehensive design plan that takes advantage of the strengths and minimizes the weaknesses of each method.**

Rationale

Pretesting techniques often complement one another and can logically be combined to maximize the efficiency of the pretest design (see Table 1). For example, to minimize cost, one can consider pretesting a questionnaire using expert review. Once the questionnaire is revised based on reviewers' comments, participants for cognitive interviews can be recruited, or a pilot study can be launched. On the other hand, studies comparing multiple pretesting methods have found that different pretesting methods produced different and sometimes even contradictory results regarding the performance of survey questions (Fowler & Roman, 1992; Presser & Blair, 1994; Willis & Lessler, 1999; Rothgeb, Willis, & Forsyth, 2001; Forsyth, Rothgeb, & Willis, 2004; DeMaio & Landreth, 2004; Jansen & Hak, 2005; Beatty & Willis, 2007; Yan, Kreuter, & Tourangeau, 2012). Therefore, it is of great importance that techniques are selected with sufficient consideration of each candidate method's strengths and weaknesses.

In addition, it is important to take language, cultural norms and traditions, as well as interviewer characteristics (see Data Collection: General Considerations and Interviewer Recruitment, Selection, and Training), into account when choosing pretesting methods. The most appropriate combinations of pretesting techniques may vary across countries involved in the study. This should be taken into account when results from the different pretests are evaluated and compared.

Procedural steps

2.1 Begin with pretesting methods that focus on specific aspects of the study (for example, wording of particular questionnaire items, comprehensibility of the informed consent, procedures for interviewers to follow in administering the survey) before moving to techniques that pull all aspects of the project into a more comprehensive study.

2.1.1 For example, consider a focus group or in-depth interviews for initial development of constructs, cognitive interviews for questionnaire development and refinement, and a field pilot study for an overall test of the survey instrument and field procedures. Often, a pilot study with a robust sample can be the best way to test the survey instrument as data analyses with sufficient power can be the most effective way to ascertain if the questionnaire is working as intended.

2.2 Discuss every round of changes introduced to the questionnaire with the coordinating center and test again—consider several iterations of testing, rather than one large scale pretest.

2.3 Be prepared to do multiple rounds of pretesting.

Lessons learned

2.1 In preparation for the shift from a paper-and-pencil instrument to a computer-assisted instrument incorporating a large audio computer-assisted self-interview (A-CASI) component, the U.S. Substance Abuse and Mental Health Services Administration (SAMHSA) implemented a comprehensive pretesting plan (Gfoerer, Eyerman, & Chromy, 2002). The overarching goal of the pretesting was to develop an optimal computerized instrument on the sensitive topic of drug usage. It was also essential that any differences in reporting due to the mode change to A-CASI be identified so that data users would understand how to interpret trend lines from the data. Pretesting work first concentrated on small-scale cognitive laboratory testing to determine the best way to structure the instrument, to train respondents to use the computer for the A-CASI components, to determine the voice to be used for the audio component, and to assess respondents' ability to enter different types of data into the computer (e.g., open-ended responses). Based on results from these laboratory studies, a pilot study was conducted to evaluate interviewer training materials and to collect sufficient data to determine how the mode change impacted reporting. After changes were made based on this field pilot study, a larger pilot study, incorporating an experimental design, was conducted. Finally, the revised instrument and procedures were implemented in a split-sample comparison with the original paper-and-pencil instrument during data collection to allow researchers to assess the impact on the trend lines.

2.2 The General Social Survey (GSS) does a “full pretest,” which tests all new items in a realistic field situation with representative respondents, between cognitive pretesting and a pilot study.

3. Train or hire staff members who are able to adequately implement the chosen pretesting technique(s).

Rationale

The selected pretesting procedures may require skills not possessed by the available interviewers. For example, cognitive interviewing requires a discursive interviewing style which is different from traditional standardized interviewing and requires additional training. Sufficient time and effort should be allowed to train staff members and develop protocols that correspond to the selected pretest design.

Procedural steps

3.1 Select staff members who are fluent in the language of the pretest and sensitive to cultural and linguistic nuances. If different pretest designs are employed in different countries, select interviewers, training, and protocol that match the chosen technique; when the same techniques are used in various countries, harmonize all procedures.

3.2 Train staff members for the pretest.

3.3 Consider interviewer characteristics as they may affect the outcome of a pretest in some cultures more than others (e.g., conversational styles in many cultures are largely determined by the education, gender, or status of the actors in the social hierarchy).

3.4 Monitor interviewer behavior to ensure data quality.

Lessons learned

3.1 Ample time is needed to train local interviewers who may have little or no experience with cognitive interviewing. In the World Health Organization Model Disability Survey, five half-days of training were scheduled to train local Nepali interviewers on how to conduct cognitive interviews. However, early on in the training, it became apparent that even though the interviewers were experienced in standardized interviewing, cognitive interviewing was a new concept. The interviewers had difficulty shifting from standardized interviewing to the protocol of probing the respondent for think-a-loud answers. A training day was added to the agenda to give the interviewers extra practice on the probing protocol. The interviewers also had difficulty understanding that getting the respondent to give a codable response was less important than knowing what the respondent was thinking when formulating their answer. This became apparent after several cognitive interviews were completed. During the daily debriefing an interviewer revealed that a respondent was having difficulty giving a codable answer and she probed until she received a codable answer, but failed to probe what the respondent was thinking.

4. Conduct the pretest in the same mode of data collection (interviewer administered or self-administered) as the main survey.

Rationale

Whatever the eventual mode of data collection, the early stages of research design—testing the construct itself—typically uses face-to-face, laboratory, methods such as focus groups, cognitive interviews, or vignettes. (See Gerber (1999) for a discussion of developing an instrument prior to testing that instrument.)

Once a draft questionnaire has been developed, however, it should be tested in the same mode of data collection as the final survey. There are several significant differences between interviewer- and self-administered surveys. Respondents listen to the questions in interviewer-administered surveys; they read the questions in self-administered surveys. Interviewer-administered surveys involve social interaction between the interviewer and the respondent; self-administered surveys do not. In interviewer-administered surveys, the interviewer handles routing through the questionnaire; self-administered surveys require the respondent to navigate through the questionnaire. Interviewer-administered and self-administered questionnaires also produce different context effects (e.g., recency and primacy) and may also result in differences in socially desirable responding (see Study Design and Organizational Structure and Data Collection: Face-to-Face Surveys). In order to determine how well proposed procedures will work in the field, pretesting should be conducted in the same mode as the final survey.

Procedural steps

4.1 If different modes of data collection are going to be employed across countries, pretest in the respective modes.

4.2 Some pretest techniques are not portable across modes (for example, behavior coding); others require modification. Adapt pretesting techniques to better match the mode of survey data collection (e.g., Redline, Smiley, DeMaio, & Dillman, 1999).

4.3 Use the latest version of the instrument and the respective materials (e.g., show cards, event history calendars).
  4.3.1 Use version control to manage revisions to documents and other materials.

4.4 Use field administration procedures planned for production data collection.

Lessons learned

4.1 Since each mode of data collection has its specific characteristics, it is important to pretest the survey instrument and procedures in every mode that will be used, whether or not the survey questionnaire is translated to a different language. In fact, a change in mode may necessitate changes in wording or changes in design in order to achieve measurement equivalence. For example, cognitive testing for the 2001 U.S. Census showed that more redundancy was needed in the instructions to the "respondent race" question for the respondents to be able to follow the "select one-or-more" option in telephone administration (Davis & DeMaio, 1993). A slightly reworded version of the instructions and question stem resulted in better understanding of the intent of the question over the phone compared to what was needed when asking the question as it appeared in the mail questionnaire (Martin & Gerber, 2004).

5. Conduct the pretest with the same target population as the target population for the survey.

Rationale

To most effectively pretest the survey instrument or field procedures, pretest respondents from the intended target population or, if appropriate, a sub-group within the target population (Willis, 2005). Ideally, the natural flow of the survey instrument should be tested for each culture and language to avoid awkward conversational situations, question order with unpredictable culture-dependent context effects, question repetition not intended in the source, or other culture-specific problems. The population of a pilot study should be an adequate reflection of the survey target population. For example, if the survey design involves oversampling of certain ethnic groups, the pretest sample should also include reasonable representation of these groups. A pretest with sample persons from the target population will most accurately reflect what will happen during actual data collection in terms of cooperation, respondent performance, total interview length, questionnaire performance, survey costs, etc.

Procedural steps

For all pretesting techniques:

5.1 Tailor subject or respondent recruitment to the population of interest.

5.2 Prepare all necessary materials that would be used in the main survey, including an informed consent form that reflects the goals and risks of the pretest study (which may be different from the main survey).

5.3 Select a sample size that is suitable for the chosen pretesting method.

5.4 Apply quotas or use a random sample of the target population to control the demographic make-up of the sample.

5.5 Monitor pretest participant recruitment to ensure best use of the chosen pretesting method.

For pilot studies:

5.6 Select a sample large enough to provide sufficient statistical power to answer the research questions identified in your pilot study analysis plan. Allow for nonresponse, noneligibility, etc.

5.7 Follow the sample selection protocol planned for the final study.

5.8 Monitor the sample selection

Lessons learned

5.1 Select respondents from the survey target population; however, keep in mind that sometimes “survey-trained” respondents may be needed to detect potential problems. A study on pretesting by Hunt, Sparkman, & Wilcox (1982) demonstrated that the general population may not be a good judge of the quality of survey questions, even when this is the target population. The researchers introduced obvious errors in the short questionnaire (e.g., missing response alternatives, inappropriate vocabulary) and asked respondents to be critical of the questions while answering them. Only a third of the sample noticed a missing response alternative; almost no one commented on “double-barreled” questions and “loaded” words. One possible explanation is that all of the respondents had roughly the same low level of survey experience.

5.2 Work conducted by the U.S. Census Bureau to develop a bilingual (English/Spanish) decennial census form has involved cognitive testing to identify potential problems with the layout of the form, to test respondents' ability to correctly navigate through the form, and to assess the quality of the Spanish translation (Goerman, Caspar, Sha, McAvinchey, & Quiroz, 2007). Testing did not directly assess the English questions, as the wording of the English items had already been nearly finalized. As part of one particular study, cognitive interviews were conducted with monolingual Spanish speakers and bilingual Spanish-dominant speakers to focus on translation issues. Results from the testing indicated specific questions that were problematic for Spanish speakers. However, because there was no comparable group of English speakers included in the testing, it was difficult to determine whether the problems were confined to the translated items or would also be problematic for respondents who read the English wordings. To eliminate this problem, in a second round of testing, monolingual English respondents were included as well. The inclusion of these respondents allowed the researchers to identify where problems with the Spanish translation was due to specific choices made in the translation and where concepts were unclear for the Hispanic respondents as opposed to questions that were equally unclear for both English and Spanish speakers.

5.3 Large established cross-cultural studies vary in the type and amount of pretesting they do.

5.3.1 Prior to the start of Round 1, the European Social Survey (ESS) source questionnaire was pretested using "interaction analysis" (i.e., behavior coding) to identify questions which were problematic for the interviewer or respondent. Problem questions were modified and the questionnaire was translated into various languages. In accordance with ESS Round 5 specifications, each participating country was required to pretest its translated questionnaire on a quota controlled, demographically balanced sample of around 50 people. The aims of pretesting were, at a minimum, to check routing and comprehension. Ideally the pretests could also be used to check for equivalence between the translated version of the questionnaire and the source. Countries were encouraged to audio record interviews, conduct respondent and/or interviewer debriefings, and use cognitive interviewing to test for equivalence. The specifications note that these pretests occurred after the source questionnaire had been finalized and that opportunities to amend the source questionnaire were extremely limited at this point (Dorer, 2014).

5.3.2 The Survey of Health, Ageing and Retirement in Europe (SHARE) utilized a four-stage questionnaire development process. In the first stage, working groups produced an English-language draft questionnaire which drew from preexisting survey instruments. The draft questionnaire was piloted in the UK in September, 2002. Based on the lessons from this pilot, the English-language questionnaire was revised and translated into all of the SHARE languages. In the second stage, the translated questionnaires were simultaneously piloted in all SHARE countries, each testing a quota sample of 75 persons. In the third stage, after further revisions to the survey instrument, the full questionnaire was tested in all countries using probability samples (some 100 primary respondents per country plus their spouses). This all-country pretest also tested the country-specific logistics and the procedures to achieve probability samples. During the fourth stage, pilot and pretest results were statistically analyzed, leading to the final design of the questionnaire (Borsch-Supan, n.d).

6. Evaluate the results of the pretest.

Rationale

The goal of the pretest is to identify problems in the questionnaire and study design in each country. The results of the pretest have to be evaluated to determine the best way to fix existing problems without introducing new ones. Changes to the survey instrument and design should be considered in the context of the whole study -- changes that fix a problem in one country may introduce a problem in another. The coordinating center should decide whether minor differences that still preserve the measurement equivalence of the survey instrument across countries can be tolerated (see Translation: Overview and Study Design and Organizational Structure). Any introduced changes in instrument design should also be pretested to avoid unforeseen errors (also see Instrument Technical Design).

Procedural steps

6.1 Examine the findings of each pretesting technique used and identify the causes of the any problems discovered.

6.1.1 Decide in advance what constitutes a problem. For example, the 10%+ rule is often used in behavior coding to flag questions: if a question is misread or misunderstood by over

10% of respondents, then it is considered problematic. The appropriate threshold for any particular study is often determined from the distribution of coded errors (which is dependent on the coding scheme and instructions for code assignments).

6.1.2 Look for problems that are common across interviews, but also be aware that a problem may be important even if it occurred in only one interview. This is especially important when qualitative techniques are used – in order to determine what constitutes a problem, all possible factors that play a role in the pretest should be considered.

6.1.3 Examine in what situations and with what types of respondents problems occur.

6.2 If a pilot study has been conducted:

6.2.1 Review response distributions and item nonresponse for key study variables.

6.2.2 Review interview length.

6.2.3 Review satisficing behaviors.

6.2.4 For attitudinal and value variables, check whether items group together as intended in the survey (e.g., perform confirmatory factor analysis, latent class analysis (Yan, Kreuter, & Tourangeau, 2012), analysis of variance (Van de Vijver & Leung, 1997).

6.2.5 Solicit and review feedback from interviewers and respondents.

6.3 Report the results and proposed changes to the coordinating center. It is important that the timing and documentation of the pretest are coordinated across participating countries to allow overall comparison of results and propose meaningful changes.

6.4 If changes are introduced to the questionnaire or design procedures, plan for another pretest.

Lessons learned

6.1 Pretesting techniques and the results they yield are meaningful only when the selected procedures are culturally appropriate. Not many pretesting techniques have been tested and studied across countries; thus, some may not be successfully implemented and lead to meaningless results in certain cultures.

6.1.1 Studies in psycholinguistics, for example, have demonstrated different cognitive tendencies between Chinese and English speakers in counterfactual reasoning (Bloom, 1981). When asked what their thoughts would have been on a hypothetical legislation by their government, Hong Kong respondents consistently responded that the government has not proposed such legislation. Chinese speakers were less attuned to hypothetical thinking because their language does not mark counterfactuals differently from conditional statements. Such examples suggest that certain cognitive laboratory methods (for example, vignettes) may be of limited use in some cultures. On the other hand, Gerber (1999) suggests that vignettes may help assess "the cultural sensitivity of a questionnaire."

6.1.2 There are certain error sources that are unique to cross-national questionnaires, or occur less frequently in single nation studies. Tools that help to identify these errors and separate them from measurement errors that only occur in single nation studies assist the cross-national survey researcher in producing a higher quality source questionnaire. In turn, this supports translators in producing functionally equivalent translations that work well in the target languages and cultures. The Cross-National Error Source Typology (CNEST) was developed as a tool for improving the effectiveness of cross-national questionnaire design and has proved useful when applied to categorizing and analyzing the results of cognitive interviews (Fitzgerald, Winstone, & Prestage, 2014).

6.2 The analysis of some pretesting methods can be very labor intensive. For example, transcription is often required for focus groups and cognitive interviews. Analyzing this type of qualitative data requires extensive effort. One simpler approach is to review all interviews, looking for patterns, and then randomly select a few cases for deeper analysis (Pan et al., 2010).

7. Fully document the pretesting protocol and findings.

Rationale

Providing a permanent record of problems encountered during the pretest(s) and any changes made to the questionnaire, respondent materials, and field procedures aids staff and researchers working on similar studies or on later rounds of the same study.

Procedural steps

In a manner consistent across countries, document:

7.1 The pretest sample selection and recruitment method, including the sampling frame and sample size.

7.2 The use of incentives.

7.3 The geographical location of the pretest.

7.4 Respondent characteristics.

7.5 Mode(s) of pretest administration.

7.6 Dates of data collection and organization(s) conducting the interviews.

7.7 Types of staff conducting the pretest (e.g., experienced interviewers, supervisors) and the training they received.

7.8 All materials used in the pretest.

7.9 Pretest findings and their implications.

7.10 Any changes made to the survey instrument and the pretesting source that lead to these changes.

7.11 The number and types of pretests.

Lessons learned

7.1 The documentation can serve as a resource for future studies. For example, researchers within a U.S. Federal Interagency Group have developed Q-BANK (http://wwwn.cdc.gov/qbank/home.aspx), a database of questions for national health surveys maintained by their Questionnaire Design Research Laboratory (QDRL) at the National Center for Health Statistics, Center for Disease Control (CDC). The database catalogues tested questions and links each question to cognitive testing findings. Questions are searchable not only by content or subject matter (e.g., asthma questions, cancer questions, demographics), but also by question type (e.g., objective characteristics, behavioral reports, attitudes), response category type (e.g., yes/no, open-ended, quantity), and response error type (e.g., problems with terms, recall problems). A statistical tool has been developed that performs basic statistical procedures on questions in the database.

Q-BANK, when completed, will centralize cognitive testing reports with links to specific questions and topic areas and will advance the field by: 1) serving as a resource in the development of new questions, 2) allowing question and response error comparisons across studies, 3) performing analysis on the characteristics of questions contributing to specific response errors, and 4) serving as a research tool investigating response error.

Q-BANK is available to any interested researcher. Researchers are also encouraged to contribute their own research reports to the catalogue to strengthen the utility of the site.

Table 1. Pretesting methods, their strengths, and weaknesses. (These can be iterative and can be used in combination)

Approach	Pretesting Method	What it is	Strengths	Weaknesses	Most Common Use
Field Methods	Field pilot study (for an overview, see Groves, et al. (2009))	A miniature version of the main data collection	Realistic; allows for testing all field procedures; allows for feedback from interviewers, field managers, respondents, and data analysts	Costly; requires large sample size relative to the other techniques, needs to be planned and conducted in advance to allow time for changes	Field work test
	Interviewer debriefings (for an overview, see Goerman, et al. (2007))	Small group discussion with interviewers to talk about their experiences	Uses interviewers' expertise on what makes a question difficult in a particular situation and with particular types of respondents	Interviewers themselves may be responsible for the respondents' confusion/problem with a question	Field work test
	Respondent debriefings	Respondents' comments on specific questions or the survey as a whole (usually collected during a field pilot study as a separate interview);	Cheap - conducted as part of the field pilot study; allows for identification of question-specific problems; large sample size allows for confidence in results; realistic (field setting)	In some cultures, respondents may not want to admit confusion and inability to understand a question; increases respondent burden as the length of the interview increases; may be hard to recall items that were problematic	Field work test
	Behavior coding (e.g., Mangione, Fowler, & Oksenberg (1992); also, Groves, et al. (2009))	Systematic coding of the interviewer-respondent interaction in order to identify problems that arise during the question-answer process	Direct observation of the question-answer process; comparability when standard codes are employed; replicable; allows for use of universal codes, but also study specific; quantitative; requires medium sample size (30 interviews are considered sufficient to detect problems)	Time and labor intensive; requires well trained coders and consistent use of the coding scheme; does not identify the exact problem in a question with many codes	Questionnaire testing; field management

Approach	Pretesting Method	What it is	Strengths	Weaknesses	Most Common Use
	Focus groups (see Davis & DeMaio (1993) for an overview; also Groves, et al. (2009))	Small group of people brought together to discuss specific topics in a relatively unstructured manner, led by a moderator who ensures the flow of the conversation is in the intended direction	Useful when there is no information on the topic of interest; uses the same types of respondents who are the target population for the survey; allows for immediate follow up; requires small group size (10-12 participants)	Mainly qualitative; results should be carefully interpreted due to small sample size; requires well trained moderators; small group dynamics may influence the results	Questionnaire development
Cognitive Laboratory Methods (for an overview, see Goerman, et al., 2007))	Vignettes (e.g., Rossi & Anderson (1982))	Brief stories/scenarios describing hypothetical situations or persons and their behaviors to which respondents are asked to react in order to allow the researcher to explore contextual influences on respondent's response formation processes	Allows for quantitative analyses; suitable for sensitive topics; requires small sample size relative to the other techniques	Disconnect between a hypothetical situation and respondent's actual views and behaviors; cultures may differ in their ability to think hypothetically (e.g., Bloom (1981))	Questionnaire development; concept understanding test
	Concurrent think-aloud (see Bickart & Felcher (1996), Davis & DeMaio (1993))	Respondents' report of the thoughts they are having while answering a survey question	Open format with potential for unanticipated information; lack of interviewer bias when probes are not used	Unnatural; high respondent burden; may affect the natural response formation process, thus provide unrealistic picture of how respondents answer questions in the field; coding may be burdensome; assumes respondents are able to identify and report what information they used to come up with a response to the survey question; respondents may begin to overinterpret the questions and come up with problems that do not exist in the natural context	Questionnaire development

Approach	Pretesting Method	What it is	Strengths	Weaknesses	Most Common Use
	Retrospective think-aloud (see Belson (1981))	Interview with respondents after they have completed a survey about how they came up with answers to specific questions	Does not interfere with the response formation process	Assumes respondents are able to identify and report what information they used to come up with a response to the survey question; assumes information is still available in short-term memory	Questionnaire development
Other	Expert review (for an overview, see Groves, et al. (2009))	Review of draft materials by experienced methodologists, analysts, translators	Cost efficient; quick; can identify a wide variety of problems in the survey questionnaire (from typos to skip patterns); requires very small sample of experts (usually 2-3)	Subjective; no "real" respondents involved	Questionnaire development
	Question Appraisal System (for example, Willis & Lessler (1999))	A systematic appraisal of survey questions that allows the user to identify potential problems in the wording or structure of the questions that may lead to difficulties in question administration, miscommunication, or other failings.	Cost efficient; provides sense of reliability due to standardization	Identifies a problem without pointing out to a solution	Questionnaire development
	Usability Testing (see Hansen & Couper (2004), Tarnai & Moore (2004))	Testing of the functionalities of CAPI, CATI, sample management systems or printed materials such as respondent and interviewer booklet, show cards, etc.	Direct user assessment of the tools that will be used during data collection; can be cheap - can be conducted with employees of the survey organization; usually requires small sample sizes	Time consuming	Field work test

Approach	Pretesting Method	What it is	Strengths	Weaknesses	Most Common Use
Statistical Modeling	Multi-trait-multi-method (MTMM) Database (see Saris, van der Veld, & Gallhofer (2004))	Database of MTMM studies that provides estimates of reliability and validity for over 1000 questionnaire items	Provides quantitative measures of question quality	Costly and labor intensive; questions are considered in isolation, so question order effects might be ignored	Questionnaire development
	Item Response Theory (IRT) Approach (see Reeve & Mâsse (2004))	Statistical models that allow examination of ways in which different items discriminate across respondents with the same value on a trait	Provides a quantitative measure of item functioning; suitable for scale development	Requires data collection; questions considered in isolation	Questionnaire development
	Latent Class Analysis (LCA) (see Yan, Kreuter,& Tourangeau (2012), Kreuter, Yan, & Tourangeau (2008))	Statistical models that allow examination of error rates associated with different items	Provides a quantitative measure of error rates; suitable for comparing different candidate items measuring the same underlying construct	Requires data collection; questions considered in isolation;	Questionnaire development

References

Beatty, P. C., & Willis, G. B. (2007). Research Synthesis: The practice of cognitive interviewing. *Public Opinion Quarterly, 71*, 287– 311.

Belson, W. A. (1981). *The design and understanding of survey questions*. Aldershot, UK: Gower.

Bickart, B., & Felcher, M. (1996). Expanding and enhancing the use of verbal protocols in survey research. In N. Schwarz and S. Sudman (Eds.), *Answering questions: Methodology for determining the cognitive and communicative processes in survey research* (pp.115-142). San Francisco, CA: Jossey-Bass.

Bloom, L. (1981). *The linguistic shaping of thought*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Borsch-Supan. *History of the development process: Pilots, pre-tests, and main study*. Retrieved from http://www.share-project.org/t3/share/uploads/tx_sharepublications/CH_7.1.pdf

Coronado, L., & Earle, D. (2002*). Effectiveness of the American Community Survey of the U.S. Census in a borderlands colonial setting*. Research report submitted to the U.S. Census Bureau, March 19, 2002.

Davis, W., & DeMaio, T. (1993). *Comparing the think-aloud interviewing technique with standard interviewing in the redesign of a dietary recall questionnaire*. Alexandria, VA: ASA Section on Survey Research Methods.

DeMaio, T., & Landreth, A. (2004). Do different cognitive interview techniques produce different results? In S. Presser et al. (Eds.), *Methods for testing and evaluating survey questionnaires* (pp. 891-08). Hoboken, NJ: John Wiley & Sons.

Devins, G. M., Beiser, M., Dion, R., Pelletier, L. G., & Edwards, R. G. (1997). Cross-cultural measurements of psychological well-being: The psychometric equivalence of Cantonese, Vietnamese, and Laotian translations of the affect balance scale. *American Journal of Public Health*, *87*(5), 794–799.

Dorer, B. (2014). European Social Survey, Round 5. *Report on changes to existing translations applied by national teams in the ESS Round 5 questionnaire*. Retrieved from http://www.europeansocialsurvey.org/docs/round5/methods/ESS5_changes_to_existing_translations.pdf

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. Lausanne: FORS.

Forsyth, B., Rothgeb, J., & Willis, G. (2004). Does questionnaire pretesting make a difference? An empirical test using a field survey experiment. In S. Presser, et al. (Eds.), *Questionnaire development, evaluation, and testing* (pp. 525-546). Hoboken, NJ: John Wiley & Sons.

Fowler, F. J., & Roman, A. M. (1992). *A study of approaches to survey question evaluation*. Final Report for U.S. Bureau of the Census, Boston: Center for Survey Research.

Gerber, E. R. (1999). The view from anthropology: Ethnography and the cognitive interview. In M. G. Sirken, D. J. Herrmann, S. Schechter, N. Schwarz, J. M. Tanur, and R. Tourangeau (Eds.). *Cognition and survey research* (pp. 217-234). New York, NY: Wiley.

Gfroerer, J., Eyerman, J., & Chromy, J. (Eds.) (2002). Redesigning an ongoing national household survey: Methodological issues (DHHS Publication No. SMA 03-3768). Rockville, MD: Substance Abuse and Mental Health Services Administration, Office of Applied Studies.

Goerman, P. L. (2006). An examination of pretesting methods for multicultural, multilingual surveys: The use of cognitive interviews to test Spanish instruments. In J. A. Harkness (Ed.), *Conducting cross-national and cross-cultural surveys* (ZUMA Nachrichten Spezial Band 12) (pp. 67-79). Mannheim, Germany: ZUMA. Retrieved April 14, 2010, from http://www.gesis.org/fileadmin/upload/forschung/publikationen/zeitschriften/zuma_nachrichten_spezial/znspezial12.pdf

Goerman, P. L., Caspar, R., Sha, M., McAvinchey, G., & Quiroz, R. (2007). *Census bilingual questionnaire research* (Final round 2 report). Retrieved June 3, 2008, from http://www.census.gov/srd/papers/pdf/ssm2007-27.pdf

Goerman, P. L., & Caspar, R. (2010). Managing the cognitive pretesting of multilingual survey instruments: A case study based on the pretesting of the U.S. Census Bureau bilingual Spanish/English questionnaire. In J. A. Harkness, M. Braun, B. Edwards, T. P. Johnson, L. E. Lyberg, P. P. Mohler, B-E. Pennell, T. Smith (Eds.), *Survey methods in multinational, multiregional, and multicultural contexts*. New York, NY: John Wiley & Sons.

Goerman, P. L., & King, R (2014). Adaptation of standard cognitive interview methodology for use with Spanish-speaking respondents. *Proceedings of the American Association for Public Opinion Research Annual Conference*, Anaheim, CA.

Groves, R. M., Fowler, F. J. Jr., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2009). *Survey methodology* (2nd ed.). Hoboken, NJ: John Wiley & Sons.

Hambleton, R. K., Yu, J., & Slater, S. C. (1999). Fieldtest of the ITC guidelines for adapting educational and psychological tests. *European Journal of Psychological Assessment*, *15*(3), 270-276.

Hansen, S. E., & Couper, M. P. (2004). Usability testing as a means of evaluating computer assisted survey Instruments. In S. Presser, J. M. Rothgeb, M. P. Couper, J. T. Lessler, E. Martin, J. Martin, et al. (Eds.), *Methods for testing and evaluating survey questionnaires*. New York, NY: John Wiley & Sons.

Helitzer-Allen, D., Makhambera, M., & Wangel, A-M. (1994). Obtaining sensitive information: The need for more than focus groups. *Reproductive Health Matters*, *3*, 75-82.

Hui, C. H., & Triandis, H. C. (1985). Measurement in cross-cultural psychology: A review and comparison of strategies. *Journal of Cross-Cultural Psychology*, *16*(2), 131–152.

Hunt, S. D., Sparkman, R. D., & Wilcox, J. B. (1982). The pretest in survey research: Issues and preliminary findings. *Journal of Marketing Research*, *19*, 269-273.

Jansen, H., & Hak, T. (2005). The productivity of the three-step test-interview (tsti) compared to an expert review of a self-administered questionnaire on alcohol consumption. *Journal of Official Statistics*, *21*, 103-120.

Kelley, J., Cibelli Hibben, K. L, Pennell, B-E., & Yan, T. (2015). The effect of providing think-aloud examples and practice on cognitive interviewing in Nepal. *Proceedings of the Annual Conference of the European Survey Research Association*, Reykjavik, Iceland.

Kreuter, F., Yan, T., & Tourangeau, R. (2008). Examining the effectiveness of latent class analysis to item evaluation. *Journal of the Royal Statistical Society*, *171*(Series A), 723-738.

Lanham, R. A. (1974). *Style: An anti-textbook*. New Haven and London: Yale University Press.

Mangione, T., Fowler, F. J., & Oksenberg, L. (1992). Question characteristics and interviewer effects. *Journal of Official Statistics*, *8*(3), 293-307.

Martin, E., & Gerber, E. (2004). *Results of recent methodological research on the Hispanic origin and race questions*. Paper presented to the Population Association of America, Boston, MA.

Miller, K., Fitzgerald, R., Caspar, D. M., Gray, M., Nunes, C., Padilla, J. L., et al. (2008). *Design and analysis of cognitive interviews for cross-national testing*. Paper presented at the International Conference on Survey Methods in Multinational, Multiregional and Multicultural Contexts, Berlin, Germany.

Oremus, M., Cosby, J. L., & Wolfson, C. (2005). A hybrid qualitative method for pretesting questionnaires: The example of a questionnaire to caregivers of Alzheimer disease patients. *Research in Nursing and Health*, *28*(5), 419-430.

Pan, Y. (2004). Cognitive interviews in languages other than English: Methodological and research issues. *Proceedings of the Section on*

Survey Research Methods. Alexandria, VA: American Statistical Association.

Pan, Y. (2008), Cross-cultural communication norms and survey interviews. In H. Sun & D. Kädär (Eds.), *It's the dragon's turn—Chinese institutional discourse(s)* (pp. 17-76). Berne: Peter Lang.

Pan, Y., Craig, B., & Scollon, S. (2005). Results from Chinese cognitive interviews on the Census 2000 long form: Language, literacy, and cultural issues. *Statistical Research Division's Research Report Series* (Survey Methodology #2005 - 09). Washington, DC: U.S. 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 J. A. Harkness, M. Braun, B. Edwards, T. P. Johnson, L. E. Lyberg, P. P. Mohler, et al. (Eds.), *Survey methods in multinational, multiregional, and multicultural contexts* (chapter 6). New York, NY: John Wiley & Sons.

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, et al. (Eds.), *Total survey error in practice*. New York, NY: John Wiley & Sons.

Presser, S., & Blair, J. (1994). Survey pretesting: Do different methods produce different results? *Sociological Methodology*, *24*, 73-104.

Rammstedt, B. (2014). *Measurement quality in the social sciences – Standards for quality assessment*. Paper presented at the ESS seminar: Highlights in Survey Research, Universitat Pompeu Fabra, Edifici Roger de Llúria. Abstract retrieved from http://www.upf.edu/survey/Seminar/Beatrice.html

Redline, C. R., Smiley, M. L., DeMaio, T., & Dillman, D. A. (1999). *Beyond concurrent interviews: An evaluation of cognitive interviewing techniques for self-administered questionnaires*. Paper presented at Proceedings of American Statistical Association: Survey Research Methods Section.

Reeve, B., & Mâsse, L. (2004). Item response theory (IRT) modeling for questionnaire evaluation. In S. Presser, J. M. Rothgeb, M. P. Couper, J. T. Lessler, E. Martin, J. Martin, et al. (Eds.), *Methods for testing and evaluating survey questionnaires*. New York, NY: John Wiley & Sons.

Rossi, P., & Anderson, A. (1982). The factorial survey approach: An introduction. In P. Rossi and S. Nock (Eds.), *Measuring social judgments: The factorial survey approach*. Thousand Oaks, CA: Sage Publications.

Rothgeb, J., Willis, G., & Forsyth, B. (2001). Questionnaire pretesting methods: Do different techniques and different organizations produce similar results? *Proceedings of the Section on Survey Methods*. Alexandria, VA: American Statistical Association.

Saris, W., van der Veld, W., & Gallhofer. I. (2004). Development and improvement of questionnaires using predictions of reliability and validity. In S. Presser, J. M. Rothgeb, M. P. Couper, J. T. Lessler, E. Martin, J. Martin, et al. (Eds.), *Methods for testing and evaluating survey questionnaires*. New York, NY: John Wiley & Sons.

Scollon, R., & Scollon, S. W. (2001), *Intercultural communication: A discourse approach* (2nd ed.). Oxford: Basil Blackwell.

Smith, T. (2009). *An analysis of computer assisted recorded interviews (CARI) on the 2008 General Social Survey*. GSS Methodological Report No. 117. Chicago: NORC.

Smith, T., & , & Sokolowski, J. (2011). Using audio-visuals in surveys. In S. Hesse-Biber (Ed.), *The handbook of emergent technologies in social research*. Oxford: Oxford University Press.

Song, M-K., Sandelowski, M., & Happ, M. B. (2010). Current practices and emerging trends in conducting mixed methods intervention studies in the health sciences. In A. Tashakkori and C. Teddlie (Eds.), *SAGE handbook of mixed methods in social and behavioral research* (2nd ed., pp. 725-748). Thousand Oaks, CA: Sage.

Tarnai, J., & Moore, D. (2004). Methods for testing and evaluating computer-assisted questionnaires. In S. Presser, J. M. Rothgeb, M. P. Couper, J. T. Lessler, E. Martin, J. Martin, et al. (Eds.), *Methods for testing and evaluating survey questionnaires*. New York, NY: John Wiley & Sons.

van de Vijver, F. J. R., & Leung, K. (1997). *Methods and data analysis for cross-cultural research*. Newbury Park, CA: Sage.

Willis, G. B. (2005). *Cognitive interviewing: A tool for improving questionnaire design*. Thousand Oaks, CA: Sage.

Willis, G. B., & Lessler, J. T. (1999). *Question appraisal system BRFSS-QAS: A guide for systematically evaluating survey question wording*. Report prepared for CDC/NCCDPHP/Division of Adult and Community Health Behavioral Surveillance Branch. Rockville, MD: Research Triangle Institute. Retrieved April 14, 2010, from http://appliedresearch.cancer.gov/areas/cognitive/qas99.pdf

Yan, T., Kreuter, F., & Tourangeau, R. (2012). Evaluating survey questions: A comparison of methods. *Journal of Official Statistics, 28*(4), 503-529.

Further Reading

Willis, G. B. (2016). Questionnaire pretesting. In C. Wolf, D. Joye, T.W. Smith, & Y. Fu (Eds.), *The SAGE handbook of survey methodology*. Thousand Oaks, CA: Sage.

Interviewer Recruitment, Selection, and Training

Kirsten Alcser, Judi Clemens, Lisa Holland, Heidi Guyer, and Mengyao Hu, 2016

Introduction

Interviewers play a critical role in surveys, as members of the research team who implement the survey design. They are often required to perform multiple tasks with a high level of accuracy. In a face-to-face survey, the interviewer may be required to physically locate the sampled household and to update the sample frame. In both telephone and face-to-face surveys, the interviewer has to contact the household, explain the purpose of the study, enumerate household members, select the respondent, motivate the respondent to participate, ask questions in the required manner, put the respondent at ease, and accurately record the respondent's answers as well as any other required information. Depending upon the survey topic and survey context, the interviewer may be required to perform additional tasks, such as biomeasure collection.

Interviewers can influence responses through their personal attributes and their behaviors, otherwise known as an interviewer effect (or interviewer effects). These guidelines present strategies to optimize interviewer efficiency and minimize the effect that interviewer attributes have on the data through appropriate recruitment, selection, and case assignment; they also present strategies to minimize the effect that interviewer behaviors have on sampling error, nonresponse error, measurement error, and processing error through interviewer training. Note that multinational, multicultural, or multiregional surveys, which we refer to as "3MC" surveys, present a particular challenge as the recruitment, selection and training of interviewers can vary greatly among different countries, due to differences in the cultural environment, existing infrastructure, and resources available (Smith, 2007).

Guidelines

Goal: To improve the overall quality of the survey data by minimizing interviewer effects while controlling costs by optimizing interviewer efficiency.

1. **Determine the structure and composition of the interviewing staff.**

Rationale

The structure and composition of the interviewing staff must be established during the design and planning phases of the project because these decisions will determine the number and type of interviewers required, training protocol, sample assignment, and most efficient methods of supervision. See also Study Design and Organizational

Structure and Tenders, Bids, and Contracts for discussion on decisions about interviewing staff.

Procedural steps

1.1 Consider such parameters as sample size and, for face-to-face studies, geographic distribution; the timing and duration of the data collection period; budget constraints; and the language(s) in which interviewing will occur (Pennell, Harkness, Levenstein, & Quaglia, 2010).

1.2 For face-to-face studies, decide whether interviewers will travel, either individually or in teams with a supervisor, or be locally assigned. See also Data Collection: Face-to-Face Surveys for additional discussion.

1.2.1 Factors favoring the use of traveling interviewers include:

- Lower training costs compared to using local interviewers, as there are fewer interviewers to train and trainers do not have to travel to as many different locations.
- Breach of confidentiality is less of an issue than with local interviewers because interviewers are unlikely to know the respondent personally.
- Respondents may be more willing to participate in sensitive-topic surveys if the interviewers are strangers or “outsiders” (Lee, 1993).

1.2.2 Factors favoring the use of traveling teams rather than traveling individual interviewers include:

- Traveling as a group may be safer than traveling individually.
- Monitoring and supervision are easier since the supervisor is part of the group and is in close daily contact with the interviewers.
- Interviewers have more opportunity to share experiences, learn from one another, and support one another than they would if traveling individually.
- If multiple household members need to be surveyed, different interviewers can speak to them concurrently.
- Similarly, if privacy is difficult to achieve, one interviewer can speak to the respondent while another engages other household members.
- It is easier to implement interpenetrated sample assignments for research purposes than it would be with individual traveling interviewers (Groves et al., 2009a). It is important to note that the cluster design of most area probability sample surveys confound the sampling and non-sampling (i.e., interviewer) variances.

"Interpenetrated sample assignments" are necessary to measure interviewer variance by removing the effects of real differences among respondents assigned to different interviewers. However, interpenetrated workloads are typically not feasible due to the added travel costs and logistics. For this reason, interpenetrated designs are typically only employed for research purposes. However, the use of interviewing teams allows for partial interpenetration, permitting estimation of measurement error introduced by the interviewer. Multi-level modeling is a data analysis technique that makes it possible to estimate interviewer and design effects simultaneously without an (or only a partial) interpenetrated design (O'Muircheartaigh and Campanelli, 1998). See Statistical Analysis for further discussion.

1.2.3 Factors favoring the use of local interviewers include:
- Employing a larger number of interviewers, each with a smaller workload, reduces the interviewer design effect (Kish, 1962; Office of Management and Budget, 2006). See Appendix A for a discussion of the interviewer design effect.
- With a larger field staff, data collection can be completed within a shorter period of time, although the effect is not linear.
- More call attempts can be made per case, since the interviewer remains in the area throughout the data collection period.
- Local interviewer assignment reduces the need for interviewers to travel large distances, thereby reducing travel costs and time expended.
- Local interviewers are familiar with the area and are more likely to share the language and customs of respondents; they may achieve higher response rates than would a stranger or "outsider."

1.3 For telephone studies, decide whether interviewers will conduct the survey from a central telephone facility or from their homes (that is, decentralized telephone interviewing). See also Data Collection: Telephone Surveys for further discussion.

1.3.1 Factors favoring the use of centralized telephone interviewing include:
- Training can be easily centralized.
- Monitoring and supervision can be easier and less expensive, since the supervisor is in close daily contact with the interviewers and may, as a result, have access to more information of relevance such as interviewer

schedules, vocal patterns, and techniques used when addressing different situations.
- It is easier to transfer sample units among interviewers.
- Cost controls are more efficient.

1.3.2 Factors favoring the use of decentralized telephone interviewing include:
- A dedicated telephone facility is not required.
- Interviewer working hours may be more flexible.

1.3.3 Some organizations already have a system in place which mixes centralized and decentralized telephone interviewing.
- In these cases, retaining the combination of centralized and decentralized interviewing may minimize disruption and maintain flexibility.
- Establishing a sample management system that pulls together information from the two into a single report can be a challenge.

1.4 Estimate the Hours Per Interview (HPI). The HPI includes time spent traveling to all sample units, attempting to contact them, documenting contact attempts, and working on project-related administrative duties, as well as conducting the interview with those respondents who agree to participate. The HPI, combined with the hours per week that each interviewer is expected to work on the project and the total number of weeks planned for data collection, helps determine the number of interviewers required (see Appendix B for an example).

1.5 Consider whether any specialized skills or knowledge are required. This may include language skills, the use of special equipment, prior collection of biomeasures, or any physical requirements.

1.51 With a steadily growing interest in the association between social science data and biological data, interviewers are increasingly being called upon to collect biomeasure data such as height, weight, blood spots, saliva samples, and other measures. It is important to consider such tasks in the recruitment stage and to communicate these, or any other tasks that deviate from standard data collection, to prospective interviewers at the time of recruitment.

1.6 Utilizing the results of feasibility assessments (see Data Collection: General Considerations), consider any special requirements of the study, such as:

1.6.1 How many languages are spoken and in what regions?

1.6.2 Would interviewer familiarity with the topic introduce bias or enhance an interviewer's ability to collect data?

1.6.3 Do cultural norms or the nature of the topic necessitate matching interviewers and respondents by gender, dialect, religion, race, ethnicity, caste, age, etc.?

1.6.4 Is physical stamina a consideration (e.g., if interviewers will be required to walk, ride, or bicycle long distances) (Nyandieka, Bowden, Wanjau, & Fox-Rushby, 2002)?

1.6.5 Is the sample widely dispersed, making interviewer access to a car or reliable public transportation a consideration?

1.6.6 Is interviewer safety an issue? For example, should interviewers be advised to travel to the area accompanied, only visit the segment in daylight hours, be prepared to deal with stray dogs, or be coached on dressing for the environment and keeping their equipment, such as a laptop or cell phone, out of view? Interviewer safety must be addressed if interviewers will be required to work in known areas of high crime or active conflict.

Lessons learned

1.1 Many organizations use a combination of interviewer assignment protocols. For example, they may hire local interviewers to make initial contact with sample households, select the respondent, and, if he or she is willing, administer the survey. Later in the data collection period, special traveling interviewers (for instance, experienced interviewers who have proven to be especially skillful at gaining cooperation or relating to particular types of respondents) can be brought in to persuade those selected individuals who have expressed a reluctance to participate. Alternatively, local interviewers might be hired in heavily populated areas while traveling interviewers are sent to more remote regions.

1.2 If traveling teams of interviewers are used, the interviewer may not always be conversant in the respondent's language, and local interpreters may be needed to facilitate data collection. For example, the French Institut National d'Etudes Démographiques has collected data in several Bwa villages in Mali for over 15 years. Although French is the official language of Mali, most villagers speak only Boma, so interpreters were essential for collecting data. The interviewer was responsible for administering the questionnaire, while the interpreter's job was to act as a neutral intermediary between the interviewer and respondent, conveying the words and the concepts associated with them to the two speakers (Quaglia, 2006) (see also Translation: Overview for more information on unwritten translation).

1.3 Matching interviewer and respondent characteristics may improve cooperation but only appears to impact survey data quality if the topic of the survey is related to an identifiable and stable interviewer attribute.

1.3.1 Indonesian researchers felt that matching interviewers with respondents in terms of age, marital status, and child-rearing experience improved rapport and willingness to participate during in-depth interviews (Papanek, 1979).

1.3.2 Several studies indicate that when the topic of the survey (e.g., racial attitudes or women's rights) is related to a fixed interviewer attribute (e.g., race or gender), the interviewer attribute can affect respondents' answers (Davis, 1997; Groves et al., 2009a; Hatchett, & Schuman, 1975; Kane & Macaulay, 1993; Schaeffer, 1980; Schuman & Converse, 1971).

1.3.3 If the topic of the survey is not related to a fixed interviewer attribute, matching the interviewer and respondent on the attribute does not appear to affect data quality. Axinn (1989) found that matching Nepalese interviewers and respondents by gender and ethnicity for a health survey did not decrease the number of technical errors and "don't know" responses or reduce incorrect information gathered during the interview.

1.3.4 Recent research shows that interviewers' religious appearance can affect responses to religion-related questions. For example, studies have found that interviewers wearing Islamic symbols received higher reports of religiosity from respondents in Turkey (Koker, 2009) and that reported religiosity was affected by the interplay between interviewer religious appearance and respondents' characteristics (Blaydes & Gillum, 2013; Benstead, 2014), and the interviewers' own religious, cultural and political attitudes (Mneimneh, de Jong, Cibelli Hibben, & Moaddel, 2015).

1.3.5 Attempting to match interviewer and respondent characteristics may strain the project's resources, particularly if this is not an established practice in the locale. Gender matching is essential in some Muslim countries, but unexpected challenges may arise even when a project has planned and budgeted for such matching (Pennell et al., 2010). The Saudi National Health and Stress survey (SNHS) in the Kingdom of Saudi Arabia has found, for instance, that male interviewers must make the initial contact with a household and seek permission to interview both an eligible male and a eligible female household member. Once cooperation has been secured, a female interviewer must arrange to visit the household to interview the female respondent. In this context, cultural norms also preclude the recording of a female voice, limiting the use of this method of quality monitoring. More complexity is introduced as female interviewers may not travel without a male family member. Supervisors are randomly assigned to observe interviews but must also be gender matched (Mneimneh et al., 2015).

1.4 Using non-clinical interviewers for the collection of biomeasures can be an efficient cost-savings measure.

1.4.1 In a pilot study for the UK Household Longitudinal Study (UKHLS), Understanding Society, researchers successfully used non-clinical interviewers for collection of biomeasures including saliva, a finger-prick, blood pressure and body mass. The research team experienced one barrier in that the interviewers were required to have certain immunizations themselves in order to collect the blood samples. The immunization series took a number of weeks to complete, and this had a negative impact on the survey organization's ability to recruit (McFall, Conolly, & Burton, 2012).

1.4.2 The Health and Retirement Study (HRS), the Survey of Health, Ageing and Retirement in Europe (SHARE) and European Longitudinal Study of Aging (ELSA) were designed to collect similar biomeasures for comparison purposes. Two of the studies, HRS and SHARE, trained interviewers to collect the measurements during the interview; ELSA employed nurses to collect the measures in a separate visit to a health center. Response rates and the distribution of measurements varied depending upon whether trained interviewers or nurses collected the measures. For example, walking speed and grip strength were more variable when measured by interviewers. While response rates in general were lower when respondents had to schedule a separate visit outside of the home, response rates for the more invasive measures were higher for nurses—probably because respondents had more confidence in the medical training of the data collector. However, cost considerations must also be taken into account when considering nurses versus interviewers in collecting such measures (Guyer, Ofstedal, Lessof, Cox, & Juerges, 2010).

2. Determine the pay structure for the data collection staff.

Rationale

Since data collection staff quality has a major impact on the quality of the data collected, it is important to attract and retain the most qualified interviewers possible and to structure compensation accordingly.

Procedural steps

2.1 Interviewer pay structures vary greatly across countries in a 3MC survey. Depending on local labor laws, set interviewer pay comparable to the pay for other jobs requiring similar skills, ideally adjusted for regional cost of living standards.

2.2 Keep in mind local research traditions, the mode of the survey, and local labor laws. The two standard policies are to pay interviewers an hourly rate or to pay per completed interview (European Social Survey [ESS], 2004; Pennell et al., 2010).

2.2.1 Factors favoring payment per interview:
- It is most feasible if each completed interview takes approximately the same amount of interviewer effort, as is more likely in a telephone survey (Pennell et al., 2010).
- It is easier to monitor and control interviewer costs than when paying by the hour (Pennell et al., 2010; Sudman, 1966).

2.2.2 Factors favoring an hourly rate:
- It is most feasible if the effort to complete an interview varies widely, as is common in face-to-face surveys (Lavrakas, 1993; Pennell et al., 2010).
- Interviewers have less incentive to perform hurried, sloppy work or even to fabricate interviews when paid hourly than when paid per interview (Pennell et al., 2010; Sudman, 1966).
- Interviewers are less likely to focus on easy cases while neglecting those who are hard to reach or hard to persuade to participate than when paid by the completed interview (ESS, 2004; Pennell et al.,2010).
- Interviewers may be more willing to spend time on other important tasks (e.g., completing a thorough screening interview and entering comprehensive, accurate contact attempt records) than when paid by the completed interview.

2.3 When determining pay, consider the length and complexity of the interview, the expected difficulties of obtaining cooperation, and the amount of record-keeping demanded of the interviewer (ESS, 2004).

2.4 Pay interviewers for time spent in both initial interviewer training as well as any necessary refresher training.

2.5 Adjust the pay rate based on interviewer experience and any special skills they may possess and require (e.g., bilingual interviewers, phlebotomists, etc.).

2.6 Consider offering incentives for work above a certain target (e.g., response rate, contact rate, refusal conversion rate) as a way to keep interviewers motivated (ESS, 2004; Weisberg, 2005).
 2.6.1 Incentives can be extra pay, prizes, or special rewards.
 2.6.2 Overreliance on interviewer incentives for completed interviews may give interviewers a reason to fabricate interviews (Weisberg, 2005).
 2.6.3 Any bonus system must be perceived by the interviewers as being fair. For example, different sample assignments can vary considerably in the challenges they pose for interviewers (Cannell, Marquis, & Laurent, 1977).

Lessons learned

2.1 Most survey organizations have a standard policy concerning pay arrangements (either paying per interview or paying by the hour) which they may be unwilling to change (Cannell et al., 1977).

2.2 If interviewers are paid by the interview instead of by the hour, they may rush the critical respondent-interviewer rapport-building process. It is especially important for face-to-face interviewers to spend the time necessary to develop this rapport so that respondents feel comfortable reporting honestly, as this leads to higher-quality responses. For example, when approaching a household, face-to-face interviewers need to conform to the culture’s introductory customs, such as drinking tea or meeting elders, which require additional time spent by the interviewer (Hursh-César, 1976).

2.3 To discourage hurried, sloppy work when paying per interview, some organizations set a cap on the number of interviews that each interviewer is allowed to conduct in a day or during some other time frame. Another strategy is to offer bonuses for high quality work. For example, set a basic pay per interview plus an additional 10% if the interviewer makes fewer than some predetermined number of errors. This requires the survey organization to have a monitoring system in place, which can distinguish between minor and more serious interviewer errors and can identify errors that cannot be attributed to the interviewer but rather to system factors, such as question wording and technology failures. See Survey Quality and Paradata and Other Auxiliary Data for further discussion on such systems.

2.4 In contrast to face-to-face interviewing, an experiment with telephone interviewers found that their productivity increased when they were paid per interview as opposed to being paid per hour (Cantave, Kreuter, & Alldredge, 2009).

3. Recruit and select an appropriate number of qualified interviewers.

Rationale

The quality of an interviewer-administered survey depends, to a large extent, on the quality of the interviewers and their supervisors. It is important, therefore, to recruit and select the best possible people for the job. In addition, selecting candidates who are well suited for the job may lead to lower interviewer turnover and reduced survey costs.

Procedural steps

3.1 Recruit applicants.
- 3.1.1. Often times, the research organization who will be conducting the data collection in an individual country will have its own in-house interviewing staff from which to select suitable interviewers for the particular survey.
- 3.1.2. The research organization may also have to implement outreach measures to find additional interviewers, such as asking local contacts for suggestions, placing flyers in strategic locations, and advertising in local papers or online. If this is necessary, recruitment and training will take longer and the cost may increase. The cost will vary by method as well.
- 3.1.3. The interviewing component of the study may also be subcontracted to an external survey organization with an existing pool of interviewers. If this will occur, it should be specified in the initial contract (see Tenders, Bids, and Contracts).

3.2 Target sources where potential interviewer candidates might be located, keeping in mind any special considerations, as described in Guideline 1. Professionals, such as traveling nurses, can be a good source of interviewers for health studies; teachers, or others with substantive knowledge of the study topic, may also be good candidates. However, these professionals must be willing to set aside other knowledge and training received if it differs from the study protocols.

3.3 Keep cultural norms and logistical factors in mind when recruiting interviewers. For example, it may not be acceptable in some cultures for young people (e.g., college students) to interview older persons or for women to interview men and vice versa. Similarly, persons with other jobs may not be available to work on the study at the times when respondents are most likely to be at home.

3.4 Clearly describe all requirements of the interviewing position in recruitment materials. In addition to reducing training costs, this can reduce the interviewer attrition that can occur when interviewers are not fully informed of study requirements and responsibilities until they attend training or begin field work.

3.5 Recruit more than the number of interviewers needed for data collection to allow for attrition and the dismissal of candidates who prove to be unsuitable.

3.6 Prepare an application form to use in prescreening interviewer candidates before they are invited to an in-person or telephone job interview as appropriate.

3.7 Consider interviewing applicants in the mode of the study. For example, hold telephone screening interviews for a telephone survey and face-to-face screening interviews for a face-to-face study. This also provides the applicant with the opportunity to demonstrate his or her use of a PAPI or a CAPI instrument, depending on the needs of the study.

3.9 Evaluate each candidate.
 3.8.1 If appropriate, conduct a criminal background check, particularly if the interviewers will handle sensitive information or come into contact with vulnerable populations (e.g., the young, the old, the infirm, etc.).
 3.8.2 Criteria for employment commonly include interviewing skills, language skills, computer or technical skills, organizational skills, education, availability, location, the ability to meet production (i.e., data collection) goals, and the capacity to handle potentially emotional or stressful interactions with respondents (Pennell et al., 2010).
 3.8.3 When possible, select interviewers who have previously worked on similar studies and have good recommendations based on their performance. Experienced interviewers require

less training and are likely to achieve higher response rates (Cannell et al., 1977; Fowler & Mangione, 1985).

3.8.4 Evaluate the accuracy and clarity with which each potential candidate can read and process the survey questions in the language(s) of the interview and make sure that he or she is comfortable reading out loud. Ideally, language proficiency should be formally assessed by an outside expert or language assessment firm and should include evaluation of (Pennell, Harkness, & Mohler, 2006):
- Conversational skills (e.g., comprehension level, comprehension speed, speech level, speech speed, and accent)
- Writing skills (e.g., grammar, spelling, and the ability to enter responses)
- Reading skills (e.g., reading aloud)

3.8.5 Realize that poor eyesight can lead to difficulty reading computer screens (Shirima et al., 2007).

3.8.6 If using a paper instrument, ensure that the applicant can follow questionnaire logic and instructions; if using a computerized interview, test applicants' computer skills.

3.8.7 Select interviewers who are punctual and have good organizational skills (e.g., are able to handle forms and keep track of paperwork).

3.8.8 Select interviewers who have completed the full period of required schooling within their country.

3.8.9 For face-to-face studies, assess applicants' ability to read or use maps or mapping software.

3.8.10 See Data Collection: Face-to-Face Surveys for additional discussion on interviewer recruitment and training considerations, particularly when the data collection instrument has any technological component (i.e., a table, laptop, smartphone, etc.).

3.9 Give the candidates a realistic preview of the job including the survey topic and the type of questions that will be asked; describe any non-traditional interviewing tasks (e.g., collecting biomeasures) in the recruitment description and the screening interview.

3.10 Clearly present the candidates with study expectations for workload (weekly, monthly, including evening work and possibly weekend work).

3.11 Obtain the candidates' written commitment to work at the expected level of effort for the duration of the data collection period.

3.12 Base selection on an objective evaluation of the candidate's abilities rather than his or her relationship to survey staff or favoritism (Nyandieka et al., 2002; Vaessen, Thiam, & Le, 2005; Afrobarometer Survey, 2010).

Lessons learned

3.1 Vaessen, Thiam, and Le (2005) suggest that study managers recruit at least 10 to 15 percent more than the number of interviewers ultimately needed for field work to allow for attrition and the dismissal of candidates who prove to be unsuitable.

3.2 A variety of selection criteria have been used successfully by established 3MC studies.

- 3.2.1 In the Afrobarometer Survey, interviewers (preferably women) usually hold first degrees in social sciences and have some university education, strong facility in the local language, and the ability to relate to respondents in a respectful manner. Selection is on a competitive basis and may include reading, speaking, and comprehension of national and local languages, and competence at following detailed instructions (Afrobarometer Survey, 2010).
- 3.2.1 The Asian Barometer recruits interviewers from among university graduates, senior social science undergraduates, and professional survey interviewers (Asian Barometer, 2010).
- 3.2.3 The European Social Survey highly recommends using experienced interviewers (ESS, 2010).
- 3.2.4 The Living Standard Measurement Study Survey requires that interviewers have completed secondary education and recommends fluency in two or more languages (Living Standard Measurement Study Survey, 1996).
- 3.2.5 The coordinating center for the Survey of Health, Aging and Retirement in Europe (SHARE) selects survey research organizations for all participating countries and requires interviewers to have extensive face-to-face experience (Survey of Health, Ageing and Retirement in Europe, 2010).
- 3.2.6 In the World Mental Health Survey, some participating countries use field staff from established survey organizations, while others recruit new interviewers from the general population or among college students. Interviewer criteria vary among participating countries and may include interviewing experience, language skills, technology skills, education, and capability to handle potential sensitive situations with respondents (Kessler, Ustun, & World Health Organization, 2008).

3.3 Students can be a good source of interviewers.
   3.3.1 In an experiment using interviewers from "scholarly networks" (senior or graduate students), government organizations, and survey firms, the research team from the Chinese General Social Survey found that student interviewers were the most trustworthy. Because it was somewhat easier to exert control over the student interviewers, they were able to monitor survey quality most easily within this group. The drawback of using the student interviewers was their limited availability (Bian & Li, 2012).
   3.3.2 In a study of childhood behaviors in Turkey, researchers posted announcements at various academic institutions. They successfully recruited senior psychology and counseling students to conduct semi-structured face-to-face interviews about behaviors related to a number of neurodevelopmental and neuropsychiatric disorders. (Cevikaslan, Evans, Dedeoglu, Kalaca, & Yazgan, 2013)

3.4 As revealed in a survey on data collection as part of the International Social Survey Program (ISSP), different types of people can be employed as interviewers, such as full-time professionals, part-time professionals, students and others (who are not in the labor force, and likely to work temporarily) (Smith, 2007). The types of people employed as interviewers differ greatly across cultures. For example, as mentioned in Smith (2007), in the ISSP, "a quarter of the countries use no part-time professionals and another quarter employ all part-timers. Likewise, over half of all countries have no full-time professionals, while almost a quarter have full-timers making up half or more of their staff. Similarly, over a third of countries use no student interviewers, while almost a fifth have a majority of interviewers who are students." The differences are likely due to different resources available for each country and local traditions.

3.5 In a panel study, it may be helpful to keep the same interviewer with the same respondent across rounds of the study. In New Zealand, focus groups were conducted with interviewers who worked on the Prospective Outcomes of Injury Study (POIS). The interviewers explained that the "personal connection generated between the interviewers and participants was important, and enabled successful follow-up rates for the study." They felt this connection allowed them to "negotiate the requirements of the interview within a relationship they form with participants" (Derrett & Colhoun, 2011).

3.6 Liamputtong, a professor in the School of Public Health at La Trobe University, argues that bicultural researchers who are familiar with both the local and mainstream cultures of communities in the study are ideal (Liamputtong, 2010).

3.7 As noted in Guideline 1, it is not always possible to recruit interviewers who are fluent in the language(s) preferred or needed by respondents. In this case, other arrangements must be made. Options may include working with interpreters, data collection by proxy, using a bridge language if available, or using a self-adminstered mode if literacy levels permit (see also Data Collection: Self-Administered Surveys).

3.7.1 A study was conducted during the 2010 Census in the United States to investigate Non Response Follow-up (NRFU) interviews with households that speak languages other than English in with heavy concentrations of residents with limited English proficiency. The researchers found that enumerators were far more likely to go off script in interviews they conducted with respondents in other languages than they were in interviews with English-speaking respondents. Interviewers relied on on-the-fly translation and use of interpreters – practices which enabled enumerators to complete nonresponse follow-up interviews but posed a potential threat to data quality. Issues observed include: (1) inaccurate and incomplete translation of census questions; (2) modifying census questions or skipping some questions completely; and (3) having someone, especially a child, act as an ad hoc interpreter, which created some communication problems or placed a cognitive and emotional burden on the under-age interpreter (Pan & Lubkemann, 2013)

3.8 If the topic is sensitive (e.g., domestic violence), empathy and strong interpersonal skills may be more important than high levels of education or previous interviewing experience (Jansen, Watts, Ellsberg, Heise, & García-Moreno, 2004). This holds true for both interviewers and any interpreters being used.

3.9 If the project's interviewing protocol differs significantly from previous studies, experienced interviewers may find it difficult to change their habits, leading to what is known as "veteran effects". In this case, it may be preferable to recruit and train new interviewers. Similarly, interviewers who have worked for an organization with low quality standards may have to unlearn some behaviors and adapt to new standards.

4. Provide general basic interviewer training.

Rationale

Newly hired interviewers and supervisors require basic training in techniques for successful interviewing before they receive specific training on the study on which they will be working. Research indicates that general interviewer training (GIT) helps improve the quality of survey data by: (1) reducing item nonresponse (Billiet & Loosveldt, 1988), (2) increasing the amount and accuracy of information obtained (Billiet & Loosveldt, 1988), and (3) increasing survey participation by teaching interviewers how to identify and respond to respondents' concerns (O'Brien, Mayer, Groves, & O'Neill, 2002).

Procedural steps

4.1 Allow sufficient time to adequately cover general interviewing technique (GIT) material. One option is to provide materials for interviewers to read and complete prior to attending in-person training. Training content can be provided electronically—either accessible online or on a DVD or CD that can be reviewed on a personal computer. Interviewers can read materials provided in advance, view videos or tutorials, and answer questions electronically prior to attending in-person interviewer training.

4.2 Select appropriate trainers. These may include research staff, project managers, project management assistants, supervisors who directly oversee data collection staff, and experienced interviewers.

4.3 Provide the following general information to the interviewers at the beginning of the training:
- 4.3.1 An overview of the survey research organization and introduction to all trainers present.
- 4.3.2 The roles of the interviewer and the supervisor in the research process.
- 4.3.3 The format of the survey interview.
- 4.3.4 An overview of different interview modes (face-to-face, telephone, computer-assisted, observation, and delivering self-administered survey materials such as diaries) and the tasks each poses for the interviewer.
- 4.3.5 An overview of the sample design and associated implications and tasks for the interviewer.
- 4.3.6 Interviewer evaluation procedures and criteria.

4.4 Include the following prescribed procedures in the general interviewer training (Fowler & Mangione, 1990):

4.4.1 Standardized question-asking. Train interviewers to read each question exactly as written and to read the questions slowly. They should ask all questions exactly in the order in which they are presented in the questionnaire (Doyle, 2004; Groves et al., 2009a) (see Guideline 5 for exceptions).

4.4.2 Questionnaire format and conventions. Teach interviewers how to enter the answers to both open- and closed-ended questions. Train them to follow interviewing conventions such as emphasizing words in the questionnaire which appear in bold or are underlined, recognizing and not reading aloud interviewer instructions, reading or not reading optional words as appropriate, and selecting correct fill choices (e.g., he/she, has/have, etc.).

4.4.3 Clarification. If the study staff has not prepared a stock definition, train interviewers to repeat all or a specified part of the question verbatim when respondents ask for clarification. Interviewers should not make up their own definitions to any word, phrase, or question in the questionnaire (Cannell et al., 1977). Train interviewers to notify their supervisors about any questions which are confusing to respondents and require further clarification.

4.4.4 Probing. If a respondent's answer is inadequate and it is legally and culturally permissible to probe (see Ethical Considerations), train interviewers to employ unbiased techniques to encourage answers that are more complete, appropriate, and thoughtful (Cannell et al., 1977; Groves et al., 2009a). Probes must be neutral; that is, they must avoid "sending a message" about what is a good or a bad response. Such strategies of probing for more information may include:

- A pause to encourage the person to fill the silence or a direct request for further information.
- Verbal probes chosen from a stock list of phrases such as "Could you explain what you mean by that?" or "Can you tell me anything else about ___________ ?"

4.4.5 Feedback. Train interviewers to provide their respondents with culturally appropriate feedback when they are doing well in order to encourage them to listen carefully and to give thoughtful answers (Cannell et al., 1977).

- This feedback may be in the form of a nonverbal smile or nod or a short encouraging phrase.
- Verbal feedback should be selected from a prepared list of stock phrases such as "That's useful information" or "Thank you, that's helpful" to ensure that the feedback is not evaluative of the content of the answer. For example,

in English the word "okay" is discouraged for use in feedback because it could be construed as agreement with or approval of the respondent's answer.

- As a general rule, give nonverbal or short feedback to short answers and longer feedback phrases to longer answers.

4.4.6 Recording answers. To reduce measurement error, train interviewers to record answers exactly as given.

- If the question offers fixed alternatives, teach interviewers to get respondents to choose one of the fixed alternatives; interviewers should not infer which alternative is closest to what the respondent actually says (Groves et al., 2009a).
- If the question requires a narrative response, teach interviewers to record the answer in as near verbatim form as possible (Groves et al., 2009a).

4.4.7 Confidentiality. Train interviewers to keep confidential all identifying respondent contact information as well as respondents' answers to survey questions. See Ethical Considerations and Data Collection: Face-to-Face Surveys for additional discussion on confidentiality.

4.4.8 Any Computer Assisted Personal Interviewing (CAPI) conventions used in the survey instrument (see Instrument Technical Design and Data Collection: Face-to-Face Surveys).

4.4.9 Completing contact attempt records. Teach interviewers to record when each contact was attempted, any pertinent respondent comments (e.g., the best time to reach him or her or reasons for reluctance to participate), and the result of each contact attempt, using disposition codes (further information on contact attempt records and disposition codes can be found in Data Processing and Statistical Adjustment; examples of contact attempt records can be found in Data Collection: General Considerations).

4.4.10 Recording time and meeting production goals. Teach interviewers how to record the time they spend on each defined aspect of their work for the study, both for their remuneration and to allow supervisors to monitor their progress and efficiency during data collection.

4.5 If legally and culturally permissible, teach interviewers non-coercive persuasion techniques and practice counter replies to common statements of reluctance.

4.5.1 Discuss optimal times and modes for contacting target persons.

4.5.2 Train interviewers to tailor their initial interactions with respondents by developing the following skills (Groves & McGonagle, 2001; O'Brien et al., 2002):

- Learning the classes of concerns ("themes") that respondents might have.
- Classifying the respondent's wording into the appropriate theme.
- Addressing the concern, using their own words.

4.5.3 Employ hands-on practice exercises so that the trainees become proficient in quickly identifying respondent concerns and quickly responding to them.

4.6 For best overall results, employ a training format that combines lecture with visuals and small-group practice sessions.

4.6.1 Mixing the format keeps the trainees engaged and acknowledges that different people learn in different ways (Galbraith, 2003).

4.6.2 Through practice, trainees move from procedural knowledge (knowledge of how to perform a task) to skill acquisition (the ability to perform the task almost automatically) (O'Brien et al., 2002).

4.6.3 Although the class can be large for lecture sessions, trainees should break up into smaller groups for hands-on practice.

4.7 Be sensitive to the local culture.

4.7.1 Educate trainers in cultural sensitivity.

4.7.2 Take religious holidays into consideration when scheduling training sessions.

4.7.3 Make every effort to accommodate dietary restrictions when planning meals or snacks for the training.

4.7.4 Be aware that conventions regarding breaks during training vary among cultures.

4.8 At the end of basic interviewer training, evaluate the knowledge of the interviewer candidates. This can be done by written test, conducting a scripted certification interview with a supervisor, audio taping, or observing the interviewer conduct an actual practice interview.

Lessons learned

4.1 If the interviewer candidates have access to the necessary equipment, some basic interview training material can be presented in the form of audio- or video-recordings for home study (California Health Interview Survey, 2007; Federal Committee on Statistical Methodology, 1998). Other training options include telephone and video conferencing and self-study using paper materials.

4.2 West & Olson (2009) found that interviewer-related variance on survey items may be due to nonresponse error variance rather than measurement difficulties. That is, different interviewers may successfully contact and recruit respondents with different characteristics (e.g., age, race), even though their sample pools start out the same.

4.3 Interviewer training and the interviewer manual need to be adjusted to be culturally sensitive to the population under study:

4.3.1 Textbook instructions on handling reluctance to participate and provide accurate information rely to a large extent on Western experiences. When possible such procedures should be modified so that they include culturally acceptable and suitable tactics. Researchers conducting a women's health study on the Apsáalooke native American reservation in southeastern Montana, U.S.A., felt that standard Western tactics for handling reluctance would be offensive in that culture. They therefore did not attempt to persuade reluctant respondents to participate. In addition, interviewers were encouraged to display a compassionate attitude and interest in the women, rather than the standard recommended neutral voice tone and lack of responsiveness to respondent answers, to minimize eye contact, and to accept offers of food and drink – all to be more consonant with the Apsáalooke culture (Christopher, McCormick, Smith, & Christopher, 2005).

4.3.2 In some countries, a Western trainer may be respected but resented. Researchers in Puerto Rico found allowing interviewer trainees to provide input about the local culture and supplementing trainer criticism with peer criticism helpful (Stycos, 1952).

4.3.3 The World Mental Health study added country-specific topics to their general interviewer training sessions. In New Zealand, they included cultural empathy to Maori and Pacific Islander households; in Colombia, they provided special training on interacting with governmental authorities and armed guerrilla and paramilitary groups (Pennell et al., 2010).

4.4 In the SHARE data collection, interviewers were trained to record all contacts and contact attempts and to take notes in order to tailor approaches for maximizing contact. This information allowed the researchers to observe considerable variation by country in how contact strategies were implemented (Alcser, Benson, & Guyer, 2011).

4.5 In an initial pilot study, the Japanese Study of Aging and Retirement (JSTAR) suffered low response rates in urban areas, due to the inability to access respondents in locked buildings. Staff from the University of Michigan provided additional training to field managers in accessing housing units in locked buildings. The response rates in a second pilot increased significantly after interviewers were trained and comfortable using the specific techniques for approaching locked units, gaining access, and returning to households where access was not previously obtained. These included identifying and talking with building managers, providing additional study information to building managers to gain their confidence in the research while still protecting respondent confidentiality, approaching other housing units when access to one unit is gained, returning to the building at different times of day and varying days of the week, and noting patterns of building resident arrival and departure times (Guyer, Alcser, & Kirgis, 2008).

5. **Provide study specific training for all interviewers and supervisors.**

Rationale

Interviewers and supervisors need to be very familiar with the study's protocols and questionnaire in order to carry out their tasks. Depending upon the survey, they may need to learn the instrument's branching structure, the study's requirement for field coding, or the use of a respondent booklet, show cards, or other visual materials. There may be special instructions for implementing all or part of the survey that deviate from the standardized interviewing covered in general interviewer training. Interviewers should also be knowledgeable about the project objectives so that their actions help, not hinder, the overall goals. Both newly hired and experienced interviewers as well as supervisors require training specific to the study at hand.

Procedural steps

5.1 Allow sufficient time for study-specific training, depending upon the complexity of the study (see Appendix C for a sample training agenda).

5.2 When possible in a 3MC survey, have the same team from the coordinating center train all interviewers from all study countries to ensure standardization of study-specific protocols (Ustun, Chatterji, Mechbal, & Murray, 2005) (see Study Design and Organizational Structure for additional discussion on the role of the coordinating center). The team may provide regional trainings, traveling to where interviewers are located.

5.3 Select appropriate trainers. These may include research staff, project managers and people on their staffs, supervisors who directly oversee data collection staff, experienced interviewers, and consultant(s) hired to assist with interviewer training.

5.4 Include a large amount of practice and role playing using the questionnaire (Ustun et al., 2005).

5.4.1 Consider having the interviewers complete a self-interview to become familiar with the survey instrument.

5.4.2 Hands-on training may include round-robin practice sessions (i.e., scripted practice sessions where interviewers take turns administering survey questions to the trainer in a group setting), mock one-on-one interviews (i.e., sessions where interviewers interview each other), listening and discussing taped interviews, and live practice with potential respondents.

5.4.3 For role playing to be effective, prepare different scripts in advance so that the different branching structures of the interview, the nature of explanations that are permitted, and anticipated problems can be illustrated.

5.4.4 Consider making a video to illustrate the correct administration of any biomeasures, if applicable. This ensures that the material is consistently taught, especially if training is conducted at multiple times or in various locations.

5.5 Provide interviewers with an Interviewer Project Manual/Study Guide that has been prepared by the coordinating center, with input from local collaborators. The manual is an important part of training and will serve as reference material while the survey is underway (Glewwe, 2005).

5.5.1 Complete and review the manual before training begins (Glewwe, 2005).

5.5.2 When appropriate, translate the manual into the languages used in the geographical areas encompassed by the study.

5.5.3 Include the following content in both the training agenda and the project manual:

- General information about the project (e.g., the study's background and goals, funding sources if relevant, and principal investigators).
- How to introduce the survey to respondents.
- Eligibility and respondent selection procedures, if applicable. Sampling and coverage errors can occur if interviewers fail to correctly locate sample households, determine eligibility, or implement the respondent selection procedure (Martin, 1996).
- Review of the survey instrument, highlighting the content of the various sections and the types of questions being

asked.

- Data entry procedures for the mode(s) of instrument used (e.g., paper-and-pencil, CAPI, etc.). Measurement error can occur if interviewers do not record responses in the appropriate manner.
- Computer hardware and software usage, if appropriate (e.g., use of the laptop computer, email, and any other software packages).
- Use of the sample management system.
- Review of interview procedures and materials (e.g., informed consent materials and respondent incentive payments).
- Review of study-specific probing conventions (e.g., when to probe a "don't know" response and an open-ended response).
- Techniques for handling reluctance that are specific to the study (e.g., recommended responses to frequently asked questions) and are approved in advance by an ethics review committee (see Ethical Considerations). Nonresponse bias can occur if interviewers are unable to persuade reluctant persons to participate in the survey.
- Nonstandardized interviewing, if appropriate for the study (e.g., event history calendars, time diaries, or conversational interviewing) (Beatty, 1995; Belli, Shay, & Stafford, 2001; Conrad & Schober, 1999; Groves et al., 2009a; Suchman, & Jordan, 1990). (See Data Collection: General Considerations for a discussion about combining qualitative and quantitative data collection methods.)
- Any observational data which interviewers will be required to enter (e.g., observations of the respondent or the neighborhood). See Paradata and Other Auxiliary Data.
- Any specialized training for the study (e.g., procedures for taking biomeasures, instruction on interviewing minors or interviewing on sensitive topics, proxy interview protocol, interviewing in unsafe neighborhoods, and protocol for handling respondent or interviewer distress).
- Procedures to be used for unusual cases, including general principles to be applied in dealing with unforeseen problems (e.g., how to report abuse of children or others that is observed while conducting an interview in the respondent's home). See Ethical Considerations for additional discussion of interviewer obligations in the course of fieldwork.
- Production goals and maintaining productivity.
- Proper handling of equipment and survey materials.
- The structure of the survey team and the role of all

members of the team.
- Procedures for editing and transmitting data. Processing error can occur if interviewers do not correctly edit and transmit the completed questionnaire (see Data Processing and Statistical Adjustment for other potential sources of processing error).
- Any other required administrative tasks. See Data Collection: Face-to-Face Surveys and Data Collection: Telephone Surveys for additional discussion on mode-specific material to include during interviewer training.

5.5.4 The Project Manual/Study Guide must be especially clear and self-contained if it is impossible to train interviewers in person (e.g., if interviewers must be trained via conference call or video).

5.6 Develop and implement interviewer training appropriate to the instrument.

5.6.1 Include equipment-specific training, such as an overview of the hardware and software systems, password use, stylus use, if needed, questionnaire access, entering responses, charging the battery, general care and maintenance, and how to insert and remove memory cards.

5.6.2 When using new technology, interviewers tend to focus on the technology rather than the respondent; this should be addressed during interviewer training.

5.6.3 Determine whether paper questionnaires will be available in the event of equipment malfunction; if this is the case, training on the PAPI instrument is also essential.

5.6.4 See Data Collection: Face-to-Face Surveys and Data Collection: Telephone Surveys for additional discussion on mode-specific material to include during interviewer training.

5.7 Collect and analyze written evaluative feedback (i.e., provide the opportunity for trainees to give written feedback on trainer performance, the sufficiency of time allocated to different topics, and the adequacy of practice exercises).

5.8 Certify the interviewers; see Appendix D for a sample interviewer certification form. Certification for study-specific tasks should include:

5.8.1 A complete role-play interview with a supervisor.

5.8.2 Certification by an appropriate trainer for any biomeasures are included in the study (see Appendix E for a sample certification checklist for taking physical measurements).

5.8.3 Language certification, as appropriate (see Translation: Overview).

5.9 Supplement the initial training with periodic in-person seminars, telephone conference calls, and periodic bulletins or newsletters (Pennell et al., 2010).

5.10 If data collection will extend for a long period of time, hold a brief refresher training course towards the middle of the data collection period (Office of Management and Budget, 2006).
5.10.1 This refresher training session is an opportunity to review various aspects of data collection, focusing on difficult procedures or on protocols that are not being adhered to sufficiently by interviewers.
5.10.2 The session can also be used to provide feedback on what has been achieved to date.
5.10.3 Require even experienced interviewers and supervisors to attend refresher training sessions, including sessions on standardized techniques.

Lessons learned

5.1 Most of the time it is not feasible for the same team to train all interviewers, particularly in very large 3MC studies. If this is the case, other steps must be taken to ensure the standardization of study-specific protocols:
5.1.1 One approach is the "train-the-trainer" (TTT) model.
- 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 the interviewers.
- This model allows for tailoring at the country or cultural group level.
- The TTT program in SHARE is one example of this approach (Alcser & Benson, 2005; Alcser & Benson, 2008; Börsch-Supan, Jürges, & Lipps, 2003; Survey of Health, Ageing and Retirement in Europe, 2010). The University of Michigan's Survey Research Center, under contract to SHARE, created the TTT program. Each participating country sent a Country Team Leader, a member of his or her staff, and 2-3 trainers to the TTT sessions. Once the trainers had completed the TTT program, they used the training materials provided, translated if necessary, to conduct country-level interviewer training (see Appendix C for the SHARE Model Training Agenda).
- The SHARE team also found that under the TTT model,

having the use of training session observations at the regional level was effective for identifying deviations from the project objectives that could potentially contribute to systematic interviewer effects (Alcser et al., 2011).

- The Chinese General Social Survey provides another example of regional trainings being conducted for supervisors and local trainings being conducted for interviewers. In this model, interviewers were not selected to work until training was completed and interviewers had demonstrated successful interviewing skills (Bian & Li, 2012).
- The World Mental Health Survey gives two TTT sessions for interviewer supervisors, lasting, on average, six days. Interviewer supervisors in turn, train the interviewers in general interviewing techniques (on average 20 hours) and Composite International Diagnostic Interview (CIDI) specific training (on average 30 hours). Before progressing to CIDI specific training, interviewers must demonstrate competence, in the form of role playing, tests, and/or supervised respondent interaction, and in general interview techniques. All interviewers must be tested and certified before they are authorized for production work (Kessler et al., 2008).

5.1.2 Another approach is the training center model (Pennell et al., 2009).

- A centralized training course is held, but language "regions" are represented rather than countries.
- This model is effective when it is not possible for every country to send trainers who are functional in the central trainer's language.
- The training center model was used in the World Health Organization's Composite International Diagnostic Interview training sessions. For example, trainers from Lebanon were trained in the United States and subsequently trained the trainers in Lebanon, Oman, Jordan, Palestine, Saudi Arabia, and Iraq.

5.1.3 Organizing training in steps (first training the trainers and then having them train the interviewers) increases the overall time needed for training, which should be factored into the project timeline.

5.1.4 All step-wise training results in a certain loss or distortion of information as it is passed along. Trainers should be aware of this and take precautions, such as providing approved standardized training materials.

5.2 If interviewers are being hired for one study only, basic interviewer training techniques can be incorporated into study-specific training.

5.3 The amount of time devoted to training varies among large established 3MC surveys. The variation is likely due to different constitutes of the interviewer (full-time, part-time or students and others), different resources available (e.g., whether the survey has its own field staff or not), organizational structures, and cultural traditions.

5.3.1 Glewwe (2005) recommends up to a month of intense interviewer training (general and study specific) for inexperienced interviewers in a face-to-face survey.

5.3.2 Field team members for the Asian Barometer received intensive, weeklong training sessions on the questionnaire, sampling methods, and the cultural and ethical context of the interview (Asian Barometer, 2010).

5.3.3 The Living Standard Measurement Study Survey (LSMS) recommends that training take place over a four-week period and include introduction to the LSMS survey, general survey procedures, the questionnaire, sampling procedures, and data entry program error reports, with at least two observed training interviews (Living Standard Measurement Study Survey, 1996).

5.3.4 SHARE requires 16-18 hours of training spread over 2-3 days in addition to the basic interviewer techniques training for new interviewers (Alcser & Benson, 2005; Alcser & Benson, 2008).

5.3.5 Similarly, the World Health Survey (WHS) recommends three full days of study-specific training (Ustun et al., 2005).

5.3.6 Round 4 of the Afrobarometer Survey held a six-day training workshop for all persons involved with the project, including interviewers and field supervisors. The Afrobarometer protocol requires holding a single national training workshop at one central location. Interviewers must complete at least six practice interviews before they leave for the field: at least one mock interview in the national language, at least one mock interview in each of the local languages they will use in the field, and at least four training interviews in a field situation (Afrobarometer Survey, 2010).

5.4 In addition to general interview training, all interviewers for Round 5 of the European Social Survey were briefed by the National Coordinator or a research team member regarding respondent selection procedures, registration of the calling process, response rate enhancement, coding of observation data, documentation, and questionnaire content (ESS, 2010).

5.5 If the topic is extremely sensitive, additional specialized training may improve response rates and data quality. The WHO Multi-Country Study on Women's Health and Domestic Violence, fielded in multiple culturally diverse countries, found that previously inexperienced interviewers who had received specialized training obtained a significantly higher response rate and significantly higher disclosure rate of incidences of domestic violence than did experienced interviewers who had not received the additional training (Jansen et al., 2004).

5.6 Training interviewers in adaptive behavior, such as tailoring responses to respondent concerns or nonstandardized conversational interviewing, can be time-consuming and could increase training costs (O'Brien et al., O'Neill, 2002).

5.7 If data are to be collected via a computerized instrument (i.e., laptop, table, smartphone, etc.) Personal Digital Assistant (PDA)'s, interviewers will need training on their use.

5.7.1 In a survey in Bolivia using PDAs, interviewers, who had limited previous experience with the technology, wanted additional practice time. They particularly wanted additional instruction on the use of a stylus, since the keyboards on handheld devices can be cumbersome (Escandon, Searing, Goldberg, Duran, & Monterrey Arce, 2008).

5. 7.2 Analyses of inter-observer accuracy and performance revealed a considerable range in a survey in Burkina Faso. Some interviewers clearly worked faster with the PDAs than others, though these were not necessarily the ones who covered the greatest number of households per day worked. However, those who carried out interviews relatively quickly were generally also the ones who made the fewest input errors. In surveys of this kind, where competence in local languages is important, there is often a limited pool of potential interviewers. Different types of interviewers can be considered including students and part-time professionals. It is found that "school leavers" in one of the world's poorest societies were, in general, able to do a good job interviewing using PDAs (Byass et al., 2008).

5. 7.3 Training on the proper handling and care of equipment is very important, particularly in a rural context where the equipment must be transported through rough terrain, the power supply is unstable, and unexpected rain is a concern. In the DHS survey in Nepal, teams were provided with generators, rain shields, umbrellas, and other items to manage these challenges. Enforcing joint responsibility for the theft of, or damage to, the tablet PCs among the interviewer teams

helped to ensure security of the tablets during transport and storage. With proper care and maintenance, tablet PCs (and portable generators) can be reused in future surveys, resulting in additional cost savings over the long term (Paudel, Ahmed, Pradhan, & Dangol, 2013).

5.8 In the pilot for the UK Household Longitudinal Study (UKHLS), interviewers reported that the certification process for biomeasure data collection enhanced their confidence in being able to execute these tasks accurately (McFall et al., 2012).

5.9 Field interviewers often work some distance away from their trainers and supervisors. Before sending the interviewers to their assigned areas, some organizations have found it useful to have them conduct a few interviews close to the training locale. Afterward, they meet with the trainer, discuss their experiences, and check their questionnaires. Any problems or misunderstandings can be identified and rectified more easily than if they had occurred in a more remote area.

5.10 During pretesting for the Tamang Family Research Project, investigators trained interviewers in a Nepalese village that was not in the sample. The investigators and interviewers lived together during this period and throughout data collection. This allowed for the continuous assessment of interviewers who were let go if they were not completing quality work (Axinn, Fricke, & Thornton, 1991).

6. Institute and follow appropriate quality control measures.

Rationale

Quality control (QC) is a procedure or set of procedures intended to ensure that a product or service adheres to a defined set of quality criteria or meets the requirements of the study (see Survey Quality). The implementation of quality control measures enhances the accuracy, reliability, and validity of the survey data and maximizes comparability of these data across cultures. To implement an effective QC program in a 3MC survey context, the coordinating center must first decide which specific standards must be met. Then real-world data must be collected and the results reported back to the coordinating center. After this, corrective action must be decided upon and taken as quickly as possible. Finally, the QC process must be ongoing to ensure that remedial efforts, if required, have produced satisfactory results.

Procedural steps

6.1 Assess the cost and success rates of different recruitment avenues to determine which are the most fruitful and cost effective; use this information to guide the future allocation of resources.

6.2 Considering the factors enumerated in Guideline 3, establish a checklist of minimum interviewer candidate requirements (e.g., interviewing skills, reading/writing fluency, language skills, educational level, and computer skills).
 6.2.1 Require recruiters to complete the checklist as they screen each interviewer candidate. If specific assessment tests are used (e.g., to evaluate language skills), record each candidate's performance on the test.
 6.2.2 Accept only those candidates who meet the predetermined minimum requirements.
 6.2.3 To ensure accountability, require the recruiter to sign or initial checklists and assessment tests.

6.3 Survey interviewer candidates to determine what improvements could be made to the recruitment process; use this information to modify the procedure, if possible (for example, ask how the candidate heard about the position).

6.4 Take attendance at general interviewing techniques and study-specific training sessions.
 6.4.1 Dismiss candidates who fail to attend a predetermined minimum number of training sessions, or make arrangements to train them individually on the missed material.
 6.4.2 Keep a signed written record of the training completed by each candidate.

6.5 At the end of basic interviewer training, evaluate the knowledge of the interviewer candidates, as described in Guideline 4.
 6.5.1 Require all trainers to use the same evaluation criteria.
 6.5.2 Dismiss or retrain those candidates who fail to attain predetermined minimum standards.
 6.5.3 Keep a signed written record of each candidate's performance on the evaluation measures.

6.6 At the end of study-specific training, certify the interviewer candidates, as described in Guideline 5.
 6.6.1 Require all trainers to use the same evaluation criteria for certification.
 6.6.2 Dismiss or retrain those candidates who fail to attain predetermined minimum standards.

6.6.3 Keep a signed written record of each candidate's performance on the certification tests.

6.7 Debrief interviewer trainees to determine how training could be improved; use this information to modify the training protocol, if possible.

Lessons learned

6.1 Including quality control protocols as part of the overall survey design, and implementing them from the start, permits the survey organization and the coordinating center to monitor performance and to take immediate corrective action when required. For example, if many interviewer candidates fail to pass the study-specific certification test, additional training could be provided. Afterward, the candidates would be tested again. Those passing the certification test could then be sent out into the field.

7. Document interviewer recruitment and training.

Rationale

Comprehensive documentation helps analysts correctly interpret the data and assess data quality; it also serves as a resource for later studies.

Procedural steps

7.1 Document the recruitment effort for enrolling data collection staff on the project, including:
- 7.1.1 Any special criteria used in reviewing data collection staff employment applications (e.g., language proficiency and special knowledge and skills, such as taking physical/biological measurements).
- 7.1.2 The way in which language fluency was assessed, as appropriate for the study.
- 7.1.3 Recruitment scripts and sources used to recruit data collection staff, as well as an evaluation of the success of the recruitment strategies.
- 7.1.4 Interviewer characteristics (e.g., gender, age, race, education, length of tenure as interviewer).
- 7.1.5 Characteristics of the multilingual interviewing staff in terms of the percent certified to interview by language.
- 7.1.6 The minimum number of hours required, if applicable, and the average number of hours worked by an interviewer during the data collection period.

7.1.7 Interviewer pay structure (e.g., hourly or per completed interview), the pay range, and any bonus program (e.g., amount and when or under what circumstances these bonuses were offered).

7.2 Document the general and study-specific training, including:
- 7.2.1 Number of training sessions conducted.
- 7.2.2 Number of training hours, dates, and locations.
- 7.2.3 Number of trainers and trainees.
- 7.2.4 Background of the trainers, including expertise in training and in any substantive areas as applicable to the survey.
- 7.2.5 Copy of the training agenda(s) (i.e., list of topics covered).
- 7.2.6 All written materials that were used (e.g., the interviewer manual/study guide, trainer/facilitator guide and supplemental training materials).
- 7.2.7 Certification procedures (e.g., scripted certification interview with a supervisor or other staff, written or online test on general interviewing procedures, live practice interviewing with potential respondents).

7.3 Document any issues encountered (e.g., if the recruitment plan failed to produce a sufficient number of qualified interviewers or interviewer attrition was unexpectedly high, necessitating a second round of recruitment and training; the training agenda did not provide adequate time for hands-on practice; or the ratio of trainers to trainees was inadequate) and suggestions for future studies.

7.4 Document all direct measurements of data quality, all indicators of data quality obtained via quality control (QC), and any decisions made to change the protocol in order to maintain high levels of quality (see Survey Quality).

Lessons learned

7.1 Documenting the recruitment effort, including method(s) of recruiting, number of candidates recruited, and number of candidates screened, as well as post-study documentation of interviewer retention, is also useful for other future projects. This information can guide future recruitment strategies and help estimate the number of recruits needed to provide a sufficient number of interviewers for data collection in similar studies.

7.2 Documentation of general and study-specific training can pinpoint areas needing improvement in future training efforts.

Appendix A

Interviewer Design Effect (Groves et al., 2009a; Kish, 1962; Office of Management and Budget, 2006; O'Muircheartaigh & Marckward, 1980)

Research indicates that the interviewer design effect may be even larger than the design effect attributable to geographic clustering (Schnell & Kreuter, 2005). This is especially true in some 3MC studies where cultural and other factors contribute to large interviewer variances; and this variance can differ between countries as well. Interviewer variance occurs when response errors of persons interviewed by the same interviewer are correlated; therefore interviewer variance is part of the correlated variance component of the total variance (other correlated variances stem from coders, editors, supervisors and crew leaders).

The intra class coefficient, ρint, is a measure of the ratio of interviewer variance to the total variance and is defined as:

$$\rho_{int} = \frac{\text{(between-interviewer variance)}}{\text{(between–interviewer variance)} + \text{(within-interviewer variance)}}$$

The value of ρint is theoretically always between 0 and 1 although calculated estimates of ρint may sometimes be negative. In this case, they are usually treated as zeros. When ρint for a particular variable is 0 or is negative, we interpret this to mean that the interviewers have no effect on the variance of responses to that variable; the larger the value of ρint, the larger the effect of interviewers on the variance of the particular variable.

The interviewer design effect (deffint) is a measure of the effect of interviewers carrying out multiple interviews, compared to what you would get if there was a different interviewer for each respondent, all else being equal (if the addition of more interviewers increases costs such that supervision or training must be reduced to compensate, interviewer variance may actually increase).

$$deff_{int} = 1 + \rho_{int}(m-1)$$

where m is the average number of interviews per interviewer.

Thus, even a small interviewer variance (ρint,) can have a significant effect on the variance of a survey estimate if m is large. The interviewer variance contribution is usually not included in textbook variance estimation formulas. Interviewer variance leads to a loss of sample information when the effective sample size neff, defined as n/deffint, is smaller than the actual sample size n.

Standardized interviewing aims to reduce interviewer variance.

For specification of a mathematical model of response errors when interviewers are used, see Hansen, Hurwitz, Marks, & Mauldin (1951); for further discussion of interviewer variance see Biemer and Lyberg (2003) and Groves (1989).

See Statistical Analyses for a discussion on incorporating interviewer effects into multi-level models.

Appendix B

Estimating the number of interviewers needed for a study

The following example shows how to calculate the number of interviewers required for a hypothetical study. The example makes the following assumptions:

1. Interviewers and respondents do not need to be matched on any attributes.
2. The average number of hours worked per week is the same for all interviewers.
3. The expected number of completed interviews is 500.
4. The estimated Hours Per Interview (HPI) is 5.
5. The projected data collection period is 5 weeks.
6. Each interviewer is expected to work 15 hours per week (based on the optimal hours of work during the times the respondents are expected to be at home).

Make the following calculations:

1. Total hours to complete the study = (500 interviews * 5 HPI) = 2500 hours.
2. Average interviewer hours per week = (2500 total hours/5 weeks) = 500 hours per week.
3. Number of interviewers needed = (500 hours per week/15 hours per interviewer per week) = 33 interviewers.

To determine the optimum number of interviewers based on interviewer variance and cost, see Hansen et al. (1951).

Appendix C

Example of a Training Agenda (Alcser & Benson, 2008)

The Survey of Health, Ageing and Retirement in Europe (SHARE) utilizes a model agenda for training interviewers in participating countries. While the content of this agenda is SHARE study specific, it might provide a useful basic template for other similar cross-national survey efforts. Organizations may add country-specific items to the model training agenda (e.g., tracing/locating steps that should be followed in their country and any relevant cultural considerations).

Note that SHARE is a longitudinal (i.e., panel) study. However, new countries join at each wave, or a refresher sample is recruited – hence SHARE provides training for panel study and baseline study at most of its trainings.

The model training agenda assumes that interviewers have already received basic training in General Interviewing Techniques (GIT). However, since SHARE wants to make sure that certain specific GIT interviewing conventions are always implemented, SHARE spends part of the study-specific training reviewing those.

See the SHARE website for details about the study (Survey of Health, Ageing and Retirement in Europe, 2010).

SHARE Model Training Agenda (na=not applicable)

Topic	Purpose	Panel: Time (minutes)	Baseline: Time (minutes)
Introductions, Welcome, and Logistics	Set the stage for this intense training.	15	30
SHARE Project and Questionnaire Overview	Explain the goals of the project and the importance of baseline and longitudinal sample.	45	45
Sample Overview	Understand how the sample was selected, sample eligibility, and response rate requirements.	30	60
GIT Requirements	Cover minimal GIT requirements, including when and how to contact sample, probes, feedback, etc.	60	60
Overview of the Sample Management System	Learn how to operate the SHARE electronic sample management system, assign result codes, and enter call notes. Introduce noncontact mock scenarios and test results.	60	90
Longitudinal Sample Management System	Introduce splitters, deceased, new eligible respondents, and additional result codes.	30	Na

Proxy Interviews	Explain how to identify and interview proxy respondents.	30	45
Nursing Homes	Explain how to contact respondents in nursing homes and to work with gatekeepers / potential proxy respondents.	30	Na
Overview of the Blaise Program	Explain the Blaise program conventions, including different types of questions, question wording, data entry, interviewer instructions, etc.	45	45
SHARE Questionnaire Walk-Through	Describe SHARE modules. Conduct a scripted review of the questionnaire, including spawning of additional line. Address main questions and issues that arise with different sections. Longitudinal: Describe longitudinal differences. Explain preloads. Address different questions arising from reinterviews.	330	240
End-of-Life Interviews (EOL)	Cover the concept of the EOL interview, approaching respondents, and administering the interview. Explain how to record these in the Sample Management System (SMS).	30	Na

Total Time Training for the Panel Model: 1080 minutes (18 hours, 0 minutes)
Total Time Training for the Baseline Model: 1120 minutes (18 hours, 40 minutes)

Appendix D

Example of an Interview Certification Form

NOTE:

- The aim of certification is to assess the interviewer's conduct of the interview, including introducing the study, doing the interview itself, using all appropriate respondent materials and interviewer aids, closing out the interview, and recording all required information in the sample management system.
- Specific studies should modify items in this form, as needed, to ensure that all key elements are measured in the certification.
  - For example, the template below assumes that an electronic sample management system (SMS) is used; if a paper coversheet is used to manage the sample, one should develop items appropriate for that system.
  - Similarly, the template assumes that the interview is programmed in Blaise, which is a data collection software program; if data is collected via paper and pencil, one should check the interviewer's comfort in following routing instructions, choosing appropriate fills, etc.
- Additional items may be included on the form and scoring may be changed to suit the situation. Some potential additions might include (a) professionalism (e.g., pace, tone and emphasis of speech), (b) establishing rapport with the respondent, (c) introducing the study to the respondent, and (d) the administration of specific areas in the instrument, such as cognitive tests or mental health questions.
-

Certifier Notes for Individual Certifications			
Interviewer:		**Certifier:**	
Time:		**Location:**	
CERTIFIER INSTRUCTIONS: Score each item 0, 1, or 2. 0 = Inadequate performance; 1 = Needs Improvement; 2 = Met Expectations. Use the Errors column to tally the number of times the interviewer makes general interviewing technique (GIT) errors in reading, probing, feedback, or clarification. Note question numbers of errors when possible.			
Interviewing Skill	**Score**	**Errors**	**Comments**
On time and prepared for certification			Sample Management System running and ready to interview; for face-to face interview, have respondent materials ready, including copy of letter and brochure.

Correctly completing household listing/enumeration and screener			Make sure that the interviewer has completed the household listing/enumeration correctly; if not, tell him/her how to correct and proceed. **The interviewer will have to re-certify on the screener portion if this happens.**
Use of GIT probes and clarification			Should use standard GIT protocol as indicated; 1 - 2 errors - score 1; 3+ errors - score 0.
Use of neutral feedback			Interviewer should provide feedback for at least 30% of responses. Non-standard feedback counts as an error.
Verbatim question reading			Include pronunciation and emphasis in evaluation; 1-3 errors - score 1; 4+ errors - score 0.
Data Entry			General comfort with navigating in Blaise.
Post-interview process & contact person information			Interviewer should confirm all contact information for respondent and enter information for required number of contact persons.
Contact attempt record			Interviewer should enter a final contact attempt note which you will check before scoring. If 1 or 2 items are missing - score 1. If more than 2 items are missing - score 0.
TOTAL SCORE	0		
Total possible = 16 Certified = 12or higher Re-Certify = 10-11 Administrative Review will be required if score is less than 10.			
GENERAL COMMENTS: Provide specific examples and question numbers of problem areas when possible. Note the way in which the interviewer administered the informed consent and reads the script to explain the need for obtaining information for contact persons.			
Debriefing with Interviewer by [NAME]:	Date:		
Notes: Include summary of recertification plan and retraining or practice interviews needed. Make note of areas that need close review on taped interviews.			

Appendix E

Example of a Certification Checklist for Biomeasures

NOTE:

- Interviewers should act as though this is a real interview. It is recommended that the person performing the certification ("certifier") observe a pair of interviewers where one acts as the respondent and the other is the interviewer being evaluated. If the "interviewer" asks questions during the certification, such as "should I ask/do this...", neither the certifier nor the "respondent" should respond.
- Make sure that the list of supplies is first checked off (interviewer has all materials ready or not).
- Observe that the interviewer reads all instructions and explanations to the respondent and enters the values correctly.
- If the interviewer performs a given activity correctly, make a "check" in the column labeled "Correct" for that activity; if the interviewer does not perform the activity correctly, circle the number in the column labeled "Incorrect" for that activity.
    - The numbers in the "Incorrect" column indicate the importance of the activity as defined by the researcher. Individual researchers can establish the relative "weight" of the error score, as necessary.
- For each physical measurement total the circled numbers and enter the sum in the row labeled "Total Incorrect;" also enter the total on the "Total Incorrect" line below the table. Assess whether or not the interviewer has passed the section. To be certified, the interviewer must successfully pass all sections.
    - Say, for example, the interviewer failed to correctly perform the Blood Pressure activities "arm on table" and "use correct cuff size." The certifier would circle the Incorrect scores of 2 and 4 respectively, for a Total Incorrect score of 6. Since the "Max incorrect to pass" is 3, the interviewer would not pass this section and would need to be re-trained and re-certified.
- At the end, be prepared to provide feedback regarding the certification items and whether the interviewer passed certification. Make a decision about whether to permit recertification (for all measures or for only those that the interviewer did not pass) and be sure to let some time pass before attempting a recertification.
- Retain final, signed records as documentation of this certification. Some 3MC studies may require some form of documentation on interviewer certification levels across member countries.

BIOMEASURES CERTIFICATION CHECKLIST

Interviewer's Name: ______________________________

Certifier's Name: _____________________________

Date of Certification: ___________________________

Blood Pressure

Total incorrect Blood Pressure: ___________

Max incorrect to pass: 3 (4 or more needs re-certification)

Activity	**Correct**	**Incorrect**
Feet flat on floor/legs uncrossed		**2**
No smoking		**2**
Loose clothing/ no more than one layer		**4**
Arm on the table (or supported) at heart level		**2**
Use of correct cuff size		**4**
Tube of cuff hanging at inner crease of arm		**4**
Start at 180 SBP (Systolic Blood Pressure)		**2**
Re-inflation no sooner than 30-45 sec.		**4**
Re-inflation to first SBP + 20		**1**
Total Incorrect:		

Height

Activity	Correct	Incorrect
Shoes off		**4**
Heels to wall		**2**
Place sticky properly on wall		**2**
Orange triangular ruler on top of head, parallel to floor (fat edge against the wall)		**4**
Place metal tape measure properly and straight for accuracy in measuring height		**4**
Remove sticky from wall when done		**1**
For leg length, ask respondent to locate bony prominence and hold metal tape in place there; keep tape straight		**4**
Total Incorrect:		

Total incorrect (Height): ___________

Max incorrect to pass: 3 (4 or more needs re-certification)

Weight

Activity	Correct	Incorrect
Place scale on firm floor		**4**
Shoes off		**4**
Remove bulky clothes		**2**
Tap red label on scale; wait for “000.0”		**4**
Total Incorrect:		

Max incorrect to pass: 2 (3 or more needs re-certification)

Waist

Activity	Correct	Incorrect
Ask respondent to identify umbilicus (navel) and hold cloth tape in place there		**3**
One layer of clothing		**3**
Tape snug but not tight		**1**
Check that tape is horizontal all around the R		**3**
Ask respondent to take normal breath and exhale, holding breath at end of exhalation		**1**
Record to nearest *centimeter*		**3**
Total incorrect:		

Total incorrect (Waist): ___________

Hip

Max incorrect to pass: 2 (3 or more needs re-certification)

Activity	Correct	Incorrect
Take measurement at respondent's side		**2**
Place cloth tape at level of maximal protrusion of gluteal muscles		**3**
Tape snug but not tight		**1**
Check that it is horizontal all around the respondent		**3**
Move tape up and down to make sure measurement is taken at greatest diameter		**1**
Ask the respondent to take normal breath and exhale, holding breath at end of exhalation		**1**
Record to nearest *centimeter*		**3**
Total incorrect:		

Total incorrect (Hip): ___________
Max incorrect to pass: 2 (3 or more needs re-certification)

References

Afrobarometer Survey. Retrieved June 1, 2016, from http://www.afrobarometer.org/sites/default/files/survey_manuals/ab_r6_survey_manual_en.pdf

Alcser, K. H., & Benson, G. D. (2005). The SHARE train-the-trainer program. In A. Boersch-Supan & H. Juerges (Eds.), *The Survey of Health, Ageing And Retirement in Europe – Methodology* (Ch. 6). Mannheim, Germany: Mannheim Research Institute for the Economics of Aging.

Alcser, K. H., & Benson, G. D. (2008). Training for SHARE Wave 2. In A. Boersch-Supan, A. Brugiavini, H. Juerges, J. Mackenbach, J. Siegrist, & G. Weber (Eds.), *Health, ageing and retirement in Europe (2004-2007)* (Ch. 8.3). Mannheim, Germany: Mannheim Research Institute for the Economics of Aging.

Alcser, K., Benson, G., & Guyer, H. (2011). Pilot Testing the SHARE Quality Profile. In M. Schröder (Ed.), *Retrospective data collection in the Survey of Health, Ageing and Retirement in Europe. SHARELIFE Methodology* (pp. 37-43). Mannheim, Germany: MEA. Retrieved June 14. 2016, from http://www.share-project.org/uploads/tx_sharepublications/FRB-Methodology_2011_Ch.5.pdf

Asian Barometer. Retrieved July 1, 2010, from http://www.asianbarometer.org/intro/program-overview

Axinn, W. G. (1989). Interviewers and data quality in a less developed setting. *Journal of Official Statistics*, *5*(3), 265-280.

Axinn, W. G., Fricke, T. E., & Thornton, A. (1991). The micro-demographic community study approach: Improving survey data by integrating the ethnographic method. *Sociological Methods Research*, *20*(2), 187-217.

Beatty, P. (1995). Understanding the standardized/non-standardized interviewing controversy. *Journal of Official Statistics, 11*(2), 147-160.

Belli, R., Shay, W., & Stafford, F. (2001). Event history calendars and question lists. *Public Opinion Quarterly, 65*(1), 45-74.

Benstead, L. J. (2014). Does interviewer religious dress affect survey responses? Evidence from Morocco. *Politics and Religion*, 1–27.

Bian, Y., & Li, L. (2012). The Chinese General Social Survey (2003–8): Sample designs and data evaluation. *Chinese Sociological Review*, *45*(1, Fall), 70–97.

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

Billiet, J., & Loosveldt, G. (1988). Improvement of the quality of responses to factual survey questions by interviewer training. *Public Opinion Quarterly*, *52*(5), 190-211.

Blaydes, L., & Gillum, R. (2013). Religiosity-of-interviewer effects: Assessing the impact of veiled enumerators on survey response in Egypt. *Politics and Religion*, *6*(3), 459–482.

Börsch-Supan, A., Jürges, H., & Lipps, O. (2003). *SHARE: Building a panel survey on health, aging and retirement in Europe.* Mannheim, Germany: Mannheim Research Institute for the Economics of Aging. Retrieved March 26, 2010, from http://www.mea.mpisoc.mpg.de/uploads/user_mea_discussionpapers/k3pzhnwk1t1zjymk_dp32.pdf

Byass, P., Hounton, S., Ouédraogo, M., Somé, H., Diallo, I., Fottrell, E., … Meda, N. (2008). Direct data capture using hand-held computers in rural Burkina Faso: Experiences, benefits and lessons learnt. *Tropical Medicine & International Health*, *13*(Suppl 1), 25–30.

California Health Interview Survey (CHIS). (2007). *CHIS 2005 Methodology series: Report 2 — Data collection methods*. Los Angeles, CA: UCLA Center for Health Policy Research.

Cannell, C. F., Marquis, K. H., Laurent, A. (1977). *A summary of studies of interviewing methodology.* Vital and Health Statistics: Series 2, Data Evaluation and Methods Research (No. 69. DHEW Publication No. (HRA), 77-1343). Washington, DC: U.S. Government Printing Office. Retrieved December 29, 2009, from http://www.eric.ed.gov/ERICWebPortal/custom/portlets/recordDetails/detailmini.jsp?_nfpb=true&_&ERICExtSearch_SearchValue_0=ED137418&ERICExtSearch_SearchType_0=no&accno=ED137418

Cantave, M., Kreuter, F., & Alldredge, E. (2009). *Effect of pay structure on interviewer productivity.* Poster presented at the 2009 AAPOR meetings.

Cevikaslan, A., Evans, D., Dedeoglu, C., Kalaca, S., & Yazgan, Y. (2013). A cross-sectional survey of repetitive behaviors and restricted interests in a typically developing Turkish child population. *Child Psychiatry and Human Development*, *45*, 472–482.

Christopher, S., McCormick, A. K. H. G., Smith, A., & Christopher, J. C. (2005). Development of an interviewer training manual for a cervical health project on the Apsáalooke reservation. *Health Promotion Practice*, *6*(4), 414-422.

Conrad, F., & Schober, M. F. (1999). *Conversational interviewing and data quality*. Proceedings on the Federal Committee on Statistical Methodological Research Conference. Retrieved February 14, 2008, from http://www.bls.gov/ore/abstract/st/st990250.htm

Davis, D. W. (1997). Nonrandom measurement error and race of interviewer effects among African Americans. *Public Opinion Quarterly, 61*(1), 183-207.

Derrett, S., & Colhoun, S. (2011). Being a quantitative interviewer: Qualitatively exploring interviewers' experiences in a longitudinal cohort study. *BMC Medical Research Methodology*, *2011*, *11*(165).

Doyle, J. K. (2004). Introduction to interviewing techniques. In D. W. Wood (Ed.), *Handbook for IQP advisors and students* (Ch. 11). Prepared for the Interdisciplinary and Global Studies Division). Worcester, MA: Worcester Polytechnic Institute.

Escandon, I. N., Searing, H., Goldberg, R., Duran, R., & Monterrey Arce, J. (2008). The use of PDAs to collect baseline survey data: Lessons learned from a pilot project in Bolivia. *Global Public Health*, 3(1), 93-104.

European Social Survey. (2004). *Field procedures in the European Social Survey: Enhancing response rates*. Retrieved January 24, 2008, from http://www.europeansocialsurvey.org/docs/round4/methods/ESS4_response_enhancement.pdf

European Social Survey (2010). Retrieved June 1, 2016, from http://www.europeansocialsurvey.org/docs/round5/methods/ESS5_response_enhancement_guidelines.pdf

Federal Committee on Statistical Methodology. (1998). Interviewer training. In *Training for the future: Addressing tomorrow's survey tasks* (pp. 50-61). (Report number NTIS PB99-102576). Washington, DC: U.S. Government Printing Office. Retrieved July 20, 2010, from https://fcsm.sites.usa.gov/files/2014/04/spwp27.pdf

Fowler, F. J. Jr., & Mangione, T. W. (1985). *The value of interviewer training and supervision*. Final report to the National Center for Health services Research. Boston, MA: Center for Survey Research.

Fowler, F. J. Jr., & Mangione, T. W. (1990). *Standardized survey interviewing: Minimizing interviewer-related error*. Beverly Hills, CA: Sage Publications.

Galbraith, M. (2003). *Adult learning methods* (3rd ed.). Malabar, FL: Kreiger Publishing Co.

Glewwe, P. (2005). Chapter IV: Overview of the implementation of household surveys. In United Nations Statistical Division, United Nations Department of Economic and Social Affairs (Eds.), Household surveys in developing and transition countries. New York, NY: United Nations. Retrieved March 26, 2010, from http://unstats.un.org/unsd/HHsurveys/pdf/Household_surveys.pdf

Groves, R. (1989). *Survey errors and survey costs*. New York, NY: John Wiley & Sons.

Groves, R. M., Fowler, F. J. Jr., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2009a). *Survey methodology* (2nd ed.). Hoboken, NJ: John Wiley & Sons.

Groves, R. M., & McGonagle, K. A. (2001). A theory-guided interviewer training protocol regarding survey participation. *Journal of Official Statistics*, *17*(2), 249-265.

Guyer, H., Alcser, K., & Kirgis, N. (2008). *Adapting a U.S. based general interviewer and study specific training sessions for multinational studies*. 3MC Conference Proceedings.

Guyer, H., Ofstedal, M-B., Lessof, C., Cox, K., & Juerges, H. (2010). *The benefits and challenges of collecting physical measures and biomarkers in cross-national studies*. Population Association of America.

Hansen, M. H., Hurwitz, W. N., Marks, E. S., & Mauldin, W. P. (1951). Response errors in surveys. *Journal of the American Statistical Association, 46*(254), 147-190.

Hatchett, S., & Schuman, H. (1975). White respondents and race-of-interviewer effects. *Public Opinion Quarterly*, *39*(4), 523-528.

Hursh-César, R. P. (1976). *Third world surveys: Survey research in developing nations*. Delhi: Macmillan Company of India.

Jansen, H. A. F. M., Watts, C., Ellsberg, C., Heise, L., & García-Moreno, C. (2004). Interviewer training in the WHO multi-country study on women's health and domestic violence. *Violence against Women*, *10*(7), 831-849.

Kane, E. W., & Macaulay, L. J. (1993). Interviewer gender and gender attitudes. *Public Opinion Quarterly, 57*(1), 1-28.

Kessler, R. C., Ustun, T. B., & World Health Organization. (2008). *The WHO World Mental Health Surveys: Global perspectives on the epidemiology of mental disorders*. Cambridge; New York: Geneva: Cambridge University Press; published in collaboration with the World Health Organization.

Kish, L. (1962) Studies of interviewer variance for attitudinal variables. *Journal of the American Statistical Association*, *57*(297), 92-115.

Koker, T. (2009). *Choice under pressure: A dual preference model and its application* (Economics Department Working Paper). New Haven, CT.: Yale University. Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1368398

Lavrakas, P. J. (1993). *Telephone survey methods: Sampling, selection, and supervision* (2nd ed.). Thousand Oaks, CA: Sage Publications.

Lee, R. M. (1993). *Doing research on sensitive topics*. London, UK: Sage Publications.

Liamputtong, P. (2010). *Performing qualitative cross-cultural research*. Cambridge, UK: Cambridge University Press.

Living Standard Measurement Study Survey. (1996). *Working paper 126: A manual for planning and implementing the Living Standard Measurement Study Survey.* Retrieved July 10, 2010, from http://www-wds.worldbank.org/external/default/WDSContentServer/WDSP/IB/2000/02/24/000009265_3961219093409/Rendered/PDF/multi_page.pdf

Martin, E. (1996). *Household attachment and survey coverage*. Proceedings of the Section on Survey Research Methods (pp. 526-531). Alexandria, VA: American Statistical Association.

McFall, S., Conolly, A., & Burton, J. (2012). *Collecting biomarkers using trained interviewers: Lessons learned from a pilot study*. Understanding Society Working Paper Series (No. 2012 – 05). Colchester, England: Institute for Social and Economic Research, University of Essex.

Mneimneh, Z., de Jong, J., Cibelli Hibben, K. L., & Moaddel, M. (2015). *Interviewer gender, Islamic hijab, and attitude effects on respondents' politico-religious and religious attitudes in Tunisia.* In Conference Proceedings. Hollywood, Florida.

Nyandieka, L. N., Bowden, A., Wanjau, J., & Fox-Rushby, J. A. (2002). Managing a household survey: A practical example from the KENQOL survey. *Health Policy and Planning*, *17*(2), 207-212.

O'Brien, E. M., Mayer, T. S., Groves, R. M., & O'Neill, G. E. (2002). *Interviewer training to increase survey participation*. Proceedings of the Section on Survey Research Methods. Alexandria, VA: American Statistical Association.

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.

O'Muircheartaigh, C., & Marckward, A. M. (1980). *An assessment of the reliability of World Fertility Study data.* Proceedings of the World Fertility Survey Conference, Vol. 3, 305-379. The Hague, The Netherlands: International Statistical Institute.

Office of Management and Budget. (2006). *Standards and guidelines for statistical surveys*. Washington, DC: Office of Information and Regulatory Affairs, OMB. Retrieved January 10, 2010, from http://www.whitehouse.gov/omb/inforeg/statpolicy/standards_stat_surveys.pdf

Pan, Y., & Lubkemann, S. (2013). *Observing census enumeration of non-English speaking households in the 2010 census: Evaluation report*. Research and Methodology Directorate, Center for Survey Measurement Study Series (Survey Methodology #2013- 02). Washington, DC: U.S. Census Bureau.

Papanek, H. (1979). Research on women by women: Interviewer selection and training in Indonesia. *Studies in Family Planning*, *10*(11/12), 412-415.

Paudel, D., Ahmed, M., Pradhan, A., & Dangol, R. L. (2013). Successful use of tablet personal computers and wireless technologies for the 2011 Nepal Demographic and Health Survey. *Global Health: Science & Practice*, *1*(2):277-284.

Pennell, B-E., Harkness, J., Levenstein, R., & Quaglia, M. (2010). Challenges in cross-national data collection. In J. Harkness, B. Edwards, M. Braun, T. Johnson, L. E. Lyberg, P. Mohler, … T. Smith. (Eds.), *Survey methods in multicultural, multinational, and multiregional contexts*. Hoboken, NJ: John Wiley & Sons.

Pennell, B-E., Harkness, J., & Mohler, P. (2006). *Day 8: Data collection, interviewers, and interviewer training*. SRC Summer Institute Presentation. University of Michigan, Ann Arbor, MI.

Pennell, B-E., Mneimneh, Z., Bowers, A., Chardoul, S., Wells, J. E., Viana, M. C., … Vilagut, G. (2009). Implementation of the World Mental Health survey initiative. In R. C. Kessler & T. B. Üstün (Eds.), *Volume 1: Patterns of mental illness in the WMH surveys.* Cambridge, MA: Cambridge University Press.

Quaglia, M. (2006). *Interviewers and interpreters: Training for a biographical survey in Sub-Saharan Africa*. Paper presented at Proceedings of Q2006 European Conference on Quality in Survey Statistics.

Schaeffer, N. C. (1980). Evaluating race-of-interviewer effects in a national survey. *Sociological Methods & Research*, *8*(4), 400-419.

Schnell, R., & Kreuter, F. (2005). Separating interviewer and sampling-point effects. *Journal of Official Statistics, 21*(3), 389–410.

Schuman, H., & Converse, J. M. (1971). The effects of black and white interviewers on black responses in 1968. *Public Opinion Quarterly*, *35*(1), 44-68.

Shirima, K., Mukasa, O., Schellenberg, J. A., Manzi, F., John, D., Mushi, A., … Schellenberg, D. (2007). The use of personal digital assistants for data entry at the point of collection in a large household survey in southern Tanzania. *Emerging Themes in Epidemiology*, *4*:5. doi: 10.1186/1742-7622-4-5.

Smith, T. W. (2007). Survey non-response procedures in cross-national perspective: The 2005 ISSP non-response survey. *Survey Research Methods, 1*(1), 45–54.

Stycos, J. M. (1952). Interviewer training in another culture. *Public Opinion Quarterly, 16*(2), 236-246.

Suchman, S., & Jordan, B. (1990). Interactional troubles in face to face survey interviews. *Journal of the American Statistical Association, 85*, 232-241.

Sudman, S. (1966). New approaches to control of interviewing costs. *Journal of Marketing Research*, *3*(1), 55-61.

Survey of Health, Ageing and Retirement in Europe. (2010). Retrieved July 1, 2010, from http://www.share-project.org

Ustun, T. B., Chatterji, S., Mechbal, A., & Murray, C. J. L. (2005). Chapter X: Quality assurance in surveys: Standards, guidelines, and procedures. In United Nations Statistical Division, United Nations Department of Economic and Social Affairs (Eds.), *Household surveys in developing and transition countries*. New York, NY: United Nations. Retrieved March 26, 2016, from http://unstats.un.org/unsd/HHsurveys/pdf/Household_surveys.pdf

Vaessen, M., Thiam, M., & Le, T. (2005). Chapter XXII: The demo-graphic and health surveys. In United Nations Statistical Division, United Nations Department of Economic and Social Affairs (Eds.), *Household surveys in developing and transition countries*. New York, NY: United Nations. Retrieved March 26, 2010, from http://unstats.un.org/unsd/HHsurveys/pdf/Household_surveys.pdf

Weisberg, H. (2005). *The total survey error approach*. Chicago, IL: University of Chicago Press.

West, B., & Olson, K. (2009). *How much of interviewer variance is really nonresponse error variance?* Michigan Program in Survey Methodology Brownbag Presentation*.* Ann Arbor, MI: University of Michigan.

Data Collection: General Considerations

Julie de Jong, 2016
(2010 Version: Beth-Ellen Pennell, Rachel Levenstein, and Hyun Jung Lee)

Introduction

Collecting comparable data in the context of multinational, multiregional, and multicultural surveys ("3MC" surveys), is a highly complex task, in which one can expect to encounter a variety of languages and cultural contexts. Even in a single country, the target population may not be linguistically, ethnically or culturally homogenous. Such cultural heterogeneity could manifest itself through a wide variety of dimensions that could impact data collection efforts. For example, languages spoken may not have a standard written form, or respondent literacy rates may be vastly different. The geographic topography may be difficult (e.g., remote islands, deserts, or mountainous regions), and 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).

Countries also vary widely in both their survey research infrastructures and in their laws, norms, values, and customs pertaining to data collection and data access. Certain modes of administration may be inappropriate or not feasible in some situations. In addition, the size and composition of nonresponse will likely vary due to differences in cooperation and ability to contact respondents. Some countries officially prohibit survey research (e.g., North Korea) or severely restrict data collection on some topics.

While a survey conducted in a single country might face one or more of the challenges mentioned above, the probability of encountering multiple hurdles 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. Because of the heterogeneity of target populations in cross-cultural surveys, allowing some flexibility in data collection protocols can reduce costs and error.

In some cases, a coordinating center dictates data collection decisions across all countries involved. The European Social Survey (ESS), for example, mandates the mode in each country, while the International Social Survey Programme

(ISSP) allows a certain amount of flexibility. See Study Design and Organizational Structure for more details.

These guidelines are intended to advise data collection decision-makers as they consider the issues and requirements relevant to different data collection modes and provide extensive recommendations for the practical implementation of data collection in different modes. Because the guidelines and lessons learned vary greatly depending on the specific mode of data collection, we begin with general considerations relevant for data collection in any mode and then provide further guidelines and lessons learned in three subsequent chapters for the main modes of data collection used for 3MC surveys as follows:

Data Collection: General Considerations (these guidelines)
Data Collection: Face-to-Face Surveys
Data Collection: Telephone Surveys
Data Collection: Self-Administered Surveys

For a discussion of the advantages and disadvantages of specific modes, key factors involved in mode choice, and whether to standardize mode across locations, see Study Design and Organizational Structure].

Because difficulties in data collection can be extreme in countries where infrastructure is limited, these guidelines heavily emphasize the challenges of data collection in such contexts.

Guidelines

Goal: To achieve an optimal cross-cultural data collection design by maximizing the amount of information obtained per monetary unit spent within the allotted time, while meeting the specified level of precision and producing comparable results.

1. **Before beginning fieldwork, assess the feasibility of conducting the research in each target country and culture.**

   Rationale

   Local knowledge can be critical to understanding cultural traditions and customs, possible limitations, and the feasibility of the research. Experienced researchers, interviewers, and key stakeholders familiar with the topic or the population under study can help assess concerns and challenges and suggest potential solutions.

Procedural steps

1.1 Assess the appropriateness of (1) the constructs to be studied, and (2) the mode of data collection selected (van de Vijver & Leung, 1997). For detailed information about different data collection modes, see Data Collection: Face-to-Face Surveys, Data Collection: Telephone Surveys, and Data Collection: Self-Administered Surveys.

1.2 Gather information from the coordinating center on major survey design features. These might include the survey topic and questionnaire items, intended mode of administration, instrument technical design, respondent burden (e.g., length of interview, complexity of topic), and proposed methods for dealing with nonresponse.

1.3 Gather information from people who are familiar with data collection in the area and from people who may not be familiar with survey data collection but who are familiar with, represent, or may share characteristics with the population of interest. If possible, conduct focus groups and one-on-one interviews with individuals within the contracted survey organization and others who have previously collected data within the country or location.

1.3.1 Solicit the help of local collaborators or researchers. Local collaborators may have a solid understanding of relevant cultural concerns or challenges or they may be able to help gather information from other local individuals who are more familiar with data collection and the population(s) of interest.

- Provide local collaborators or researchers with a detailed description of the protocol, including the proposed mode of data collection, nonresponse reduction techniques, timing, interviewer training, remuneration, monitoring, and the general framework for data collection.
- Explain and clarify any survey terminologies to ensure common understanding.
- Request feedback on all aspects of the proposed study.
- Arrange to be present (even if by phone or other means of communication) when local collaborators are collecting information from local resources to clarify and probe when needed. However before making a decision to join those meetings, assess whether participating in those meetings might make locals uncomfortable and wary of providing information.

1.3.2 Elicit information from these local human resources and any relevant administrative bodies on:

- Population issues (e.g., local knowledge about the survey, family structure and household listing issues, literacy levels, unwritten languages and cultural norms).
- Logistical issues (e.g., seasonal accessibility, locked dwelling units, secured or dangerous areas, and connectivity issues).
- Issues related to mode choice (see Study Design and Organizational Structure, Data Collection: Face-to-Face Surveys, Data Collection: Telephone Surveys, and Data Collection: Self-Administered Surveys).
- Issues related to interviewers if an interviewer-administered mode is used (e.g., availability of interviewers, background, and safety concerns)
- Human protection issues (e.g., legal and cultural permissions which may be necessary to conduct the study) (see Ethical Considerations).

Lessons learned

1.1 While outside input is often helpful, recognize that negative feedback may, in part, reflect uncertainty rather than concrete obstacles. Such feedback can, however, alert researchers to constraints that require attention. For example, in an early survey of mass media communication behavior in the Middle East, experts predicted that data collection would not be possible in Arab countries because the experts believed the populace would think that the interviewers were agents of the government. The experts also suggested that women could not be hired as interviewers and that it would be impossible to survey previously unsurveyed groups, such as the nomadic Bedouin tribes. The research team, however, was successful in their data collection efforts (Carlson, 1958).

1.2 While a mixed-mode design can reduce the cost of data collection by allowing for increased flexibility to accommodate local contexts, it may also create an additional layer of complexity and, thus, the overall costs for the subsequent harmonization of data by coordinating centers. The Gallup World Poll implements a mixed mode design in which the telephone is used in countries where 80% or more of the target population is covered and face-to-face interviewing is used in countries with lower telephone coverage. The reported costs of telephone surveys are much lower than face-to-face modes (Biemer & Lyberg, 2003), so overall data collection costs are reduced. However, comparability problems due to different modes (phone in one country, face-to-face in another) may exist (Gallup, 2007).

1.3 In a cross-national context, the impact of mode can be confounded with cultural differences. For example, when the International Social Survey Programme (ISSP) began, the required mode was a self-administration. However, low literacy levels in some countries necessitated the use of interviewers. Both response rates and reports from substantive measures differed widely, possibly as a result of differences in mode (Skjåk & Harkness, 2003). Therefore, reported variation between countries on survey estimates may indicate substantive differences or may be a result of mode effects and interviewer effects.

2. **Decide whether the desired information can best be collected by combining qualitative methods with the standardized survey.**

Rationale

A mixed method data collection approach can increase data quality and validity in a number of ways.

First, applying a combination of research methodologies to study the same phenomenon facilitates the validation of data through cross verification, while each method counterbalances the potential limitations of the others (Hulme, 2007). Qualitative and quantitative data collection and analysis methods can be used iteratively to strengthen both approaches. For example, qualitative, less structured interviews may permit a more positive interaction between the interviewer and the respondent increasing the accuracy of the information the respondent provides as well as his or her willingness to provide such information. Qualitative methods can also place the behavior of respondents into a broader context and can improve data coding by revealing unanticipated influences.

Second, mixing qualitative and quantitative methods can address the complexity of sensitive topics or cultural factors more fully than can quantitative methods alone (Bamberger, Rao, & Woolcock, 2010). Finally, it is not necessary to draw a strict dichotomy between qualitative and quantitative approaches; researchers may remain open to switching between the two so-called paradigms within the course of a study (van de Vijver & Chasiotis, 2010).

Procedural steps

Choose data collection methods to fit the aim of the research question (Axinn & Pearce, 2006).

2.1 Consider combining less structured interviewing, field notes, observation, historical materials, or event history calendars with the standardized survey (Axinn & Pearce, 2006).
  2.1.1 In the social sciences, the term "methodological triangulation" is often used to indicate that more than two methods are used in a study to double (or triple) check results (for further information on methodological triangulation and integrating qualitative and quantitative methods in data collection, see Further Reading).
  2.1.2 Triangulation can also widen and deepen one's understanding of the phenomenon being studied.

2.2 Ethnosurveys offer an approach that combines survey and ethnographic data collection and allows each method to inform the other throughout the study. Equal weight or priority is given to each method. Quantitative data is collected in a manner that falls between a highly structured questionnaire and a guided ethnographic conversation, which is helpful in contexts where rigid structure may be inappropriate but where some standardization is needed for comparison purposes. See Massey (1987) on the theory and practice of ethnosurveys.
  2.2.1 Determine whether your study is retrospective, prospective, or both. Calendar methods are more efficient for retrospective studies while longitudinal designs are more efficient for prospective studies (Axinn & Pearce, 2006; Freedman, Thornton, Camburn, Alwin, & Young-Demarco, 1988).
  2.2.2 Remember that traditional qualitative methods can be more expensive and time consuming than a standardized survey (Massey, 1987; Morse & Niehaus, 2009).

Lessons learned

3MC projects have successfully combined qualitative and quantitative methods of data collection in many different ways.

2.1 The Tamang Family Research Project, conducted in Nepal in 1987 to 1988, studied two communities to see how family structure influenced fertility decisions. By adding less-structured ethnographic interviews to the highly structured survey, the investigators discovered that a previously unknown factor, the Small Farmers Development Program (SFDP), had a significant influence on fertility decisions (Axinn, Fricke, & Thornton, 1991; Fricke, 2005).

2.2 The event history calendar method is easily adaptable to fit cultural needs. Some tribes in the Chitwan Valley Family Study (CVFS), conducted in Nepal, had no conception of time measurement.

Researchers successfully used local and national events as landmarks to help respondents accurately recall their life course history (Axinn & Pearce, 2006; Axinn, Pearce, & Ghimire, 1999; Belli, 1998).

2.3 To look at trends in household poverty, Krishna (2007) followed seven steps in a Stages of Progress method:
   2.3.1 Assembled a "representative community group" (p. 2);
   2.3.2 Presented objectives;
   2.3.3 Collectively described the construct;
   2.3.4 Used current definitions of households as the unit of analysis, inquired about the status of the construct at present and 25 years ago;
   2.3.5 Assigned households to categories;
   2.3.6 Asked about reasons for descent into poverty among a sample of households within each poverty category (relative to previous and current poverty status); and
   2.3.7 Interviewed household members.

2.4 Broom (2005) believes that health research is best conducted using in-depth interviews, rather than being driven by the questionnaire and preconceived notions. He argues that qualitative methods allow for a more thorough analysis and holistic understanding of the patients' decision-making processes.

2.5 Hulme (2007) describes the use of mixed methods in the context of country case studies.

2.6 The Demographic and Health Surveys (DHS) program conducts research in approximately 75 developing countries across the world (DHS, 2010). The main objectives of the DHS program are "(1) to provide decision makers in the survey countries with data and analyses useful for informed policy choices, (2) to expand the international population and health database, (3) to advance survey methodology, and (4) to develop in participating countries the skills and resources necessary to conduct demographic and health surveys." Phase II of the DHS introduced a calendar at the end of one of the core questionnaires to clarify dates relating to fertility, contraceptive, postpartum, marriage, migration, and employment history. The researchers found that the calendar provided gains in the quantity and quality of data collected, as well as increasing their analytical potential.

2.7 Hargreaves et al. (2006) used mixed methods to assess the poverty rankings of individual households in eight villages in rural South Africa. The study aimed to identify the number of poor households and to assess their level of poverty. Working with researchers, community residents drew a map of their village and located each household within its boundaries. Researchers then asked smaller groups of residents to rank pairs of randomly selected households, asking which household in the pair was poorer and which was better-off. Finally, the responses were coded. The authors found strong agreement between the subjects' coded perceptions of poverty and a household wealth index generated using statistical methods. Howe and McKay used similar methods to study chronic poverty in Rwanda (Howe & McKay, 2007).

2.8 Keller (2007) studied the influence of parents and other socialization factors on human development. Working with young infants and their families in Asia, Latin America, Europe, North America, and Africa, she successfully combined qualitative analyses of interviews and participant observation with quantitative analyses of questionnaires and videotape footage.

2.9 Implementing qualitative methods or ethnosurveys helped University of Chicago researcher Douglas Massey gain greater insight into the reasons behind migration in the U.S. (Massey, 1987).

2.10 By combining data obtained from both statistical and qualitative analyses, Sampson and Laub were able to more accurately explain and identify changes and consistencies in criminological behavior over a convict's life (Sampson, & Laub, 1998).

2.11 Bamberger, Rao, and Woolcock (Bamberger et al., 2010) suggest returning briefly to the field when writing the quantitative report for more descriptive information or to explore inconsistencies in the data.

3. Reduce the potential for nonresponse bias as much as possible.

Rationale

Optimal data collections maximize response rates and thereby decrease the potential for nonresponse bias. Nonresponse occurs when survey measures are not obtained from sampled persons, thereby increasing the nonresponse rate. Nonresponse bias occurs when the people who are non-respondents differ from respondents systematically. Although the response rate alone does not predict nonresponse bias (Groves, 2006) a low response rate can be a predictor of the potential for nonresponse bias.

Furthermore, response rates have been dropping differentially across countries due to noncontact and, increasingly, reluctance to participate (de Leeuw & de Heer, 2002).

The coordination of a cross-cultural survey can be centralized or decentralized, with a corresponding focus on either input or output harmonization, as discussed in Study Design and Organizational Structure. These differences in coordination can impact response rates and response bias differentially. For example, in a study using the output harmonization model, where each country uses their own methods and strategies to maximize response rate, nonresponse rates can be calculated and response bias can occur in different ways, whereas in a study using input harmonization, study countries will be limited in adaptation to local contexts, which in turn also impacts response rates and response bias See Wagner and Stoop (2017) for a more in-depth discussion on nonresponse and nonresponse bias in a cross-national study.

For further discussion of nonresponse bias within the survey quality framework, see Survey Quality.

Procedural steps

3.1 Consider the following steps at the community level to reduce non-response before beginning data collection.
  - 3.1.1 Depending upon cultural norms, gain the support of any "gatekeepers" (e.g., community leaders or elders) before attempting to reach individual households.
  - 3.1.2 Make all efforts to raise awareness about the need for high quality surveys and thus the need for people to take part.
  - 3.1.3. Publicize the survey locally to raise awareness and encourage cooperation.
    - If most of the population is literate, consider displaying colorful, attractive leaflets on local bulletin boards and in other public spaces.
    - Use word-of-mouth channels or local dignitaries (doctors, teachers) as appropriate.

3.2 Send pre-notification letters to sampled households if feasible.
  - 3.2.1 The letter should (1) explain the purpose of the survey, (2) establish the legitimacy of the survey organization and the interviewer, (3) assure confidentiality of answers, (4) notify the household that participation is voluntary, (5) include or announce gifts or incentives and provide information about them, and (6) provide contact information for the organization (see Appendix A for an example of pre-notification letters).
  - 3.2.2 There should be a short timespan between the arrival of the letter and first contact by the interviewer; a time span of several days is ideal. If there is a long delay between the letter

and contact, consider sending a second letter before attempting contact.

3.2.3 Personalize the advance letter with the individual name if possible and appropriate.

3.2.4 Be aware that survey sponsorship may affect both response rates and the accuracy of the actual data. For example, some respondents may fear repercussions if they do not respond to a survey sponsored by a government agency. While this fear may dramatically increase response rates, the quality of the data may be dubious; respondents may feel that their responses are not genuinely confidential if the survey is sponsored by a government agency, and they may not respond openly. In addition, ethical issues arise in such situations (see Ethical Considerations).

3.3 Nonresponse can be assessed and reduced with effective sample management and interviewer management monitoring systems and associated paradata. For an in-depth discussion on the use of responsive designs and paradata to assess nonresponse and nonresponse bias, see Paradata and Other Auxiliary Data.

3.3.1 Study structure and data collection modes may specify what sample management systems are used. In cross-cultural surveys with strong centralized control, a single sample management system may be specified in the contract with local survey organizations.

3.3.2 A good sample management system facilitates evaluating interviewer workload and performance.

3.3.3 Monitor response rates continuously, and produce reports of daily response rates in order to identify data collection procedures that are more or less successful at increasing participation.

3.4 Structure the field staff to aid them in working the sample efficiently and effectively.

3.4.1 Give supervisors the responsibility of assigning sample elements to interviewers and reassigning them when necessary.

3.4.2 Do not allow interviewers to switch sample among themselves without the explicit approval of the supervisor.

3.4.3 Ensure that sample elements are assigned in a way that minimizes travel efforts and costs.

3.4.4 Decide whether interviewers will work alone, in pairs, or in traveling teams (see above and Interviewer Recruitment, Selection, and Training).

3.4.5 Decide whether interviewers and respondents should be matched on some characteristic(s), such as gender or ethnicity, in order to increase respondent comfort and increase respondent cooperation. If the respondents' characteristics are unknown prior to data collection, develop procedures to make on-the-spot matching possible. For example, to facilitate gender matching, send interviewers into the field in male-female pairs.

3.5 Specify the minimum, the maximum number, and the timing of attempts to contact before the final disposition code is assigned to increase efficiency.

3.5.1 Interviewers should attempt to contact respondents at different blocks of time across the week to increase the probability of reaching the respondent at home.

- The times of day when persons are most likely to be at home vary by culture, location and context. For example, working respondents in the United States are more likely to be reached on evenings and weekends (Groves & Couper, 1998).
- Alternatively, specify the minimum number of times that attempts must be made during daytime hours, during evening hours, and during the weekend (see Kulka & Weeks (1988) for details on call scheduling). Incorporate culture-specific information about likely at-home patterns, such as normal workdays, normal work hours, and holidays. Beware of religious and other cultural norms that restrict interviewing at certain times.

3.6 If appropriate, offer an incentive for participation (Singer, 2002).

3.6.1 Adapt the type and amount of the incentive to local customs. Make yourself familiar with country-specific research on incentives.

3.6.2 According to US- and Canada-based research:

- Present the incentive as a "token of appreciation" for participating in the survey, not as payment for the response.
- Make the token reasonable; it should not be so large that it might raise suspicion about the researcher's or organization's motives or be somehow coercive. It should be generally proportionate to the respondent burden.
- Ideally, provide the incentive prior to the interview. Incentives promised upon the completion of the interview also increase participation, but to a lesser degree (Berk, Mathiowetz, Ward, & White, 1987; Singer, Hoewyk, Gebler, Raghunathan, & McGonagle, 1999).

3.6.3 Document the use of incentives, including 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.

- According to the existing literature, unconditional prepaid incentives seem to be more effective than conditional incentives paid upon completion of the interview (Koch, Blom, Stoop, & Kappelhof 2009). Thus, eliciting feelings of obligation from the unconditional incentive is more effective than rewarding participation.
- It may be necessary to monitor the extent to which monetary incentives disproportionately encourage the participation of people with low incomes compared to those with high incomes and thereby have an effect on nonresponse bias. If poorer people are usually underrepresented in the achieved sample, monetary incentives might reduce nonresponse bias. If poorer people are already overrepresented, incentives might even increase the nonresponse bias.
- Offering a choice of different types of incentives might attract people from a more diverse background. This can help to reduce an existing nonresponse bias and counteract the potentially selective effect of offering one specific incentive.
- For financial incentives, interviewers may be asked to record that an incentive was given to a respondent; similarly, the respondent may need to sign to indicate receipt.
- In deciding whether to use an incentive, weigh the relative time and cost advantages of using an incentive versus not. Incentives may mean less interviewer time in persuading respondents to participate or less time in refusal conversions. The reduction in interviewer time – and thus costs – must be weighed against the cost of providing incentives.
- See Ethical Considerations for further discussion on the appropriate use of incentives.

3.7 In using a face-to-face or telephone mode, train interviewers to use culturally appropriate reluctance aversion techniques (see Interviewer Recruitment, Selection, and Training).

3.7.1 Social or psychological factors (e.g., reciprocation, consistency, social validation, authority, scarcity, liking) affect respondents' decision in survey participation (Cialdini, 1988). Minimally, train interviewers how to answer anticipated respondent concerns (Groves, Cialdini, & Couper, 1992).

3.7.2 Be aware that local customs and legal limitations may prohibit any attempt to recontact someone who has declined to participate in the survey. In these cases, using appropriate reluctance aversion techniques becomes especially important.

3.7.3 Make sure that supervisors monitor interviewers closely on respondent reluctance issues.

3.8 If using a face-to-face or telephone mode, consider assigning supervisors or more experienced interviewers to cases where interviewers have been unsuccessful making contact or achieving cooperation.

3.9 Consider switching modes to increase contact and cooperation.

3.9.1 Some studies in the United States employ a mixed mode design in which the least expensive mode is used initially, after which time progressively more expensive modes are implemented in order to reduce nonresponse.

3.9.2 Different modes may produce different survey estimates. These mode-specific differences in measurement might be acceptable to the investigator if nonresponse is sufficiently reduced.

3.9.3 If more than one mode is expected to be used and budget permits, examine possible mode effects prior to the start of data collection.

- Test for mode effects by administering key questions or questionnaire sections to a randomly split sample of respondents similar to the targeted population (e.g., asking the questions on the telephone for one group and in-person for another).
- If it is not possible to test for potential mode effects beforehand, check for differences in responses at the end of data collection.
- Ascertain whether respondents surveyed in each mode produce similar response distributions on key variables before combining their responses for analysis.

3.10 Have interviewers complete a contact attempt record each time they attempt contact, whether or not the attempt is successful (see Appendix B for an example of a contact attempt record).

3.10.1 Use disposition codes to describe the outcome of each contact attempt.

3.10.2 Distinguish (1) completed interviews with eligible persons, (2) non-interviews (eligible persons), (3) non-interviews (unknown if eligible persons), and (4) non-interviews (ineligible persons).

3.11 Assign a final disposition code to each sample element in the gross sample at the end of data collection; include any new sample elements that may be created or generated during data collection (e.g., for additional family members or through half open intervals).

3.11.1 Provide a clear explanation and training to interviewers before they are allowed to assign final disposition codes.

3.11.2 Take into account that, in some survey organizations, only supervisors can assign final disposition codes.

3.11.3 See Appendices D - G for a description of disposition codes and templates for calculating response rates from the American Association for Public Opinion Research (AAPOR).

3.11.4 See also AAPOR's *Standard Definitions* publication (American Association for Public Opinion Research, 2016), which also provides definitions for final sample disposition codes and formulas for calculating response, refusal, and other rates. See also AAPOR's Response Rate Calculator (available for download at: http://www.aapor.org/Standards-Ethics/Standard-Definitions-(1).aspx?utm_source=AAPOR-Informz&utm_medium=email&utm_campaign=default).

3.11.5 Note that the list of disposition codes may need to be modified for the local situation and additional codes may need to be defined to account for local conditions.

3.12 Minimize the effects of nonresponse bias on analyses as much as possible.

3.12.1 Nonresponse bias is a function of both the response rate and the difference between respondents and nonrespondents on a particular statistic (Groves & Couper, 1998). Because nonresponse bias is statistic-specific, response rates alone do not indicate nonresponse bias. Therefore, estimate the effect of nonresponse bias on key survey estimates, if possible (see Guideline 7 below).

3.12.2 If possible, use weighting and imputation (Groves, Dillman, Eltinge, & Little, 2002) (see Data Processing and Statistical Adjustment).

Lessons learned

3.1 Differences in response rates cross-nationally can be due to many factors, including differing judgments of interviewers and other local survey staff about the efficacy and subsequent application of particular survey research techniques and protocols. A review of response rates from the 1995 round of the International Social Survey Programme (ISSP) found significant differences in response rates, with at least some of the difference likely attributable to mode (face-to-face vs. mail). Even for countries with roughly comparable response rates, sources of nonresponse differed, with noncontact contributing substantially to nonresponse in Japan, and refusal contributing to nonresponse in Russia (Couper & de Leeuw, 2003).

3.2 Response rates are not necessarily good indicators of nonresponse bias, but nevertheless tend to be used as a proxy for bias. In a health study of the elderly in Scotland, healthy individuals were more likely to participate than unhealthy individuals. Because of this difference between the respondents and nonrespondents, the estimate of health was biased even though response rates reached 82% overall (Cohen & Duffy, 2002).

3.3 While the literature has clearly established the positive effects of prepaid and cash incentives upon response in minority countries (Berk et al., 1987; Singer et al., 1999), it is possible that incentives may affect the propensity to respond differently among a population with high rates of poverty. For example, offering a choice of incentives may be more effective at increasing response rates than simply offering a prepaid incentive. Furthermore, in areas with rampant inflation, the value of cash incentives may decrease dramatically within a short period of time.

3.4 The same incentive may affect response rates differently across countries or cultures. In the German General Social Survey (ALLBUS), the same incentive (€10) was offered to all respondents. The authors examined cooperation rates for Moroccan and Turkish immigrants. The authors found that the incentive affected cooperation differently by ethnicity and gender: cooperation rates increased as a result of the incentive for Moroccan women, but did not increase for Moroccan men, Turkish men, or Turkish women (van den Brakel, Vis-Visschers, & Schmeets, 2006).

3.5 The mechanism of incentive efficacy will differ across mode. In telephone surveys, incentives are often sent to the respondent in an advance letter prior to contact, to encourage cooperation. In mail surveys, the incentive may be sent either in advance or along with the mailed questionnaire. And, in face-to-face interviews, the respondent generally receives the incentive at the conclusion of the interview, meaning that the actual transfer of the incentive, and therefore its effect on response rate, can differ across mode, leading to further differentiation in response rates cross-nationally if different countries use different modes in a cross-national survey.

3.6 Use caution when choosing to give monetary awards to study participants. Keller studied the influence of parents and other socialization factors on human development in Asia, Latin America, Europe, North America, and Africa. Respondents received a cash incentive. Keller experienced some hostility from families that were not selected for the study (and, thus, not given any monetary rewards) because they did have young children (Keller, 2007).

3.7 Some studies vary incentive use within a country; for example, offering incentives only to respondents in urban areas, where response rates are typically lower; or offering incentives only in cases of refusal, in an attempt to gain cooperation. If considering this approach, be aware of any concerns that might arise from ethics review boards.

3.8 Countries have different incentive norms.
  3.8.1 For example, in a recent study conducted in Nepal and the United States, respondents in Nepal were highly cooperative and were offered no financial incentive. In the U.S., however, potential respondents were not as cooperative or easy to contact, and incentives were required (Axinn, Chardoul, Ghimire, Gatny, & Barber, 2008).
  3.8.2 Some 3MC surveys (e.g., the European Social Survey and the Living Standard Measurement Study Survey (Living Standard Measurement Study Survey, 1996) allow each participating country to decide whether or not to offer incentives.
  3.8.3 If incentives are offered, the type may vary from one country to another. For example, the Survey of Health, Ageing and Retirement in Europe (SHARE) offers various incentives, depending on the country's culture. Incentives for the World Mental Health Survey (Keller, 2007) vary across participating countries, including but not limited to, cash (in the Ukraine and United States), an alarm clock (in Columbia), and a bath towel (in Nigeria); no respondent incentives are offered in Mexico, South Africa, Belgium, Germany, Israel, Japan, or China. In the Netherlands, flowers are a customary gift to the hostess when visiting for dinner, and therefore flowers are an effective incentive in the Netherlands.

3.9 Similarly, many cross-cultural surveys (e.g., the European Social Survey, the Living Standard Measurement Study

Survey (Living Standard Measurement Study Survey, 1996), and the World Mental Health Survey (Kessler, Ustun, & World Health Organization, 2008)) allow participating countries to vary in their use of advance letters and follow-up letters. In the Survey of Health, Ageing and Retirement in Europe (SHARE), advance letters are mailed to each household in the gross sample and follow-up letters are used with reluctant respondents.

3.10 In an experimental design in the U.S., researchers investigated the use of a novel incentive they termed “reciprocity by proxy”, wherein respondents were invited to participate in a program with the promise that their participation would result in a gift to a third party, such as a charity. Researchers found that reciprocity by proxy increased participation more than either incentive by proxy or no incentive. However, researchers caution that this approach can backfire if the target audience does not support the beneficiary of the gift (Goldstein, Griskevicius, & Cialdini, 2012). To mitigate this risk, researchers can offer to make a contribution to a charity of the respondent’s choosing.

3.11 An effective sample management system can clarify the causes of nonresponse. When the Amenities and Services Utilization Survey (AVO) was conducted in the Netherlands in 1995, interviewers were not asked to record detailed disposition codes for each call. As a result, refusals could not be distinguished from noncontacts. When the study was repeated in 1999, detailed disposition codes were collected. Researchers were then able to see that, after three unsuccessful contact attempts, refusal was the more probable explanation (Stoop, 2005).

3.12 Not all survey organizations will be familiar with sample management practices. Allow some time in training for interviewers to become familiar with the sample management system (see Interviewer Recruitment, Selection, and Training) and check completed forms.

4. Time data collection activities appropriately.

Rationale

A specific survey estimate of interest may determine the timing of data collection activities; for example, a survey about voting behavior will necessarily be timed to occur around an election. Data collection activities may be hampered by inappropriate timing. Face-to-face data collection, for example, may be impossible during a monsoon season, an earthquake or a regional conflict.

The guideline assumes that a specific start time and end time to data collection exists; this guideline does not address issues in continuous data collection.

Procedural steps

4.1 Based upon feasibility studies (see Guideline 1 above), evaluate environmental, political, and cultural considerations which might affect the timing of data collection. These could include:
  - 4.1.1 Extreme weather patterns or natural disasters.
  - 4.1.2 War, civil war, military rule, militia rule, or the possibility of hostage taking.
  - 4.1.3 Religious and secular holidays or migratory patterns of nomadic people. For example, Independence days (e.g., Bastille Day in France), New Year's Day in China, summer Christmas holiday in Australia and New Zealand, and vacations in July and August in Europe would not be a good time.

4.2 Establish a specific start and end date for data collection.
  - 4.2.1 Keep a concurrent fielding period across countries. This would guarantee the cross-national comparability. For example, the ESS requires interviewers across participating countries in Europe to collect data within a four-month period from September to December of the survey year (Koch et al., 2009).
  - 4.2.2 If the 3MC project includes countries located in both the northern and southern hemispheres, where summer and winter are in opposition, consider what field period is most feasible for all countries.
  - 4.2.3 Because unexpected events can interfere with data collection activities, remain somewhat flexible to allow for unexpected events. Include details about any deviations from the anticipated schedule in the study documentation.

Lessons learned

4.1 Coordination of data collection activities across countries or cultures can be difficult or even impossible. The Afrobarometer measures public opinion in a subset of sub-Saharan African countries. The coordinators for the Afrobarometer note that data collection is especially difficult during national election or referendum campaigns, rainy seasons, times of famine, and national or religious holidays. Since such events vary across countries and cultures, fieldwork activities are spread over a full year (Afrobarometer Survey, 2010).

4.2 Timing of data collection activities may be related to the topic of the survey or statistics of interest. The Comparative Study of Election Systems (CSES), for example, studies elections around the world and therefore must time data collection activities according to local election cycles (Howell, 2010).

4.3 The response rate for the Asian Barometer survey in Japan in 2003 was 71%. In 2007, the response rate dropped to 34.3%. One possible reason for the sharp drop in response rates in 2007 is that, in 2006, the law no longer allowed commercial surveys to use voter lists or resident registries. As a result, many people mistakenly believed that the new regulation also applied to academic research (Ikeda, Yamada, Taniguchi, Nishizawa, & Steel, 2007).

4.4 Data collection in Germany for the first European Social Survey had to be delayed due to general elections held in that autumn.

4.5 In some settings, electrical availability is dictated by the calendar and should be evaluated prior to data collection. For example, Nepal relies primarily on hydropower, and so electricity shortages increase significantly in most areas of the country during the dry season between February and April, with some areas without electricity for more than 14 hours per day. Recharging equipment in these sorts of environments can be a major impediment (Paudel, Ahmed, Pradhan, & Dangol, 2013).

5. Institute and follow appropriate quality control measures.

Rationale

If errors are caught early, they can be corrected while the study is still in the field. Improvement made during data collection may introduce some measure of inconsistency in the data, however. This trade-off should be considered before any action is taken (Groves, 2006). See also Survey Quality for a discussion of the quality control framework and Paradata and Other Auxiliary Data for a detailed discussion on using paradata in quality control and survey error reduction.

Procedural steps

5.1 Evaluate the effectiveness of data collection protocols regularly. Include:

5.1.1 Sample management systems.
5.1.2 Contact protocols.
5.1.3 Reluctance aversion protocols.

5.2 With real-time or daily data transmission, quality control routines and error detection can be implemented more efficiently
5.2.1 The use of technology for data collection allows for collecting and analyzing paradata (such as keystrokes and time-stamps, GPS coordinates) for monitoring interviewer behavior (if an interviewer-administered mode is used). This allows for early detection of interviewer deviation from interviewing protocol, for early intervention, and better data quality. [See Lessons Learned below as well as Paradata and Other Auxiliary Data]. Moreover, post-survey processing time is greatly reduced.
5.2.2 If an interviewer-administered mode is used, observe the interviewers throughout data collection (Lavrakas,1993); monitor them more frequently early in the study, less frequently as the study continues.

5.3 If an interviewer-administered mode is used, review a random sample of coversheets on an ongoing basis to ensure that correct eligibility and respondent selection procedures are being followed.

5.4 If an interviewer-administered mode is used, provide interviewers with feedback, both individually and as a group (Couper, Holland, & Groves, 1992; Lavrakas, 1993).
5.4.1 Provide immediate, individual feedback if there has been a critical error.
5.4.2 Provide routine, individual feedback for self-improvement.
5.4.3 Offer group feedback to focus efforts on improving the process.
5.4.4 Continually evaluate the following with respect to interviewers (Biemer & Lyberg, 2003):
- Knowledge of the study objectives.
- Administration of the survey introduction.
- Administration of household enumeration and respondent selection procedures.
- Reluctance aversion efforts.
- Contact efforts.
- Rapport with the respondent (e.g., having a professional, confident manner).
- Standardized interviewing techniques (e.g., reading questions as worded, probing, and clarifying).
- Data entry procedures.
- Administrative tasks (e.g., submitting timesheets in a timely fashion).

- Ability to meet production goals and maintain productivity.
- Administration of specialized study-specific procedures (e.g., procedures for taking physical measurements and administering tests of physical performance or cognitive ability).

5.5 Whenever possible, recontact or reinterview approximately 10-15% of each interviewer's completed cases, selected at random (American Association for Public Opinion Research, 2003; Office of Management and Budget, 2006).

5.5.1 If recontacting the respondent, verify that the interview took place, inquire if interviewer acted professionally, and ask factual questions (e.g., mode of data collection, interview length, incentive, household composition, and key survey topics (American Association for Public Opinion Research, 2003)).

5.5.2 If reinterviewing the respondent, ask a sample of factual questions that do not have heavily skewed response distributions, were not skipped by many respondents, are scattered throughout the questionnaire, and have answers which are unlikely to have changed between the time of the interview and the verification check (Forsman & Schreiner, 1991; United Nations, 2005).

5.5.3 Conduct reinterviews within a time period that is not so long that respondents will have forgotten about the survey or so short that respondents will remember all the details of the survey (Forsman & Schreiner, 1991).

5.5.4 Make sure recontacts and reinterviews are made with the original respondent and that questions refer to the same time period as was asked about in the original interview (Forsman & Schreiner, 1991).

5.5.5 In some countries, it is not possible to perform recontacts or reinterviews due to laws and/or local customs. Document such instances.

5.6 If feasible, audio record face-to-face interviews for review.

5.6.1 Determine whether cultural norms permit taping.

5.6.2 Inform respondents that they may be recorded for quality purposes and allow respondents to refuse to be recorded.

5.6.3 Store any tapes safely and securely (see Ethical Considerations).

5.7 Identify potential interviewer falsification.

5.7.1 Implement silent monitoring in centralized facilities, use audio-recordings and recontacts in field studies, and analyze outliers in the data to detect falsification (American Association for Public Opinion Research, 2003).

5.7.2 Check responses to stem questions for each interviewer. Questions that have a stem-branch structure—in which specific responses to "stem" questions require the interviewer to ask a number of "branch" questions—can be at increased risk for falsification. If a particular interviewer has recorded responses to stem questions that consistently preclude the interviewer from asking the branch questions, the interviewer may be falsifying data.

5.7.3 Examine paradata, such as keystroke data and time stamps, by interviewer to identify potential falsification.

5.7.4 Examine survey data for any duplicate cases, which can indicate falsification as well as data processing error.

5.7.5 If falsification of data is suspected, contact the respondents involved over the telephone (Forsman & Schreiner, 1991) If respondents cannot be reached via telephone, send out a brief mail questionnaire with a prepaid return envelope (Biemer & Lyberg, 2003).

5.7.6 If falsification of data is suspected, investigate the interviewer's other work and remove the interviewer from all data collection activities until the issues have been resolved (American Association for Public Opinion Research, 2003).

5.7.7 If irregularities or falsified data are discovered, redo the interviewer's cases and delete all of his or her recorded data (American Association for Public Opinion Research, 2003; Biemer & Lyberg, 2003).

5.8 For approximately 5% of each interviewer's finalized non-interviews, perform random checks with households to verify that ineligibility, refusal, or other status was correctly assigned. Checks may be done by telephone, in person, or by mail, as needed.

5.9 If physical measurements are being taken:

5.9.1 Periodically retest the interviewers on the use of these instruments.

5.9.2 Select equipment that can withstand the local conditions (heat, cold, altitude, etc.).

5.9.3 Document the technical specifications of the equipment chosen.

5.9.4. Re-calibrate equipment as needed throughout data collection

5.10 If the survey is being conducted in a centralized telephone facility, follow established monitoring procedures (Couper et al., 1992)

5.10.1 Monitor in relatively short (e.g., one-hour) shifts; this is cost-effective and reduces supervisor fatigue.
5.10.2 Use probability sampling to ensure that the number of interviews monitored is proportional to the number of interviewers working each hour (see Sample Design).
5.10.3 Monitor new interviewers at a higher rate than experienced interviewers.
5.10.4 Select from eligible cases in which the phone is still ringing so that the supervisor is not forced to wait for new interviews to begin in order to start monitoring.

5.11 Monitor quality indicators consistently throughout the field period; use an electronic system or note them in a daily log book (United Nations, 2005). Include the following:
5.11.1 Distributions of key variables.
5.11.2 Hours per interview per interviewer, for the study as a whole, and by respondent groups of interest.
5.11.3 Number of respondents approached, interviews completed, incomplete interviews, and contact attempts.
5.11.4 Response, refusal, and non-contact rates (United Nations, 2005) (see Data Processing and Statistical Adjustment).
5.11.5 Outcomes of all contacts and review of disposition code assignment.

5 .12 Create statistical process control charts (SPCs) to provide timely information on key aspects of the data collection process (Mudryk, Burgess, & Xiao, 1996).
5.12.1 Use the charts to detect observations that are not within pre-determined limits (often between one and three standard deviations of the mean).
- A common use of SPCs in survey organizations is to assess nonresponse reduction methods over the field period. Using these charts, the impact of interviewer effort on response rates can be easily assessed (see case studies in Survey Quality for additional discussion of SPCs).

5.12.2 Give extreme observations additional attention and try to determine the root cause.
5.12.3 Refer to the charts when deciding whether to release additional sample elements for interviewers to attempt to contact, further monitor interviewers, and offer additional training sessions.

5.13 Set contact limitations, determining:
5.13.1 The point at which additional attempts to contact a sample element are inefficient.

5.13.2 Whether respondents cooperating after a certain number of contact attempts are significantly different from others on key indicators.

Lessons learned

5.1 Process and progress indicators are often interdependent. Therefore, improving one process or progress indicator may negatively affect another, particularly in the context of attempts to achieve cross-national comparability. For example, the pursuit of higher response rates can actually increase nonresponse bias if the techniques used to obtain the higher response rates are more acceptable and effective in some cultures than in others (Groves, 2006; Harkness, 1999).

5.2 In Round 4 of the Afrobarometer Survey (2010), teams of four interviewers travel together to the field under the leadership of a field supervisor who has at least an undergraduate degree and experience in collecting data and managing field work teams or no degree but extensive experience. It is the supervisor's job to ensure quality control of survey returns on a daily basis. Interviewers are monitored at all stages and debriefed daily immediately after interviews. Completed questionnaires are checked for missing data and inconsistencies. Each field supervisor maintains a daily written log of observations on sampling and interviewing conditions and political and economic features of the area and makes daily telephone report to headquarters. A fieldwork debriefing is held after all returns have been submitted. Sampling back-checks are routinely conducted to ensure that the respondent selection is correctly done. The field supervisor also verifies basic information (e.g., respondent age and level of formal education.

5.3 The Asian Barometer Survey (2010) required all interview teams to travel together under the supervision of a field supervisor and to have a debriefing meeting each evening. Supervisors randomly checked with respondents to make sure the interviews were done properly.

5.4 In Round 5 of the ESS (ESS, 2010), quality control back-checks were performed for at least 10% of respondents and 5% of the non-respondents either in person, by telephone, or by mail. For the respondents, a short interview was conducted to confirm the interview, whether showcards were used, the approximate length of the interview, etc.

5.5 In the Living Standard Measurement Study (LSMS, 1996), each field supervisor oversees two interviewers. Each week the field supervisor observes and evaluates one interview per interviewer and documents the process for submission to the national office. Data collection is broken into two rounds; the first half of the questionnaire is completed in round one and then checked for accuracy before the second half of questionnaire is completed in round two. After the second round, only data entry errors are corrected. Check-up interviews are routinely performed in 15% to 25% of the households.

5.6 The Survey of Health, Aging and Retirement in Europe (SHARE) requires all survey agencies to use an electronic sample management system (SMS). All but three participating countries (France, the Netherlands, and Switzerland) use a "Case Management System" (CMS), developed by CentERdata. This system monitors the survey progress in real-time, including screening for eligible respondents, recording contact attempts, ensuring the correct implementation of contact and follow-up strategies, and managing refusal conversion strategies. Bi-weekly reports are generated for the coordinating team.

5.7 The recommended supervisor-to-interviewer ratio in the World Mental Health Survey is 1 for every 8 to 10 experienced interviewers, with those countries using a pencil-and-paper mode having a higher ratio than those conducting computer-assisted surveys. Supervision consists of direct observation and/or audio recording of part or all of the interview for 5% to 10% of each interviewer's work. Supervisors randomly select 10% of interviewed households, confirm the household listings and selection procedure, and repeat some of the questions. Open-ended responses and other quality control checks are reviewed on a daily basis by supervisors, and interviewers recontact respondents to obtain missing data (Kessler et al., 2004; Kessler et al., 2008).

5.8 Data falsification can be difficult to detect and there is no one identification strategy. Kuriakose and Robbins (2015) suggest researchers set a benchmark (in this example, 85%), wherein any two cases where at least 85% of responses are duplicate to be suspicious. However, this strategy has been argued to produce a large number of false positives (Bohannon, 2016, Simmons, Mercer, Schwarzer, & Kennedy, 2016), and researchers argue that each survey has unique parameters that researchers should account for when analyzing data for potential falsification.

5.9 In surveys conducted at the Allensbach Institute in Germany, researchers have used two different methods to mitigate interviewer falsification in lieu of recording respondent contact information and performing post-survey verification (Smith, 2011a). In the first method, researchers included a factual question in the survey that asked about a little-known fact that would be unanswerable to most respondents. Later in the survey, a second item provided the information that would answer the earlier factual question. In a valid interview, respondents would not be able to go back in the questionnaire to use this information to answer the first question correctly; therefore, it was expected that the vast majority of respondents would provide the wrong answer to the first question. However, an interviewer falsifying responses could potentially use the information to correctly answer the first item. Researchers could then identify any interviewer whose respondents had accurate responses for the first survey question and investigate his or her other completed interviews for a pattern indicating possible falsification. A second technique used by the researchers in Allensbach was to have respondents write responses to open-ended questions. The handwriting could then be examined to see if the interviewer was completing the interviews him- or herself.

6. Document data collection activities.

Rationale

The documentation of data collection procedures is an essential part of the data collection process. Process documentation is necessary for timely intervention. In addition, by understanding what was done in the field, the data are more easily interpreted and understood.

Procedural steps

6.1 Document the following (see Appendix C):
- 6.1.1 A summary of feedback from the feasibility studies.
- 6.1.2 The interview or data collection process.
- 6.1.3 A description of the mode(s) used.
- 6.1.4 A description of the mode-specific protocols.
- 6.1.5 A description of the sample management system.
- 6.1.6 A description of any paradata collected.
- 6.1.7 Special approaches to reduce nonresponse, including any incentives and nonresponse follow-up.
- 6.1.8 Outcome rates by key respondent groups, including response, refusal, noncontact, and other nonresponse rates.
- 6.1.9 Structure of the field staff (e.g., size of interviewer groups and supervisor/interviewer ratio).
- 6.1.10 Timing of the fieldwork for each country or cultural group.
- 6.1.11 A description of quality control procedures and protocols, including:
    - Interviewer monitoring procedures.

- Outcomes of interviewer monitoring, such as hours per interview and any falsification rates.

6.1.12 Any validation study descriptions and outcomes (see Guideline 7 below)

7. When possible, conduct validation studies to estimate bias.

Rationale

As noted in Guideline 3 above, response rates alone are not good indicators of nonresponse bias; understanding nonresponse bias and making subsequent post-survey adjustments require information about the nonrespondents. Similarly, measurement error bias can only be estimated when "true" values for survey variables are known or can be modeled (i.e., using latent class analysis). Validation studies can increase confidence in results, assist with post-survey adjustments (see Data Processing and Statistical Adjustment), and address potential criticisms of the study. However, while the interpretation of survey estimates can benefit greatly from validation studies, conducting them may be difficult and prohibitively expensive.

Survey methodological experiments are designed up front and the outcomes are carefully documented. While these experiments may or may not directly benefit a given study, they are extremely important for the development and building of a body of knowledge in cross-national survey methodology, on which future studies will be able to draw.

Procedural steps

7.1 Collect data on nonrespondents, if possible, to estimate nonresponse bias (Groves, 2006).

7.1.1 One approach is to study sample elements that initially refused to be interviewed.

- Draw a random sample of such initial nonrespondents and attempt to interview them under a modified design protocol (e.g., increased incentives or a shorter interview).
- This approach assumes that people who were initially reluctant to participate are identical to nonrespondents on key variables; this may or may not be a valid assumption (Lin & Schaeffer, 1995).
- Document the data collection procedures, including the proportion of initial nonrespondents included in the validation study, mode of administration, and any additional incentive (Groves & Heeringa, 2006).

7.1.2 A second approach is to compare respondents and nonrespondents on statistics of interest using information contained in external records (e.g., population register data).
- Complete external records for all sample elements may be difficult to find, inaccurate, or outdated.
- These benchmark data are rarely available for statistics of interest.

7.1.3 A third approach is to calculate response rates within subgroups (e.g., racial, ethnic, or gender groups).
- This approach assumes that subgroup membership is related to the propensity to respond, and assumes that biases in demographic variables are informative of biases in substantive variables.

7.1.4 A fourth approach is to compare estimates to similar estimates generated from outside surveys.
- While estimates similar to estimates from these benchmark surveys can increase credibility, the key survey variables may not exist in the benchmark survey. Furthermore, coverage, nonresponse, and measurement error differences in the benchmark survey are largely unknown.

7.1.5 A fifth approach is to examine the effect of post-survey adjustments on the estimates by comparing unadjusted and adjusted values.
- The use of this approach strongly assumes that the models used to adjust for nonresponse fully capture the nonresponse mechanisms at work. While some amount of nonresponse bias may be controlled using these adjustments, they will rarely—if ever—fully control nonresponse bias.
- See Data Processing and Statistical Adjustment for more information on post-survey adjustments for nonresponse.

7.2 Use methodological studies to assess measurement error.

7.2.1 One approach is to use cognitive laboratory techniques, such as cognitive interviews, vignettes, response latency, and behavior coding (see Pretesting), to assess potential measurement error.
- This approach assumes that laboratory measurements are comparable with those obtained in the survey.
- Many laboratory experiments do not use probability-based samples; therefore, errors detected in the self-selected laboratory sample may not be representative of errors in the target population.

7.2.2 Another approach is to check outside records for the true value, or a proxy of the true value, of the measure.

- The researcher must have access to the outside records.
- This approach assumes that the outside records are complete and error-free.
- It may be difficult to match the respondent to the outside record.
- Document record collection procedures, including a description of the records and their quality.

7.2.3 A third approach is to embed a randomized experiment within the survey to assess differences in survey estimates among different measurement conditions. In this situation, respondents should be randomly assigned to the experimental conditions (e.g., interview mode).

7.3 Consider using other methods of assessing measurement error.

7.3.1 Reinterview respondents. Reinterviews are especially useful in determining interviewer falsification (Forsman & Schreiner, 1991) but may also help assess other forms of measurement error (see Biemer, 2004; Biemer & Forsman, 1992) for details on estimating simple response variance or bias).

7.3.2 Document all aspects of the reinterview procedure, including:

- The respondents who were eligible for the reinterview component of this study (e.g., random 10% of respondents), as well as the total number of respondents selected and how many completed the reinterview.
- The questionnaire used in the reinterview.
- The mode of administration of the reinterview.
- The interviewers who administered the reinterview (e.g., any project interviewing staff, specially designated interviewers, supervisory staff, clinicians, self-administered, etc.).
- The time interval between administration of the main interview and the reinterview (e.g., reinterviews were conducted 1-2 weeks after the main study interview).

7.3.3 Collect paradata that may be correlated with measurement error (e.g., number of keystrokes, length of interview).

7.3.4 Use interpenetration to estimate correlated response variance due to interviewers.

Lessons learned

7.1 Supplemental studies can be difficult and expensive to implement, but they are useful for validating survey results. For example, a study of discharged patients at a French hospital found no difference in patient satisfaction ratings between early and late respondents. The authors interpreted this finding to indicate that there was little evidence of nonresponse bias in their estimates of patient satisfaction. However, it is unclear if the differences in estimates were due to nonresponse bias or to measurement error (Gasquet, Falissard, & Ravaud, 2001).

7.2 Try to use resources to gain knowledge on bias in an efficient way. Validation studies are expensive but come late. Therefore, one should first strive for more preventive measures that hopefully make processes almost error-free. Then paradata should be collected and analyzed so that processes can improve and display a decreased variability. Finally, some small-scale validation studies, rather than large ones, should be conducted, and used as input to more long-term improvements of processes and methods. The optimal allocation between the three is unknown but the general preferred allocation is evident, namely prevention first, then process adjustments via paradata, and lastly small validation studies.

Appendix A

Example of pre-notification letter from ESS 2008 (Forsman & Schreiner, 1991).

\

Logo and Contact information of research organization →

Logo of research organisation

Autumn 2008

[Dear .....]

European Social Survey 2008

Purpose of the survey →

[You have/your address/household has] been selected to take part in an international study on what people think about various important issues affecting [country]. The study is being carried out simultaneously in 25 countries across Europe and will help to find out how much or how little people in different countries share the same views and beliefs. It is being paid for from both [country] and European sources.

Confidentiality of answers →

The questionnaire covers a wide range of topics and no special knowledge is needed to answer any questions. Your [name/household/address] has been selected from [sampling frame] by scientific methods to ensure that we get a representative picture of people in [country]. We cannot therefore substitute any [name/household/address]. All information you provide will be treated in strict confidence and will never be linked to your name or address.

Voluntary participation →

Most people taking part in the study find it an interesting and enjoyable experience, and we hope that you will too. An interviewer will [visit] you shortly to explain more about the study and, if you agree, will arrange a suitable time for the actual interview, which will be carried out in person. Interviews normally take just under an hour. We certainly hope we can rely on your co-operation.

Meanwhile, if you wish to have any further information about the study, please feel free to contact me on the number above.

Thank you in advance for your help,

XXXX XXXXXX

Information on survey interviewer →

The interviewer who will be contacting you is:..

Appendix B

Contact attempt record (example from the University of Michigan's Institute for Social Research)

	CALL #1	CALL #2	CALL #3	CALL #4
DATE:				
DAY OF WEEK:				
EXACT TIME BEGAN:				
IWER ID:				
CONTACT WITH:	R / INF/ NO ONE	R / INF/ NO ONE	R / INF/ NO ONE	R / INF/ NO ONE
MODE OF CONTACT:	PERSONAL / TEL	PERSONAL / TEL	PERSONAL / TEL	PERSONAL / TEL
TELEPHONE NUMBER IF OBTAINED:				
HU LISTING OBTAINED:	YES / NO	YES / NO	YES / NO	YES / NO
DETAILED DESCRIPTION OF CONTACT OR CONTACT ATTEMPT				
DISPOSITION CODE:				

R = Respondent HU = Housing Unit
Inf = Informant Listing = enumeration

Appendix C

Documentation

	Details	Examples from SMDS[1]	Examples from ISSP[2]
Data collection organizations	• The number of organizations • Contact information • Type of organizations (e.g., government agency, private research company)	How many organizations conducted data collection for this study in your country? If your agency/organization contracted with another organization which provided data collection services, please include that here. ____ organizations	Please enter the name of your institute and your country: Institute: _______________ Country: _______________
Data collection methods	• The number of separate survey data collection efforts • A brief title of each survey data collection efforts Delivery of sample to interviewers (e.g., computerized sample management system) • Mode of data collection • Screening/respondent selection procedure	How were the face-to-face interviews administered in this study? Please check all that apply. - Computer-assisted personal interviewing (CAPI) - Paper and pencil interviewing (PAPI) - Other, specify: ____________________	What selection method was used to identify a respondent? Please tick all that apply. (do not answer if your sampling frame consists of named individuals — which are the target persons. Then continue with question 66) Kish grid ____ Last (or next) birthday _____ Quota _____ Other (please write in details) _____
Techniques used to maximize response rate	• Pre-notification strategies	Which, if any, of the following pre-notification strategies were used for the face-to-face interviews that were conducted in this study? Please check all that apply. - Advance letter sent prior to initial visit - Email message sent prior to initial visit - Telephone call made prior to initial visit - Announcement in local newspaper, radio, or television - Other, specify: _______________ - None of the above	Were postal or telephone components used at any point (e.g., advance contacts)? Yes - postal ?Question 41 Yes - telephone ?Question 41 No ?Question 42
	• Use of incentives • Specific incentive offers made to a particular group of sample members	How many different respondent incentives were initially used for the face-to-face interviews that were conducted in this study? For example, if half of the respondents were randomly assigned to receive â,¬15 and the other half received a gift basket, this should be recorded as two incentives; or if â,¬10 was included in the advance letter and â,¬20 was promised upon completion of the interview, this should also be counted as two incentives.	Were incentives offered? Yes, to respondent _____ Yes, to interviewer _____ No, neither to respondent nor to interviewer _____

	Details	Examples from SMDS[1]	Examples from ISSP[2]
		____ different incentives used {1-10}	
	• Refusal conversion protocols • Interviewer incentives/ bonuses	Which, if any, of the following (additional) techniques were used to maximize response rate for the face-to-face interviews that were conducted in this study. Please check all that apply. - Special refusal aversion or refusal conversion training sessions for interviewers - Specially designated interviewers for reluctant cases (e.g., flying experienced interviewers in to attempt difficult cases) - Persuasion letters sent to reluctant sample members - Increased or additional respondent incentives implemented after the start of data collection - Interviewer incentives/bonuses - None of the above	
Contact protocols	• Minimum number of contacts (on weekday, in the evening, on weekends) before a case is finalized • Maximum number of contacts after a case would be finalized • Methods used to attempt to reach sample members	Was there a minimum number of attempts required before a sample case was finalized (i.e., no more attempts were made on the case) for face-to-face data collection in this study? - Yes - No	Were interviewers required to make a certain number of calls/ visits before they stopped approaching an address or household? Minimum number of calls/visits required - please write in number _____ No minimum call requirement ____
Eligibility screening	• Minimum number of attempts for screening (on weekday, in the evening, on weekends) before the case is finalized • Maximum number of attempts for screening after a cases would be finalized • Methods for refusal conversion for eligibility screening • Methods to reach sample members for the screening • Any additional techniques that were used to increase response rate for the screening to determine eligibility	What was the minimum number of attempts required before a case was finalized (i.e., no more attempts were made on the case)? If the mode of contact was not specified, please only provide the total number of attempts below. ____ minimum face-to-face attempts {ALLOW VALUE,1-40} ____ minimum telephone attempts {ALLOW VALUE,1-40} ____ total minimum attempts (face-to-face and telephone) {ALLOW VALUE,1-40}	Was substitutionor replacement permitted at any stage of your selection process or during fieldwork? Yes? Question 67 No? Question 68
Use of special test or data collection besides survey		Besides the survey questions, did this study involve any of the following? Please check all that apply. - Tests of physical performance (e.g., walking speed, grip strength)	

	Details	Examples from SMDS[1]	Examples from ISSP[2]
questions		- Tests of cognitive ability (e.g., memory tasks, word recognition) - Physical measurements (e.g., height, weight, blood pressure) - Collection of biological specimens (e.g., blood, saliva, urine) - Collection of environmental specimens (e.g., soil, dust) - Procurement of respondent permission to access and link respondent data from other sources (e.g., government records, medical records, employment records) - Other, specify: {TEXT BOX} - None of the above	
Locating sample members	• Tracking procedures ○ leader/coordinator of tracking ○ tracking manual/tracking team ○ training ○ between wave tracking efforts ○ steps/options used in tracking (relatives, friends, neighbors, and employers)	Were any tracking activities carried out to locate sample members in this study? Please check all that apply. - Yes - No {SKIP TO DC223}	
Quality control	• Supervision • Back-checking		Were any interviews back-checked (e.g. supervisor checks later whether interview conducted)? Yes - please write in approximate proportion % ______ No ______

[1] Survey Metadata Documentation System (SMDS): a standardized web-based documentation tool which was developed by the University of Michigan's Institute for Social Research and Gesis-ZUMA.
[2] International Social Survey Programme (ISSP): see Scholz & Heller (2009), for details.

Appendix D

Disposition codes (American Association for Public Opinion Research, 2016)

- The coordinating center should provide a list of specific disposition codes with a clear description of how to code all sample elements during (temporary disposition) and at the close of (final disposition codes) the field period.

- Generally, disposition codes identify sample elements as (complete or partial) interviews or non-interviews.
  - The coordinating center should set the criteria for determining whether interviews are classified as complete or partial.
  - Non-interviews are grouped by whether the respondent is eligible, unknown eligible, or ineligible to participate in the study.

- Disposition codes are mutually exclusive. While sample elements may be assigned different temporary disposition codes throughout the field period, there will be only one final disposition code.

Appendix E

Components and descriptions of each category of response rate calculation (for a sampling frame of housing units) (American Association for Public Opinion Research, 2016)

- To standardize the response rate calculations across countries, every country should group each sample element's final disposition code into one of the following mutually exclusive and exhaustive categories:
  A. Interviews
  B. Non-interviews—Eligible
  C. Non-interviews—Unknown eligibility
  D. Non-interviews—Ineligible

A. Interviews

Component	*Description*
Complete interviews	• Respondent has finished the interview.
Partial interviews	• The survey organization (in consultation with the coordinating center) may decide prior to the start of data collection to consider an interview to be a partial interview if at least some percent (e.g., 80%) of applicable or crucial/essential questions have been answered.
TOTAL INTERVIEWS	• Sum of interviews.

B. Non-interviews—Eligible

Component	*Description*
Refusals	• It has been determined that there is an eligible respondent in the housing unit but either he/she or someone else refuses the interview request.
Non-contacts	• It has been determined that there is an eligible respondent in the housing unit but the interviewer cannot gain access to the building, no one is reached at the housing unit, or the respondent is never available when the interviewer attempts an interview.
Other	• It has been determined that there is an eligible respondent in the household (eligibility determined as of a particular date, e.g., the date that the household listing is taken) but at some time after the determination of eligibility, the respondent is unable

	to complete the interview due to reasons other than a refusal or is unable to be reached after repeated attempts. For example, the respondent may have died, been incarcerated or hospitalized, or left the country. • It has been determined that there is an eligible respondent in the household, but he/she does not speak any of the study language(s) or is permanently incapable of participating in the interview due to a physical or mental condition (e.g., senility, blindness, or deafness). Note: Sample elements may be considered ineligible if the target population is defined such that respondents who do not speak the study language(s) or respondents who are unable to hear are explicitly excluded from the target population to which the study plans to makes inferences. • Any other eligible non-interview status.
TOTAL NON-INTERVIEWS—ELIGIBLE	• Sum of eligible non-interviews. • If the survey organization is unable to provide separate counts of each component but the survey organization can provide the total number of eligible non-interviews, use the total.

C. Non-interviews—Unknown eligibility

Component	*Description*
Unknown if household/occupied housing unit	• The sample elements have not been attempted or worked (e.g., no interviewer is available in area or replicates are introduced too late to work all sample elements). • Interviewer is unable to reach the housing unit due to weather or concerns about safety in a dangerous neighborhood. • Interviewer is unable to locate the housing unit (e.g., inaccurate or inadequate address/locating information).
Unknown if eligible respondent is in unit/no screener completed	• It has been determined that there is an eligible housing unit but the interviewer is unable to determine whether there is an eligible respondent in the unit. For example, a household member may refuse to complete the screener or no one is available to complete the screener when the interviewer visits the household. Note:

	These sample elements are not considered refusals, since only elements where it has been determined that there is an eligible respondent can be classified as refusals.
Other	• Any other status for which eligibility is unknown
TOTAL NON-INTERVIEWS—UNKNOWN ELIGIBILITY	• Sum of non-interviews of unknown eligibility • If the survey organization is unable to provide separate counts of each component, but the survey organization can provide the total number of non-interviews of unknown eligibility, use the total.

D. Non-interviews—Ineligible

Component	*Description*
Not an eligible housing unit	• The sample elements are out-of-sample housing units or housing units that are incorrectly listed in the address frame (e.g., housing units are outside the primary sampling unit in which they are thought to be located). • The sample elements are non-residential units (e.g., businesses, government offices, institutions, or group quarters). • Housing units are vacant on the date that eligibility is determined. Note: Sample elements may be considered eligible non-interviews if someone is present at the housing unit on the date that eligibility is determined, even if when the interviewer returns the household has moved and the unit is vacant. • Households are temporary, seasonal, or vacation residences (i.e., not the usual place of residence).

No eligible respondent	• It has been determined that there is an eligible housing unit, but there is no eligible respondent in the unit. For example:
	  • Residence with no one 18 years of age or older.
	  • Respondent does not speak any of the study language(s) and the target population is explicitly defined such that respondents who do not speak the study language(s) are not considered part of the target population to which the study plans to make inferences (may also hold for physical or mental conditions, if the target population is explicitly defined to exclude persons who are blind, deaf, senile, etc.).
	  • Respondent died before eligibility is determined.

	• Respondent is incarcerated or hospitalized (i.e., institutionalized) at the time that eligibility is determined, and remains institutionalized throughout the data collection period.
Other	• Respondent is in a group/cell for which the quota has already been filled. • Any other ineligible non-interview status.
TOTAL NON-INTERVIEWS—INELIGIBLE	• Sum of ineligible non-interviews. • If the survey organization is unable to provide separate counts of each component but the survey organization can provide the total number of ineligible non-interviews, use the total.

Appendix F

Recording counts of response rate categories template (for a sampling frame of housing units) (American Association for Public Opinion Research, 2016)

- Use the template below to help determine the number (or weighted count, if appropriate) of sample elements finalized in each of the categories and, thus, the total number/weighted count of sample elements fielded. The total number of sample elements is the sum of all categories of the final disposition codes.
    - First, enter the number of sample elements finalized as each given category component. If no sample elements are finalized as a particular category component, enter "0" in the "Count" column.
    - Next, total the components for each category by entering the sum on the longer of the "Count" column lines.
    - Finally, total the sums of each category by entering the overall sum on the last "Count" column line.
    - Use the "Additional Information" column to provide any information that will assist in interpreting the figures provided, particularly the study's definition of partial interviews or descriptions of "Other" classifications specific to the study.

Category (with Components)	*Count*	*Additional Information*
A. Interviews		
Complete interviews	___	____________________
Partial interviews	___	____________________
TOTAL INTERVIEWS	_____	
B. Non-interviews—Eligible		
Refusals	___	____________________
Non-contacts	___	____________________
Other	___	____________________
TOTAL NON-INTERVIEWS—ELIGIBLE	_____	____________________
C. Non-interviews—Unknown eligibility		
Unknown if household/occupied housing unit	___	____________________
Unknown if eligible respondent in unit/no screener completed	___	____________________
Other	___	____________________
TOTAL NON-INTERVIEWS—UNKNOWN ELIGIBILITY	_____	____________________
D. Non-interviews—Ineligibility		

Not an eligible housing unit No eligible respondent Other TOTAL NON-INTERVIEWS—INELIGIBILITY	____ ____ ____ _______	____________________ ____________________ ____________________ ____________________
TOTAL NUMBER OF SAMPLE ELEMENTS	_______	____________________

Appendix G

Recording counts of response rate categories for additional eligible respondents template (for a sampling frame of housing units) (American Association for Public Opinion Research, 2016)

- Use the template below to help determine the number (or weighted count, if appropriate) of additional respondents in each of the categories and, thus, the total number/weighted count of additional respondents. The total number of additional respondents is the sum of only the eligible categories of the final disposition codes; if a household was not eligible, no respondents—let alone additional respondents—were selected.
    - First, enter the number of additional respondents finalized as each given category component. If no additional respondents are finalized in a particular category component, enter "0" in the "Count" column.
    - Next, total the components for each category by entering the sum on the longer of the "Count" column lines.
    - Finally, total the sums of each category by entering the overall sum on the last "Count" column line.
    - Use "Additional Information" column to provide any information that will assist in interpreting the figures provided, particularly the study's definition of partial interviews or descriptions of the "Other" classification specific to the study.

Category (with Components)	*Count*	*Additional Information*
A. Interviews Complete interviews Partial interviews TOTAL INTERVIEWS	 ___ ___ ______	 ____________________________________ __________________
B. Non-interviews—Eligible Refusals Non-contacts Other TOTAL NON-INTERVIEWS—ELIGIBLE	 ___ ___ ___ ______	 __________________ __________________ __________________ __________________
TOTAL NUMBER OF ADDITIONAL RESPONDENTS	______	__________________

Data Collection: Face-to-Face Surveys

Julie de Jong, 2016

Introduction

Many cross-cultural projects attempt to keep the mode of administration constant by choosing face-to-face data collection, where the survey questionnaire is administered, at least in part, by a survey interviewer. Generally, the face-to-face mode has the best sample coverage properties, highest response rates (and therefore possibly lower nonresponse bias), and does not require respondents to be literate. For a discussion of the advantages and disadvantages of the face-to-face mode of interviewing, see Study Design and Organizational Structure.

Before the advent of personal computing, face-to-face surveys were administered using a paper-and-pencil instrument (PAPI). However, laptops and other electronic instruments (e.g., tablets, smartphones, etc.) are now widely used in lieu of PAPI.

In order to collect comparable data, multinational, multiregional, and multicultural surveys ("3MC" surveys) must establish a standard data collection protocol. At the same time, the protocol will sometimes need to allow for modifications required by local norms, conditions, or customs.

The implementation of face-to-face surveys presents a number of logistical challenges not faced in other modes. This chapter first addresses issues pertaining to the face-to-face mode, regardless of the instrument used to collect the data (i.e., paper and pencil questionnaire, computerized instrument, etc.), and then presents considerations particular to each type of instrument.

Guidelines

Goal: To achieve an optimal 3MC data collection design by maximizing the amount of information obtained per monetary unit spent within the allotted time, while meeting the specified level of precision and producing comparable results, within the context of a face-to-face survey.

1. **Consider the following steps when conducting survey interviews using a face-to-face mode. Surveys conducted by interviewers face-to-face share a number of common procedural steps.**

   Rationale

   There are a number of important considerations when interviewers are contacting respondents in a face-to-face survey, whether the instrument is paper-based (PAPI) or computer-based (CAPI).

Procedural Steps

1.1 Contact local authorities for clearance for the interviewers to collect data at the sample site(s); if necessary, negotiate with local authorities or, in some cases, military authorities to gain access to sample areas.

1.2 Allow adequate time for interviewer recruitment and training
 1.2.1 Match interviewer and respondent characteristics (e.g., race, ethnicity, or gender) when cultural norms so dictate, and/or if there is reason to think that interviewer effects may occur depending on the social conditions (see Interviewer Recruitment, Selection, and Training).
 1.2.2 While interviewers might be relatively easy to recruit in some countries, in other places such as the Gulf States and Middle East region, researchers might face some challenges in recruiting qualified field interviewers (Gengler, 2013).

1.3 Take measures to ensure interviewer safety.
 1.3.1 Inquire about potential safety problems, such as civil unrest and high crime areas.
 1.3.2 Decide whether interviewers should travel in groups and be accompanied by security personnel.
 1.3.3 Have interviewers visit their work areas during the daytime before the first day of data collection. They should check for potential hazards and safe havens during this visit.
 1.3.4 Have interviewers tell their supervisors and family members when they plan to leave for the field, the location of the area, and when to expect them back.

1.4 Have interviewers carry the following items in the field to establish their legitimacy:
 1.4.1 Official identification from the survey organization.
 1.4.2 Official letters to local authorities describing the study, if appropriate.
 1.4.3 Other letters of permission or support from local authorities if appropriate and/or necessary given the local social context and governmental regulations.

1.5 Provide adequate transportation and accommodation for staff and supplies.
 1.5.1 If maps are unavailable or unreliable, consider the use of local guides or GPS instruments.
 1.5.2 Arrange to secure fuel and oil and to maintain the vehicles used by the field staff; this may present logistical problems in some countries where there are breakdowns in infrastructure.

1.5.3 Arrange for emergency transportation in the event that a field team member becomes ill or injured and needs immediate medical attention or it becomes unsafe to stay in an area.
1.5.4 Arrange for backup transportation.
1.5.5 Secure housing accommodations in more remote areas prior to fieldwork or have the team carry their own (e.g., tents or mobile homes).

1.6 If physical measurements are taken as part of the survey, check the cultural acceptance of taking such measurements.

1.7 Provide all members of the field staff with access to a reliable line of communication with their supervisor. This will allow them to report progress and problems.
1.7.1 Majority countries may have weak communication capacities, especially in rural areas.
1.7.2 Cellular or satellite phones may be a worthwhile investment for teams in remote areas.

1.8 Aim to conduct the interview in a setting which affords visual, physical, and auditory privacy.
1.8.1 Privacy is critical for keeping respondents' answers to the survey questions confidential.
1.8.2 Although complete privacy is ideal, it is impossible to achieve in some cultures. Interviewers should attempt to keep the interview as private as possible, while still respecting cultural norms. This may involve self-administration on more sensitive questions. See Guideline 4 below regarding self-administration in the context of a face-to-face interview. An alternative may be to keep any others present occupied while the targeted respondent completes the survey.
1.8.3 In some countries, it may be unacceptable to have an interviewer come to the respondent's home, or it may be unacceptable for an interviewer of the opposite sex to interview or enter the home of the selected respondent or informant. As noted above, this may necessitate interviewer-respondent gender matching.
1.8.4 Privacy increases the likelihood that respondents will answer honestly about sensitive behaviors, such as sexual practices or drug use, or about sensitive attitudes such as religion in some contexts. What is considered sensitive may vary among countries or cultures; administration practices may need to differ accordingly.

1.9 In order to reduce non-response in the face-to-face mode of interviewing, train the interviewers to make observations of the housing unit to assess likely at-home patterns.

    1.9.1 Note that in some countries interviewers are not allowed to ask neighbors about targeted but not yet contacted respondents.

Lessons learned

1.1 Because responses to some survey questions can be affected by other individuals present during data collection, it is optimal—but not always possible—to conduct face-to-face surveys in private. In a face-to-face fertility survey of women in what is now Bangladesh, privacy was difficult to establish; most interviews took place in the presence of the respondent's mother- or sister-in-law. This may have affected responses to sensitive questions (Choldin, Kahn, & Ara, 1983).

1.2 Similarly, men in some parts of Africa were found to object to confidential interviews of their wives or children. The interviewers were instructed to conduct interviews in a place that was visible to the male heads of household but out of earshot (Chikwanha, 2005).

1.3 In some rural places it might not always be feasible to conduct an interview inside a home, and may have to take place outside and in a more public setting.

1.4 In other rural places, the survey interview is still a novel concept, making interview privacy difficult to attain. In the Chitwan Valley Family Surveys in Nepal, a survey interview often draws family members and even interested neighbors, who sit with the respondent and interviewer to listen in.

1.5 Analyses using data from nine countries participating in the World Mental Health Survey Initiative provided evidence that the presence of a third party during the survey interview process affected the reporting of sensitive information, but the effect is moderated by differences in social conformity and the cultural setting from country to country (Mneimneh, Tourangeau, Pennell, Heeringa, & Elliott, 2015).

2. Consider the following steps when using a paper and pencil instrument (PAPI).

Rationale

While the use of computerized technology has increased in survey administration, the paper-and-pencil instrument continues to be used by survey projects that lack the infrastructural capacity to adopt and maintain the necessary technology.

Procedural Steps

2.1 The paper instrument should be designed so that it is visually easy for the interviewer to administer. See Instrument Technical Design for further detail.

2.2 Develop a sample management protocol for use in the field by data collection supervisors. The protocol should include instructions for passing sampling units from one interviewer to another if the need arises, as well as the corresponding documentation of such transfers (see Data Collection: General Considerations 3.3 and sample management system).

2.2.1 Use a coversheet to track each sample element during the study (see Appendix A for an example of a paper coversheet),

2.2.2 Interviewers using paper coversheets have found that they work most efficiently if they sort the coversheets by (1) appointment time and (2) geographical location.

2.2.3 Consider efficient methods that allow interviewers to fill in coversheets and do household contacting at the same time. Filling in coversheet forms after making the contact has shown to be error prone.

2.3 Train interviewers to complete household enumeration and randomly select eligible members within the household unit (Kulka & Weeks, 1988) (see Appendix B for household enumeration and Appendix C for an example of a Kish table).

2.4 Develop a distribution procedure for supplies to interviewers in the field, including a surplus of paper questionnaires to ensure a continual supply.

2.5 Develop a protocol for transferring completed paper questionnaires from interviewers to field supervisors, and from field supervisors to the head office or other location where data entry will occur.

2.6 Develop a protocol for maintaining completed questionnaires and coversheets in a secure location to ensure protection of respondent confidentiality.

Lessons Learned

2.1 Paper questionnaires and other survey materials can be misplaced, stolen, or otherwise lost in the field. Document any such circumstances and develop a protocol to determine whether affected respondents will be recontacted for a repeat interview.

2.2 Plan for adequate storage, security, filing system to get back to interviews efficiently.

2.3 In certain countries, like Ghana, weather conditions such as high humidity can destroy paper questionnaires in storage.

2.4 Researchers administering a PAPI survey of business and social entrepreneurship in the Kingdom of Tonga with complex skip patterns used a detailed skip pattern map to simplify training and questionnaire preparation. They also developed a notation system on the actual questionnaire page to assist the interviewer (Frederick, 2012).

2.5 If there are multiple components to the questionnaire, consider using paper of different colors for each component (e.g., the coversheet in yellow, interviewer-completed survey in green, self-administered section in orange, etc.).

2.6 Alternately, if the questionnaire will be administered in several different dialects or languages within a country, consider printing each dialect/language on differently-colored paper.

2.7 Consider using heavy card stock or lamination for Show Cards and other paper-based instruments that will be used for multiple respondents.

2.8 If using an event history calendar or other unusually-sized instrument, allow for adequate printing time, particularly in countries where printing of odd-sized documents may be challenging. Researchers in Nepal report having difficulties in locating printing businesses with the capacity to print the large life-history calendars designed for administration.

2.9 Researchers administering a PAPI survey in the Kingdom of Tonga faced limitations in printing in the country itself, including the lack of paper, printing cartridges, and water-resistant paper that could withstand moisture and travel. Researchers emailed the questionnaire to a specialty printer in New Zealand, and the printed questionnaires were sent via airmail back to Tonga for use (Frederick, 2012).

2.10 The cost of paper can be very expensive in some countries. If the survey instrument contains many skip patterns, there can be a lot of waste as well. For example, the PAPI version of the World Mental Health Composite International Diagnostic Interview (CIDI) 3.0 was about 400 pages in length, but contained numerous sections which began with a screener question and resulted in the respondent skipping the entire section(s) for which the questions were not applicable.

3. If an electronic instrument will be used instead of a paper-based instrument, consider the following procedural steps.

Rationale

As technology becomes more accessible and affordable, with use increasing worldwide, computer-assisted personal interviewing (CAPI) is a popular mode choice and is frequently used in lieu of PAPI. Laptop computers have generally been the instrument of choice for CAPI, but tablets, smartphones, and other handheld device are increasing in popularity.

Procedural Steps

3.1 If CAPI use is new to the study site, develop an introduction strategy for both local collaborators and study respondents.

3.1.1 Involve local collaborators in study design if possible to facilitate its adoption. The clinical and administrative staff in a rural Kenyan health center aided in the identification of appropriate data and formatting of the paper and electronic data recording interfaces. This helped reduce fears and distrust of computers and engaged the clinical staff in the clinical research project (Diero et al., 2006).

3.1.2 In settings with limited technology, computerization can stimulate survey respondent interest and add legitimacy to the interviewers. Interviewers might also be more motivated to use technology in such setting. However, at the same time and in certain cultures, the use of technology can raise suspicion among respondents (Paudel, Ahmed, Pradhan, & Dangol, 2013).

3.2 Assess technical experience at the data collection firm.

3.2.1 Critical staff should have adequate language competency. Programs interested in incorporating technology into their activities need to hire bilingual staff as trainers and programmers to improve understanding of how to use the chosen technology, and to facilitate design and analysis activities, as most technology specifications are available only in universally-used languages such as English.

3.2.2 The data collection organization needs to have technical expertise to create the questionnaire, provide technical support for interviewers, manage the in-flow and out-flow of data, manage databases, and run quality control checks. While some of these tasks could be outsourced, building local capacity is always recommended for continuity and long-term goals.

3.3 Assess available infrastructure in the study country.

3.3.1 If the data collection organization requires data to be transmitted on a regular basis for quality control, and reliable Internet connectivity needs to be in place, evaluate WiFi and other Internet connection capabilities across the geographic areas covered by the sample. Even though a country's major cities may have adequate Internet capabilities allowing for regular data transfer, rural areas may present more challenges.

3.3.2 Interviewers and field office staff need to have access to reliable electrical power sources for the interviewing as well as communication devices (e.g. mobile phones). Interviewers might need to carry multiple batteries for their devices if they are visiting areas with limited power supply. Interviewers could also be instructed to use other methods for charging batteries including in-car chargers such as cigarette-lighter adapter or portable generators (Shirima et al., 2007; Byass et al., 2008)

3.3.3 If the need to revise the questionnaire during data collection arises, computerization and connectivity allows for an easy transmission of updated questionnaires to interviewers or respondents without the need for reprinting, mailing, or personal pick-up of material. Moreover, avoiding printing any material at, before, or during production is environmentally friendly.

3.4 Choose and procure the necessary primary and auxiliary equipment.

3.4.1 Primary equipment

- There must be a good fit between the project and the technological tool. Handheld devices may be more appropriate for smaller or simpler questionnaires, and, because of their size, devices such as smartphones are not as suitable for collecting open-ended responses (Escandon, Searing, Goldberg, Duran, & Monterrey Arce, 2008).

- Purchasing equipment and accessories locally can facilitate more efficient servicing of equipment than if equipment is purchased internationally (Paudel et al., 2013).
- Although new technology may be more expensive if purchased locally within less-developed countries, the cost saved in shipping, delays, and in-country technical support can more than compensate for that difference.
- If equipment is not available locally; however, most hardware is available through collaborators in industrialized countries or can be ordered directly via the Internet.
- Consider ordering an excess supply of batteries and extra equipment (e.g., several extra laptops) in case of equipment malfunction.

3.4.2 Auxiliary equipment

- Decide on a backup and uploading process (SD cards, flash drives, automatic uploading to central system, etc.).
- Data synchronization between a mobile device and a central computer can be very time consuming in a rural, remote setting. In a survey in Zanzibar, a mobile device was used to collect data, store and copy the data from the SD card to the central computer (Thriemer et al., 2012).
- The back-up system must be carefully developed to handle possible transitions or losses. In a root-cause analysis from a survey using PDAs in Bolivia, poor back-up protocol, due to programmer error, precluded researchers from interpreting the data (Escandon et al., 2008).
- If possible, at least two separate central back-up systems should be developed, in addition to having back-up on the unit itself (i.e., memory cards) and a communal archiving system.

3.5 Select appropriate data collection software.

3.5.1 Additional attention should be given to non-Latin languages (i.e., Chinese, Arabic, Russian, etc.) when selecting technology and programming software. Not all software packages can support non-Latin script.

3.6 Select an appropriate electronic sample management system.

3.6.1 If an electronic sample management system is used, coordinating centers may play a role in monitoring fieldwork. See Study Design and Organizational Structure for details.

3.6.2 The electronic sample management system should permit interviewers to be able to sort the sample respondents by (1) appointment time and (2) geographical location.

3.7 Develop and test the CAPI instrument

3.7.1 Allow for sufficient time and budget for computerized specifications in the preproduction phase (House & Nicholls, 1988)

3.7.2 Consider using paper documents for certain aspects of the survey. For example, interviewers in China using handheld computers reported that it was overly time-consuming to read the full consent form on a small screen (Wan et al., 2013).

3.8 Develop a distribution system for supplies to the field.

3.8.1 Develop procedures for storage and transport of equipment.

3.8.2 Interviewers who are traveling long distances, through difficult terrains, or weather conditions find it easier to carry their laptop or even smaller devices (tablets) to conduct their interviews than carrying cumbersome paper questionnaires (Paudel et al., 2013).

3.9 Develop procedures for use and maintenance of technology in the field.

3.9.1 Charge batteries daily to mitigate data loss due to battery discharge. Instruct interviewers to verify daily that batteries are charged.

3.9.2 Provide interviewers with a reliable electrical source to charge both CAPI instrument and mobile phones batteries so that interviewers can contact supervisors in the event of equipment malfunction. Communication is necessary for possible instrument troubleshooting and monitoring team progress. Most technical issues are simple user errors that can be resolved with a short discussion with the supervisor.

3.9.3 Backup plans need to be designed and implemented in case of power outages, especially in resource-constrained environments.

3.9.4 It is possible for data to be lost because of hardware or software malfunction and for equipment to be lost or stolen during fieldwork. Researchers need to establish protocols for preventing and handling such situations.

3.9.5 Decide whether interviewers should be provided with paper copies of the questionnaire or some material to take notes in-case of equipment failure (Onono, Carraher, Cohen, Bukusi, & Turan, 2011). Some studies choose not to provide paper versions because they do not want to encourage use of alternate paper instrument by interviewers.

3.9.6 It is crucial to have local informatics experts for development and custom integration of databases, continued support, and adaptation to new applications. Specifically, a programmer with experience in database and systems design, implementation, and maintenance is recommended, and this resident expertise is available in most if not all countries (Avilés, Ortega, Kuan, Coloma, & Harris, 2007).

3.9.7 Equip interviewers with accessories that are needed for protecting and maintaining the equipment such as laptop bags, screen covers, sleeves, rain shields, etc.

3.9.8 Ask interviewers sign term of use agreement detailing equipment's ownership and responsibilities.

3.10 Management of data files during the field period.

3.10.1 The electronic data audit trail provides important paradata and should be uploaded and backed up as well. Determine what will happen to paradata in case of equipment failure during interview.

3.10.2 Lack of electricity and/or Internet connection can lead to delays in the backup and uploading process. For example, a survey in Kenya experienced delays in immediate transfer of data collected due to electrical instability, and data often could not be backed up in the field so was only backed up once a week at the study office (Onono et al., 2011).

3.11 Develop strategies to increase privacy

3.11.1 Though interviewer-administered computerization can in general increase the level of respondent privacy, the novelty of it in some cultures might attract bystanders, and interviewers may need additional training on how to request and achieve privacy in such situations (Paudel et al., 2013). DHS interviewers in Nepal found they often had to make extra effort to maintain privacy, which usually demanded more time to administer the questionnaire (Paudel et al., 2013). See also Ali et al. (2010).

3.11.2 Reading computer screens under direct sunlight can lead to difficulty in administering an interview and limit the options for confidential interview space. This can be a particular concern when asking sensitive questions related to sexual behavior and domestic violence.

3.12 Devote adequate time to interviewer training for CAPI-specific issues. When using CAPI, interviewer training is a two-step process, requiring technical training focused specifically on the survey instrument itself (e.g., the introduction of the instrument to respondents, how to use and care for the instrument, etc.), as well as study-specific and general interviewer training.

3.12.1 Instruct interviewers on how to introduce technology to the survey population, especially in settings where exposure to technology is more limited. This could be done by collaborating with community leaders who could act as liaisons announcing the survey and the use of technology to their community members.

3.12.2 Instruct interviewers on how to explain the use of technology to respondents during the consent process (e.g., that recording will or will not be disabled).

3.12.3 Provide training on how to handle, label, care, transport, and store equipment properly. This is especially important in contexts where technology is more novel.

3.12.4 Instruct interviewers on steps to take in case of equipment failure and theft.

3.12.5 Instruct interviewers on password use, stylus if needed, how to access the questionnaire, enter responses, and insert and remove any memory cards used.

3.12.6 Operational instructions should be in study site language and not only in English (Wan et al., 2013).

3.12.7 If paper questionnaires will be available in the event of equipment malfunction, training on the PAPI instrument is essential as well.

3.12.8 When using technology, there can be a tendency for interviewers to focus on the technology rather than the respondent, which should be addressed during interviewer training.

3.12.9 Allow interviewers ample time to practice administering the questionnaire to increase comfort with the flow of questions. Interviewers are more likely to lose track of where they are in the sequences of questions because they can see only one screen at a time, and familiarity with the instrument can decrease difficulty (Groves & Mathiowetz, 1984; House & Nicholls, 1988; Couper, 2000).

3.12.10 If using an electronic sample management system, train interviewers to complete household enumeration and randomly select eligible members within the household unit.

3.12.11 Although interviewers must be trained in the use of the specific computer program, it is crucial to devote adequate time to training on other important interpersonal aspects of survey implementation (Groves et al., 2009).

3.13 Develop interviewer management procedures for use in the field.
  3.13.1 Interviewers must have fast and regular communication of field staff with team leaders and technical support staff. This is necessary for troubleshooting and monitoring team progress.
  3.13.2 Information technologies allow implementation of a system of work ownership if all personnel are assigned a code for database entry, supervision, and analysis to maintain logs controlling data management and information flow.

Lessons Learned

3.1 Technology can be adopted even in resource-poor countries, leading to improvements in efficiency and data collection capabilities.
  3.1.1 Researchers successfully conducted a Demographic and Health Survey (DHS) in Nepal using tablet PCs. The connection to the central network took, on average, one minute, and data transfer to the server in Kathmandu took approximately 5 to 7 minutes. In contrast, in the past, paper-based surveys had to be sent to Kathmandu via pouch mail or hand-carried, which took days or even weeks. The use of CAPI reduced data collection time by one month compared to the previous survey completed by PAPI (from 6.5 months to 5.5 months). However, there were some security concerns with carrying these tablets and storing them especially in remote areas because some interviewers had to stay in community members' homes. Enforcing joint responsibility for theft of, or damage to, the tablet PCs among the interviewer teams helped to ensure security of the tablets during transport and storage. For example, interviewers were trained to lock and be aware of their tablet PCs at all times, even during meal and rest times (Paudel et al., 2013). And, in a Peruvian survey, handheld computers were inserted into a wooden and Styrofoam clipboard to shield them from possible damage and to conceal them (Bernabe-Ortiz et al., 2008).
  3.1.2 In cross-cultural surveys such as the World Mental Health (WMH) Initiative, some participating countries have been unable to implement technology-based survey instruments due to infrastructural constraints. However, the WMH Coordination Centre made the decision that those countries which can, should use technology (in this case, CAPI), as the advantages outweigh the methodological concerns of non-comparability. Other experimental studies have found few significant differences in survey estimates (Baker, Bradburn, & Johnson, 1995; Couper, 2000). The WMH Organization's current recommendation is to challenge where CAPI can and cannot be used. For example, in 2003, Columbia was able to

implement the WMH survey with great success using CAPI. Countries that used PAPI in the most recent WMH surveys expressed the wish that they had more strongly pursued CAPI, especially because of quality control and complexity of survey instrument (Pennell et al., 2008).

3.1.3 Researchers should be aware that mode differences can occur in unanticipated ways. In a meta-analysis of studies from the United States, Canada, the United Kingdom, and Italy comparing data from PDAs to PAPI, the results favor handheld computers over paper and pencil for data collection among study participants, but the data are not uniform for the different outcomes. Handheld computers appear superior in timeliness of receipt and data handling (four of four studies) and are preferred by most subjects (three of four studies). On the other hand, only one of the trials adequately compared adherence to instructions for recording and submission of data (handheld computers were superior), and comparisons of accuracy were inconsistent between five studies (Lane, Heddle, Arnold, & Walker, 2006).

3.1.4 The availability of information and communication technologies for direct data transfer has the potential to improve the conduct of research, and, especially, public health research, in resource-poor settings. Because of shortened data entry time in a vaccination survey in Zanzibar through use of CAPI, transition time to vaccination and subsequently to disease surveillance was shortened (Ali et al., 2010). As technology continues to evolve, research on its impact on survey data collection should continue.

3.1.5 In a study by Thriemer et al. (2012), a PDA-based survey in Tanzania resulted in an estimated 25% reduction in cost, compared to a paper-based survey. Elimination of questionnaire printing costs is even more significant if multiple languages/versions are needed in a country because multiple versions can be programmed into the platform (Onono et al., 2011). In another effort to reduce costs, researchers found that sending an excess supply of batteries to study sites helped decrease use of PAPI and its associated additional costs (Onono et al., 2011).

3.1.6 The use of technology can greatly increase the efficiency through which data from multiple data collection modes can be linked. Current Smartphone capabilities allow for scanning barcodes on respondent records, which has the potential to further effectively link data from multiple sources, such as completed surveys, signed letters of consent, medical charts, biomarker records, etc. (Aviles et al., 2007; Thriemer et al., 2012).

3.1.7 The use of CAPI platforms can extend potential working hours. Because CAPI can be used in low-light situations, interviewers can work during evening hours, otherwise a challenge with paper questionnaires in settings with frequent power outages.

3.2 In non-western settings, interviewers have generally reported a preference for CAPI instruments. Examples include the following from across the world:

3.2.1 CAPI was successfully implemented in a survey of malaria morbidity in Gambia, where interviewers reported a preference for CAPI over PAPI in terms of amount of work, number of errors, length of interviews, and ease of transport (Forster & Snow, 1992).

3.2.2 Handheld computers were used for a tobacco use survey in a hard-to-reach population in China where most interviewers stated a preference for handheld computers for future surveys (Wan et al., 2013).

3.2.3 In a survey in Zanzibar, acceptability of PDA use was high among staff not familiar with computers or PDAs and after an initial training period, none of the users was interested in returning to paper-based data entry (Thriemer et al., 2012).

3.2.4 In a survey in Bolivia, interviewers reported that using PDA to administer interviews stimulated their own interest in working on the survey (Escandon et al., 2008).

3.3 Use of technology has been well-received by respondents.

3.3.1 In a survey in Tanzania using PDAs, most respondents who expressed their opinions about the use of PDAs had something positive to say. For example, a 30-year old man with primary-level education said, "I was very happy to see a computer as it was my first time to see it. It simplified recording of our responses." An elderly man expressed his appreciation of having learnt what day of the week he had been born (Shirima et al., 2007).

3.3.2 In a Demographic and Health Survey in Nepal, respondents were curious about being interviewed using the tablet PCs. The interviewers perceived a high level of respect and enthusiasm from respondents, and they felt that respondents viewed them as technical employees with higher education. This was an unanticipated, but encouraging finding, especially because of respondents' limited exposure to computers. However, at the same time and in certain cultures, the use of technology can raise suspicion among respondents and although acceptability of the tablet PCs was high, there were a few cases of skepticism. As part of the informed consent process, respondents were informed that the interview would not be video- or audio-recorded and that the recording feature had been disabled on the tablet PCs. However, a few respondents were still concerned (Paudel et al., 2013).

3.3.3 Analyses from a survey in rural south Kenya using PDAs found a reduction in refusals, attributed to the perception of respondents that the PDA was more secure (Onono et al., 2011).

3.3.4 Due to the increasing use of mobile phones and other similar technologies in day-to-day life, operating a computerized questionnaire on a handheld device might be more familiar to respondents with little or no experience in the use of computers (Ali et al., 2010).

3.4 Allow for adequate project preparation before beginning fieldwork.

3.4.1 Do not underestimate the additional time needed for preparation for both initial adoption and continued use of technology. In a survey in Burkina Faso, researchers reported underestimating the amount of work required to program questionnaires, and as a result failed to maximize the use of some of the available options for input checking and other real-time quality control procedures. Village names, for example, were implemented as a text-entry field, but would have been better as a drop-down list to avoid ambiguities of spelling, etc. Combinations of input checks, plus quality control measures at the stage where data were downloaded to portable computers in the field, should have picked up concerns at an earlier and remediable stage (Byass et al., 2008).

3.4.2 Having local trained personnel is essential. Using a “train the trainers” model, technical and supervisory staffs in a public health survey in China were able to develop the questionnaire and complete the programming with minimal assistance from technical experts from the coordinating center. When problems occurred, the Chinese technical experts could then provide immediate technical guidance and trouble-shooting to interviewers and other staff (Wan et al., 2013).

3.4.3 It can be difficult to repair equipment in country. Aviles et al., (2007) recommend the implementation of preventative maintenance program.

3.4.4 Consent letters mentioning the use of technology can be helpful in reducing non-response. In a survey in Tanzania using PDAs, most respondents said that they had noted the PDA after its mention in the consent letter. Several interviewees appreciated the interviewer having introduced them to the technology during the consent procedure (Shirima et al., 2007).

3.5 When using CAPI, particularly with interviewers previously unfamiliar with computerized instruments, consider the following with regards to interviewer recruitment, training, and management.

3.5.1 Experience suggests that interviewers with little education and no experience in the use of a computer are easily able to use handheld devices for survey administration (Ali et al., 2010). With increasing use of mobile phones and other similar technologies, operating handheld devices, downloading data, and recharging batteries are becoming increasingly familiar concepts.

3.5.2 Although use is increasing, however, plan for adequate time for interviewer training. In a survey in Bolivia using PDAs, interviewers wanted additional practice time because of previously limited experience with the technology, and particularly more instruction on the use of a stylus as keyboards on handheld devices can be cumbersome (Escandon et al., 2008).

3.5.3 Analyses of inter-observer accuracy and performance revealed a considerable range in a survey in Burkina Faso. Some interviewers clearly worked faster with the PDAs than others, though these were not necessarily those who covered the greatest number of households per day worked. However, those who carried out interviews relatively quickly were generally also those who made the least input errors. In surveys of this kind, where competence in local languages is an important factor, there are often not many options in terms of who can be recruited as interviewers (Byass et al., 2008).

3.5.4 Training on proper handling and care of the equipment is also very important, particularly in a rural context where the equipment has to be transported through rough terrain, the power supply is not stable, and unexpected rain is a concern. In the DHS survey in Nepal, teams were provided with generators, rain shields, umbrellas, and several other items to manage these challenges. Enforcing joint responsibility for theft of, or damage to, the tablet PCs among the interviewer teams helped to ensure security of the tablets during transport and storage. With proper care and maintenance, tablet PCs (and portable generators) can be reused in future surveys, resulting in additional cost savings over the long term (Paudel et al., 2013).

3.6 The use of CAPI is not without its technologically-related challenges.

3.6.1 Project staff should be aware of the possibility for corrupted date/time stamps because of equipment malfunction. In a survey using PDAs in Kenya, researchers found that if the PDA lost power, it automatically reset the clock, which had effects on pregnancy data that was collected. Particular caution should be used if data is time-sensitive as in this case (Onono et al., 2011).

3.6.2 If using CAPI, the concurrent use of paper files for portions of the survey can lead to logistical challenges. Diero et al. (2006) used PDAs to follow patients who visited a rural Kenyan health center. The data entry program did not allow for entry of text field notes by the research assistants, who had to use a paper notebook for such notes. This can cause a disconnect between these text notes and the patient data to which they referred.

3.7 Researchers have used several methods for maintaining respondent confidentiality and ensuring data security when using CAPI.

3.7.1 Data can be copied and automatically saved to a SD card, after which interviewers are unable to retrieve or change an entry, with no record of the entry was retained on the PDA (Onono et al., 2011; Shirima et al., 2007). In case of equipment loss, it is then impossible to access and see the data on the SD card without a password and the requisite software. In a survey in Kenya, when one PDA was stolen in political violence, two interviews were lost on the SD card, but respondent confidentiality was maintained because of security protocols in place (Onono et al., 2011).

3.7.2 In a survey in Tanzania, data were downloaded to the laptop computers and daily summary reports produced to evaluate the completeness of data collection. Data were backed up at three levels: (i) at the end of every module, data were backed up onto storage cards in the PDA; (ii) at the end of every day, data were downloaded to laptop computers; and (iii) a compact disc (CD) was made of each team's data each day (Shirima et al., 2007).

3.8 As smartphones become more ubiquitous in daily life, their use in survey research is expected to increase. Findings from a recent survey examining the usability of smartphones versus tablets in Kenya generally favor tablets over smartphones (Hughes & Haddaway, 2014). Highlights from the study include:

3.8.1 Confidence and comfort in typing dependent on past experience with device and touchscreens

3.8.2 Interviewers felt more likely to accidentally select options on phones
3.8.3 Interviewers admitted to not scrolling completely through questions/responses on phones
3.8.4 Interviewers felt more professional with tablets
3.8.5 Interviewers felt safer with phones because of the smaller size compared to the tablets, which attracted unwanted attention
3.8.6 Smartphones were associated with more typing error
3.8.7 Long open-ended questions and long numeric strings are difficult

4. **If the questionnaire includes items of a sensitive nature, consider administering these questions in a self-administered module during the face-to-face interview.**

Rationale

Evidence suggests that increasing privacy during an interview can improve the accuracy of reporting such topics in surveys (Turner et al., 1998; Turner et al., 2002; Krawczyk et al., 2003), but achieving privacy in non-western settings varies considerably between countries (Mneimneh, 2012). For a face-to-face interview, consider administering the sensitive sections in a self-administered questionnaire (SAQ). Research indicates that respondents in an interviewer-administered, non-private setting tend to misreport information perceived to be sensitive. For example, respondents might underreport undesirable or private information such as drug use or illegal status; and they might over-report desirable information such as voting.

Many surveys include potentially sensitive questions about both respondent behavior and attitudes concerning such topics as sexual behavior and contraceptive use, substance abuse, violence, and politics. And, in non-western settings, these delicate topics are particularly susceptible to social desirability bias. However, asking sensitive questions in an SA format) has the potential to decrease bias and achieve more accurate reporting.

Procedural Steps

4.1 Assess the literacy of the target population and choose the most appropriate instrument for the SAQ.
4.1.1 The SAQ can be a paper questionnaire given to the respondent to self-complete. The paper-based SAQ should not have complex skip patterns, and the target population should have adequate literacy levels.

4.1.2 The SAQ can take the form of computer-assisted self-interviewing (CASI), where respondents use a technology platform (i.e., a laptop, tablet, smartphone, etc.) and complete the entire questionnaire, or a specific section of the questionnaire, independently. The technology therefore facilitates the administration of a complex instrument, much like CAPI facilitates administration for the interviewer.

4.1.3 Audio-CASI (A-CASI) has the advantages of CASI, but can be particularly helpful in low-literacy settings. In A-CASI, respondents listen to an audio track recording of each survey question using a headset and move through the survey at their own pace. If illiterate, survey respondents can be instructed to push color-coded buttons on a touch screen or mini-keyboard, or have graphical representations of answer categories to indicate their response to each question (see Instrument Technical Design, Appendix F for an example).

4.1.4 If using A-CASI, assess whether the setting would benefit from gender-matching in terms of the audio voice used. That is, if the recording presented to female respondents should be a female voice, while male respondents are presented with a recorded male voice.

4.2 When designing an SAQ instrument, consider the following:

4.2.1 Be mindful of survey length. Longer surveys administered using an SAQ mode may have more missing data both because of lack of interviewer probing and lack of the pressure respondents feel to cooperate with the interviewer (Hewett, Erulkar, & Mensch, 2004a).

4.2.2 Develop interviewer instructions for explaining the SAQ to the respondent.

- The detail of instructions will differ by mode, with CASI and A-CASI necessitating more explanation than a paper-based SAQ, particularly in low-literacy settings.
- If an SAQ is utilized for reasons of increased respondent confidentiality, then this rationale should be explained to respondents.
- Develop a protocol for interviewer behavior during the interview, particularly concerning the extent to which interviewers should be encouraged to help the respondent or otherwise interact with the respondent. All interactions should be documented.
- Consider adding questions at the end of the interview to assess respondents’ perceived ease of use, privacy, and truthfulness.

4.2.3 When using CASI and A-CASI modes, attention to details that facilitate the respondent experience can lead to increased data quality.

- Consider disabling the screen saver and monitor power-saver settings on the device so that screens do not go blank if a participant takes additional time to answer a question (NIMH, 2007).
- Graphical and/or audio representations of the response process can help guide the respondent through the interview. In a survey in India using A-CASI, the entry of a response was marked by the change in the color of the corresponding response bar on the screen to grey, along with a "beep" sound. A "Thank you" screen indicated the end of the survey (Bhatnagar, Brown, Saravanamurthy, Kumar, & Detels, 2013).
- If a participant did not answer a question after approximately 60 seconds, consider repeating the question and/or programming additional text. The additional text can be programmed to appear encouraging participants to answer the item(s) in a truthful manner (NIMH, 2007).
- If used, the keyboard should be user-friendly. Keyboard options can be limited to responses (e.g. YES, NO, and numbers) and larger color-coded keyboard keys could be used. Additional keyboard shortcuts to replay questions can also be marked.
- Text on the computer screen should be large enough to be easily legible for respondents.
- In an A-CASI survey in India, neither the question nor the response texts were displayed on the screen to ensure privacy and confidentiality for the respondents (Bhatnagar et al., 2013).
- Touchscreens on A-CASI instruments can be particularly helpful for less-educated populations (Lara, Strickler, Olavarrieta, & Ellertson, 2004).

4.3 Additional technologies for SAQ mode in a face-to-face interview continue to emerge, including video-computer-administered self-interview (V-CASI) (Kissinger et al., 1999; Krysan & Couper, 2003). If planning to use an SAQ, investigate the most recent literature available for further guidance.

Lessons Learned

4.1 The use of novel technology, particularly in non-Western settings, can motivate respondents to participate.

4.1.1 In a comparison of paper SAQ vs. self-administered PDA questionnaires on sexual behavior, South African adolescents reported more favorable attitudes toward the PDA mode (Jaspan et al., 2007; Seebregts et al., 2009).

4.1.2 End-of-questionnaire items measured high respondent-perceived truthfulness in a South African survey about sexual behavior (Beauclair et al., 2013) and greater preference for A-CASI compared to other modes, primarily because of perceived increased confidentiality and privacy, as well as the novelty of technology (van de Wijgert, Padian, Shiboski, & Turner, 2000; Bhatnagar et al., 2013; Lara et al., 2004; Hewett, Mensch, & Erulkar, 2004b; Gutiérrez & Torres-Pereda, 2009).

4.1.3 However, if technology is unfamiliar to the population, it may cause concern about the project activities. In a study using A-CASI in rural Kenya, interviewers and supervisors reported that the presence of computers heightened the animosity and opposition of the community to the project activities. Rumors spread that the survey was the work of devil worshipers and that interviewers were collecting the names of adolescents who would later be abducted. Many respondents believed that the computers collected information for the government. Also, respondents were angry that expensive equipment was brought into resource-starved community during a time of drought. Misinformation spread throughout the region before interviewers even entered some sampling units. Some residents thought that the computer was having a "conversation" with the respondents, despite insistence that the computer voice was taped. And, in the initial A-CASI protocol respondents' answers were read back to them after each question for verification, a protocol which needed to be discontinued because some respondents perceived the computer to be "talking to them", resulting in decreased perceptions of confidentiality (Hewett et al., 2004a).

4.2 Use of an SAQ mode can impact the length of time needed for interviewer administration depending on setting and demographics. The HIV/STD Prevention Study found that surveys using A-CASI generally took longer to administer than CAPI in China, Peru, India, and Zimbabwe. However, A-CASI took less time in Russia, where the participants had more exposure to technology and were of a younger age (NIMH, 2007).

4.3 There is evidence that using A-CASI is feasible in non-western settings.

4.3.1 The NIMH Collaborative HIV/STD Prevention Trial Group conducted a feasibility study comparing results from surveys using CAPI and A-CASI in China, India, Peru, Russia, and Zimbabwe (NIMH, 2007). Despite the varying levels of literacy and exposure to computers by country, most study participants reported that it was easy to enter their answers into the computer, that they felt comfortable doing so, and that they preferred the computer rather than an interviewer for answering questions about topics such as sexual behavior and drug and alcohol use, or had no preference. Most participants gave the same responses on both their A-CASI and CAPI interviews.

4.3.2 While A-CASI has generally been feasible in non-western settings, however, ease of use can vary by socio-demographic characteristics.

- Older and unemployed respondents report increased difficulty with A-CASI (Beauclair et al., 2013), as do less educated respondents (van de Wijgert et al., 2000).
- Women with little education (primary school or less) had considerably more problems using the computer keyboard, reading the computer screen, and correcting mistakes than women in higher educational groups (also Gutierrze & Torres-Pereda, 2009).

4.4 In regions where there are multiple languages and dialects, use of A-CASI can facilitate the interview process. A completely self-administered questionnaire can ease the logistical challenges in the field of matching a respondent with an interviewer who has the necessary language capabilities.

4.5 Use of A-CASI can lead to improvements in data quality.

4.5.1 A-CASI is a more standardized method of assessment than CAPI. Using CAPI, interviewers may use probes beyond the standard set even though they are instructed not to do so.

4.5.2 Unlike a paper-based SAQ, use of A-CASI leads to fewer data entry errors and missing data because the skip patterns are programmed into the computer and are executed as the interview is administered (van Griensven et al., 2006; Langhaug et al., 2011).

4.5.3 In a comparison study on topics related to HIV/AIDS in three cities in Vietnam, respondents assigned to A-CASI had lower item refusal rates than those assigned to a face-to-face interview or a paper-based-SAQ (Le & Vu, 2012).

4.5.4 There is evidence that using A-CASI has the potential to improve data quality through the reduction of missing data. Studies of mode differences in South Africa and Thailand have found that those respondents assigned A-CASI had less missing data than those assigned to a paper-based SAQ (Jaspan et al., 2007; van Griensven et al., 2006).

4.5.5 A short respondent-training session prior to the administration of A-CASI can improve data quality. A survey of young women in Malawi utilized headphones and an external color-coded minikeypad, with a red key to replay the question, a green key to go on to the next question, and a yellow key to skip a question. For dichotomous questions, respondents were instructed to press 1 for yes and 2 for no. Prior to the A-CASI main survey, each respondent completed three "practice" questions to evaluate her understanding of the interview process, for example, "Are you a male or a female?" For each practice question, the correct answers were previously entered by the interviewer to serve as a check against the respondent's entry. Respondents were not able to proceed to the main interview until they were able to answer all three practice questions correctly (Mensch, Hewett, Gregory, & Helleringer, 2008).

4.6 On the other hand, use of A-CASI can bring challenges to data quality as a result of decreased interviewer interaction.

4.6.1 Respondents may not understand skip patterns or other aspects of the survey but are reluctant to ask the interviewer for direct assistance given the hands-off nature of A-CASI (Lara et al., 2004).

4.6.2 An SAQ on sensitive topics may also lead to reluctance to engage the interviewer in a related question about completing the survey because, in using A-CASI, there can be an underlying perception that the topic is too delicate to discuss outright.

4.6.3 A study using SAQ in Tanzania found that about 7% of respondents selected only the first or the last response categories in a section for which such a response pattern would be inconsistent. This bias was associated with females, those less educated, and those more geographically remote (Plummer et al., 2004a).

4.6.4 Mode can impact data quality because of inconsistency in editing. A comparative survey of young women in Malawi found more consistent reporting in the face-to-face mode than in the A-CASI mode. Researchers speculated that, contrary to protocol, the face-to-face interviewers may have edited the questions for consistency post-hoc, whereas such editing was not possible in the A-CASI mode by respondents.

4.7 A-CASI is often used with the a-priori expectation that privacy for the respondent will result in increased reporting of more sensitive behaviors, with the related implication that this reporting is indeed accurate (Groves et al., 2009; Couper, 2005). However, meta-analyses using data from non-western settings are inconclusive on whether SAQ modes increase accuracy of sensitive behaviors.

4.8 Results from a meta-analysis of face-to-face and A-CASI modes in studies on sexually transmitted infections and associated behaviors in Brazil, Vietnam, Thailand, Kenya, India, Russia, Zimbabwe, Malawi, China, and Tanzania demonstrate that overall, A-CASI methods are not consistently associated with a significant increase in reporting of sensitive behaviors, but trends can be seen in certain contexts. In general, increased reporting in A-CASI has been associated with region (Asia), setting (urban), and education (secondary education) (Phillips, Gomez, Boily, & Garnett, 2010).

4.9 In contrast, a meta-analysis by Langhaug, Sherr, and Cowan. (2010) of 26 studies in developing countries on sexual behavior demonstrated that, in general, A-CASI can significantly reduce reporting bias. The results of this review as well as findings from other researchers (cited below) show that the relationship and success of novel interviewing methods has proved complex in a low- and middle-income country context and researchers should be aware of the mode differences that can result, depending on the study topic and social context (NIMH, 2007; Jaspan et al., 2007; Rathod, Minnis, Subbiah, & Krishnan, 2011; Langhaug et al., 2011; Lara et al., 2004; Mensch, Hewett, & Erulkar, 2003; Mensch et al., 2008; Mensch et al., 2011; Hewett et al., 2004b; Plummer et al., 2004a; Plummer et al., 2004b; Potdar & Koenig, 2005; Minnis et al., 2009; Jaya, Hindin, & Ahmed, 2008).

Appendix A

Cover-sheet (example from the University of Michigan's Institute for Social Research)

SPACE CAN BE USED FOR ADMINISTRATIVE INFORMATION

THIS SPACE USED TO AFFIX LABEL CONTAINING SAMPLE INFORMATION	THIS SPACE USED TO AFFIX LABEL CONTAINING INTERVIEWER INFORMATION

THIS SPACE RESERVED TO RECORD ADDITIONAL INFORMATION ABOUT THE LOCATION OF THE SAMPLE CASE IF NECESSARY (e.g., landmarks or housing unit description if address is unavailable, etc.).	Final Result Code: ___________ Date of Final Result (dd/mm/yyyy): __ / __ / _____ Length of interview: ___________ Total calls: ______ Length of edit:

Appendix B

Household enumeration table (example from the University of Michigan's Institute for Social Research)

	HOUSEHOLD ENUMERATION					RESPONDENT SELECTION		
	11 a. Household Member's First Name	11 b. HH Member's Relationship to Informant	11 c. Sex	11 d. Age	11 e. Language Spoken	11 f. Eligible	11 g. Person Number	11 h. Selected R
MALES			M					
			M					
			M					
			M					
			M					
			M					
			M					
FEMALES			F					
			F					
			F					
			F					
			F					
			F					
			F					

Instructions for household enumeration table

Column 11a (Household Member's First Name): List all members of the household, beginning with the informant. Note that **males are listed in the upper portion** of the table and **females in the lower portion**.

Column 11b (Household Member's Relationship to Informant): Record each household member's relationship to the informant (e.g., husband or wife, son or daughter, mother or father, brother or sister, friend, etc.).

Column 11d (Age): Record each household member's age.

Column 11e (Language Spoken): This column may or may not be included, depending upon the study requirements.

Column 11f (Eligible): Place a check mark in this column if, based upon the information in columns 11a-11e, the household member meets the eligibility criteria for the study.

Column 11g (Person Number): Assign a sequential number to each eligible household member. Begin by numbering **eligible males from oldest to youngest**, continue by numbering **eligible females from oldest to youngest**.

Column 11h (Selected R): Count the number of eligible persons in the household. Find that number in the Kish table in the "If the Number of Eligible Persons is:" column. The selected respondent will be the household member with the "Person Number" corresponding to the "Interview the Person Numbered:" column in the Kish table.

Appendix C

Random within house selection techniques

Kish (1949) selection tables

SELECTION TABLE A

If the number of eligible persons is:	Interview the person numbered:
1	1
2	1
3	1
4	1
5	1
6 or more	1

SELECTION TABLE B_1

If the number of eligible persons is:	Interview the person numbered:
1	1
2	1
3	1
4	1
5	2
6 or more	2

SELECTION TABLE B_2

If the number of eligible persons is:	Interview the person numbered:
1	1
2	1
3	1
4	2
5	2
6 or more	2

SELECTION TABLE C

If the number of eligible persons is:	Interview the person numbered:
1	1
2	1
3	2
4	2
5	3
6 or more	3

SELECTION TABLE D

If the number of eligible persons is:	Interview the person numbered:
1	1
2	2
3	2
4	3
5	4
6 or more	4

SELECTION TABLE E_1

If the number of eligible persons is:	Interview the person numbered:
1	1
2	2
3	3
4	3
5	3
6 or more	5

SELECTION TABLE E_2

If the number of eligible persons is:	Interview the person numbered:
1	1
2	2
3	3
4	4
5	5
6 or more	5

SELECTION TABLE F

If the number of eligible persons is:	Interview the person numbered:
1	1
2	2
3	3
4	4
5	5
6 or more	6

The tables provide unbiased estimates by giving each respondent a weight based on the number of adults in the household. This guarantees that the selection within a household is random for a combined total random sample across the housing units (addresses) that were randomly selected in the first place.

Instructions for Kish tables

1. Assigning Kish tables to the sample file: one of the twelve tables is randomly selected and assigned to the first line in the sample file. The series of twelve is then run through twice, assigning tables to the sample lines. Then again a table is randomly selected and the series is run through twice. This procedure is repeated until all sample lines have an assigned Kish table.
2. Household listing: a household listing of eligible adults (age 18 and over) who reside in that household is taken at each of the sample addresses. Usually the males are listed first in order of decreasing age, and then the females in the same order.
3. Using Kish tables: the table assigns a number to each member of the household listing. Sample Kish tables are shown above. In the first column the interviewer would circle the total number of eligible persons. The corresponding number in the second column of the Selection Table denotes the person selected to be interviewed.

Data Collection: Telephone Surveys

Julie de Jong, 2016

Introduction

If researchers wish to have survey interviews carried out by an interviewer, but face-to-face interviews are not possible, conducting interviews via telephone either through a landline or mobile telephone can be an alternative. Multinational, multiregional, and multicultural survey ("3MC" surveys) use different standards to determine whether telephone penetration is adequate in a study country. For example, the Gallup World Poll generally uses a telephone survey only in countries where telephone coverage represents at least 80% of the population (Gallup, 2015). Telephone interviews are generally less costly than face-to-face methods, and can be completed in a shorter amount of time. However, response rates are generally lower and depending on the available sampling frame for a country, a rigorous telephone-administered sample design can be difficult to develop. See Sample Design for a discussion of the challenges and limitations of a telephone-based frame and sample design.

As discussed in Data Collection: General Considerations, 3MC surveys sometimes employ mixed modes, depending on individual country constraints. However, it is important to note that mode effects may occur if the survey is carried out by telephone in some countries and face-to-face in others (see Study Design and Organizational Structure for discussion on mode effects).

Virtually all questionnaires administered by interviewers in telephone surveys are completed using an electronic computer-based instrument to record survey responses. This data collection mode is most commonly referred to as computer-assisted telephone interviewing (CATI). These guidelines assume that the interviewer will be using a computer-based instrument and will refer to the mode as CATI.

For additional discussion on the advantages and disadvantages of telephone surveys, see Study Design and Organizational Structure.

Guidelines

Goal: To achieve an optimal cross-cultural data collection design by maximizing the amount of information obtained per monetary unit spent within the allotted time, while meeting the specified level of precision and producing comparable results, within the context of a telephone survey.

1. **Develop the computer-based system(s) that the interviewers will use to administer telephone interviews.**

Rationale

Interviewers can conduct telephone interviews from either a central location or remotely. Software systems can be used to distribute sampled telephone numbers, to dial telephone numbers, to manage call records, and to record survey data. When using CATI, it is crucial to design and implement a system that interviewers can use to reliably collect survey data.

Procedural steps

1.1 Decide whether interviewers will work in a centralized and/or decentralized location.
   - 1.1.1 Many survey research firms conducting telephone interviews maintain a “telephone lab,” which is a central calling center where center supervisors oversee a variable number of interviewers. Each interviewer has access to the electronic instrument and records responses directly in the electronic file. Interviews can be monitored in real time.
   - 1.1.2 Sometimes interviewers work from other locations while having access to the electronic system set up by the survey research firm.

1.2 Develop a system and protocol for sample release management, including how cases will be transferred between interviewers when necessary.

1.3 Develop a protocol for dialing sampled telephone numbers. Some projects may use CATI systems that can dial telephone numbers automatically, while other projects may elect to have interviewers dial telephone numbers manually. In some countries it is against the law to use automation to dial specific types of telephone numbers (e.g., in the United States, it is illegal to use automation to dial mobile numbers). If using automation, be familiar with the local laws about its use.

1.4 Consider the cost structure for telephone calls in each study country. In the United States, respondents are responsible for the cost of incoming telephone calls on mobile telephones. However, in the Persian Gulf, for example, there is no charge and interviewers based in Nepal were able to telephone Nepali migrant workers living in Gulf countries for a migration survey without any cost to the respondents (Ghimire, Williams, Thornton, Young-DeMarco, & Bhandari, 2013).

1.5 Decide which telephone number and name will be displayed to the respondents in the caller ID, and whether the telephone number should be available if people call back the number.

1.6 Develop an electronic survey instrument used to record survey responses. There are numerous CATI software packages. However, it is also possible to use a web-based survey instrument, which may not be as suitable for more complex projects but is less expensive. Electronic survey instruments in a telephone survey share many of the same requirements as electronic survey instruments administered in the face-to-face mode. For in-depth discussion of these elements, see Guideline 3 in Data Collection: Face-To-Face Surveys.

Lessons Learned

1.1 While survey mode can affect survey responses, studies are not unanimous in the direction of the effect observed.
   1.1.1 A survey of HPV awareness and knowledge, including sexual behavior, was conducted in Singapore, with half participating via CATI and half through an interviewer-administered face-to-face interview. Few differences between survey modes were found in the information disclosed (Smith et al., 2009).
   1.1.2 A study in India evaluating accuracy of health data collection through several different interfaces found that telephone interviewing had greatest accuracy in phone interviews when compared to electronic forms on PDAs and text messaging (Patnaik, Brunskill, & Thies, 2009).

1.2 CATI can be particularly useful in a panel study setting, especially when there is frequent contact with respondents. Experiences vary by country, however.
   1.2.1 In a study of farmers in Tanzania, researchers gave respondents pre-paid mobile phones for the duration of the field period so that they could receive a phone call from an interviewer and complete a survey every three weeks over a ten-month period, resulting in a high quality dataset (Dillon, 2012).
   1.2.2 Researchers distributed mobile phones to female sex workers in India for use in a diary study on sexual behavior, which resulted in high response rates and high-quality data (Bradley et al., 2012).
   1.2.3 Researchers on a panel study in South Sudan using CATI found that response rates were affected by irregular fluctuations in the mobile network (Demombynes, Gubbins, & Romeo, 2013).

1.3 Beyond the traditional CATI mode, interviewing via text message has been recently used. In this mode, the interviewer sends individual survey questions by text to the respondent, who sends his or her responses back by text to the interviewer (West, Ghimire, & Axinn, 2015; Lau, Lombaard, Baker, Eyerman, & Thalij, 2016).

2. Train interviewers on interviewing strategies specific to telephone interviewing.

Rationale

The nature of the interaction between the interviewer and the respondent depends on the mode of data collection. Some interviewing strategies that are accessible in a face-to-face mode, such as interpretation of body language, are not possible to implement over the telephone, contributing in part to lower response rates and potential for non-response bias. However, there are certain telephone-specific strategies that researchers can introduce to assist interviewers in completing telephone interviews.

Procedural steps

2.1 Consider the social context of the study country when hiring interviewers to administer a telephone survey, and whether selection of interviewer based on gender or other characteristics will affect response rates. See Lessons Learned 2.1 below as well as Interviewer Recruitment, Selection, and Training for additional discussion of interviewer recruitment considerations.

2.2 Develop an introduction appropriate for the interviewer to read upon contact with the respondent.

2.2.1 The introduction is especially important and may differ depending on cultural norms, and the way the opening unfolds between the interviewer and respondent may have significant implications for both survey non-response and data quality (Couper & Groves, 2002). The context of the interview can dictate identification procedures and pace of interview.

2.2.2 Establishing and maintaining rapport is especially important in achieving a telephone survey. Particular care should be taken in the translation stage to ensure an interviewer script that does not violate cultural norms involving politeness and linguistic encoding of status and social distance (Kleiner & Pan, 2006).

2.2.3 The introduction can be particularly critical in achieving cooperation in some countries. Previous respondent exposure to the telephone as a survey mode can differ across countries, and there can be discomfort in sharing personal information over the phone (Hughes, 2004).

2.2.4 In countries where there are linguistic differences depending on actors' social status, translations must also recognize that interviewers and respondents are strangers and cannot rely on visual cues to establish social distance and appropriate linguistic level, necessitating the opportunity for some social interaction at the beginning of the survey to establish such social distance.

Lessons Learned

2.1 Gender norms of the study country can have a significant impact on response rates in CATI surveys.

2.1.1 In France, researchers have found that female interviewers generally have higher refusal rates in telephone surveys (Verger, Baruffol, & Rotily, 2001).

2.1.2 In Nepal, a highly gendered society, women generally prefer to speak to other women, and men to men, even over the telephone. However, in a CATI survey using Nepali-based interviewers contacting (mostly male) Nepali migrant workers in Persian Gulf countries, researchers obtained high response rates using predominantly female interviewers, because of the cultural perception that women would not call a male unless it was an important matter (Ghimire et al., 2013).

2.1.3 There is also anecdotal evidence that male respondents in the highly gendered countries in the Middle East are more likely to participate in a telephone survey when contacted by a female interviewer.

2.2 Immediate identification by name is standard telephone practice in the United States, but is uncommon in China (Kleiner & Pan, 2006).

2.3 Acceptable pace of the interview introduction can vary across even otherwise similar cultural contexts. For example, an examination of reaction to phone calls in Hong Kong and Beijing found that Beijing residents were more resistant to a fast-paced, business-like telephone conversation when compared to those from Hong Kong (Pan, Scollon, & Scollon, 2002). Similarly, a comparison of Greeks and Germans showed that Greeks prefer social interaction before reaching the main point of a telephone conversation, while Germans prefer to discuss the main point immediately (Pavlidou, 1994).

2.4 Acquiescence bias differs across cultures and can be particularly problematic in a telephone survey where otherwise difficult issues can be exaggerated. For example, in many Asian cultures, people tend to avoid "no" answers to yes/no questions, particularly when there is an asymmetrical relationship between speakers as in a survey interview (Kleiner & Pan, 2006).

2.5 Introductory scripts can differ dramatically across cultures. For example, in Chinese, the use of expressions like "please" and "thank you" are not normally used in daily conversation and imply a large social distance between speakers. The mandated repetitive use of such words in a survey among Chinese speakers would be detrimental, particularly in a telephone survey where rapport is especially important, in sharp contrast to a survey in American English, where such phrases are acceptable and expected (Pan et al., 2002).

3. Decide whether a subset of survey questions would best be collected in a self-administered section of the interview.

Rationale

Interviewer-administered telephone interviewing is subject to social desirability biases similar to those in face-to-face interviewing. Interactive Voice Response (IVR) is a telephone mode where the computer plays recordings of the questions over the telephone to respondents who then respond by using the keypad of the telephone or saying their answers aloud. IVR can be used as a self-administered mode (SAQ) to administer a portion of an interview, otherwise conducted by CATI, which is particularly sensitive in nature and where accuracy might improve without the presence of an interviewer. It can also be used exclusively as a self-administered mode (SAQ), with the computer automatically telephoning the respondent and then completing the questionnaire (see Data Collection: Self-Administered Surveys) for further discussion of IVR in a completely self-administered mode.

Procedural Steps

3.1 Design the IVR system so that it is technically well-integrated into the CATI system in use by the project and that switching from the CATI to the IVR system is straightforward for the interviewer.

3.2 Decide whether to program the IVR system as touchtone, voice input, or a combination of the two.

3.2.1 When deciding on the programming, consider the target population. Studies in rural India and Botswana found that respondents with less education and lower literacy do better with touchtone, and cited privacy for touchtone preference as well (Kuun, 2010; Patel et al., 2009).

3.2.2 A study in Pakistan found that a well-designed speech interface was more effective than a touch-tone system for respondents regardless of literacy level (Sherwani et al., 2009).

3.3 Devote sufficient time to the development of a high-quality IVR system to maintain respondent interest and continued cooperation.

3.3.1 The IVR system must have a high quality recording, as the respondent is likely to break off the survey if quality is poor.

3.3.2 See Oberle (2008) for a guide to the development of an IVR system and the associated speech characteristics which need consideration.

Lessons Learned

3.1 Consider the voice used for recording.

3.1.1 In a health helpline project in Botswana, researchers employed a well-known local actress for the IVR recording, and users reacted very positively (Kuun, 2010).

3.1.2 Depending on the social context, using an IVR recording of a male for male respondents and of a female for female respondents may elicit more accurate reporting, particularly of sensitive information.

3.2 Plauche, Nallasamy, Pal, Wooters, and Ramachandran (2006) developed an innovative approach to the challenge that dialectical variation and multilingualism poses to speech-driven interfaces for IVR in India, applicable to other settings as well. In their approach, people from specific villages are recorded during interactions, and their speech is semi-automatically integrated into the acoustic models for that village, thus generating the linguistic resources needed for automatic recognition of their speech.

3.3 A survey of teachers in Uganda resulted in a number of useful considerations when designing an IVR system to improve response rates and data quality (Lerer, Ward, & Amarasinghe., 2010).

3.3.1 The IVR call began with the immediate information that “This is a recorded call from Project X. You are not talking to a real person.”

3.3.2 The IVR call provided very specific instructions about whether to use keypad or to speak.

3.3.3 Respondents were initially confused by the automation of the IVR system. Researchers had better results when using a chime to get respondents’ attention before the automated voice gave instructions.

3.3.4 Leveraging conversational and turn-taking conventions of normal conversation in the IVR system lead to more success than detailed instructions in eliciting desired user behavior.
3.3.5 An IVR system which projected a loud voice, with prompts recorded as if the speaker were using a poor cell connection, resulted in a survey that was easier for respondents to follow.
3.3.6 When producing the IVR recording, use slow speech to get slow speech – respondents will emulate the voice, and resulting data will be easier to understand.
3.3.7 The IVR recording included 3 seconds of silence before the recorded speakers says “thank you” and moves onto next question, which was reported as well-received by respondents.

Data Collection: Self-Administered Surveys

Julie de Jong, 2016

Introduction

Fully self-administered questionnaires (SAQ) are not as common as interviewer administered surveys in the context of multinational, multiregional, and multicultural surveys ("3MC" surveys). However, as surveys become more costly to administer using interviewers, whether face-to-face or by telephone, more researchers are considering SAQ modes. SAQ modes include mail surveys, web surveys, and Interactive Voice Response (IVR) surveys conducted via telephone.

An important element of the self-administered mode is that there is, by definition, no interviewer involved. As discussed in detail in Survey Quality, interviewer error can contribute significantly to total survey error. In removing the interviewer from the equation, survey quality can improve. This may be particularly true if the survey topic is sensitive. Self-administered modes can also be effective when the when privacy during the survey interview is difficult to obtain.

However, the absence of an interviewer also demands a carefully designed survey instrument that is easy for the respondent to complete. Because there is no interviewer present, there is no one to assist the respondent in understanding instructions or to provide encouragement to complete the questionnaire. Differences in literacy levels among countries should also be considered in the questionnaire design phase of self-administered instruments (see Questionnaire Design for further details).

In addition, because of the lack of interviewer-respondent interaction, non-response is more difficult to assess and it is a challenge to disentangle the effects of noncontact, refusal, and a poor sampling frame. For example, non-response to a mail survey may result from misdirected mail that never arrived at the sample respondent's house, misplaced mail within the respondent's house, initial willingness to complete the survey but subsequent forgetfulness, unwillingness to complete the questionnaire (i.e., a refusal), or any number of issues. And, in a multi-person household, it may be impossible to identify who the actual respondent was. Therefore, when designing an SAQ, it is crucial to implement strategies to maximize survey quality.

For further discussion on the advantages and disadvantages of self-administered surveys, see Study Design and Organizational Structure. For additional information on sample design and related challenges for self-administered modes, see Sample Design].

Guidelines

Goal: To achieve an optimal cross-cultural data collection design by maximizing the amount of information obtained per monetary unit spent within the allotted time, while meeting the specified level of precision and producing comparable results, within the context of a self-administered survey.

1. **When a mail survey using a paper-based instrument will be sent to respondents, develop the questionnaire and protocols with consideration that the survey must be straightforward for respondents to self-administer.**

   Rationale

   Concerns about response rates, length of surveys, and quality of data have all resulted in a reduction in the use of mail surveys in recent years. However, Dillman and others argue that high quality mail surveys, with close attention to detail, can result in accurate data (Dillman et al., 2007; Dillman, Smyth, & Christian, 2014). The mail survey is becoming more widespread as the cost of interviewer-administered surveys increases. If a mail survey is the chosen mode of data collection, consider the following steps when developing the instrument.

   Procedural Steps

   1.1 Assess the postal system in the study country and use it to develop a timeline for data collection that is realistic given the local context. In a 3MC survey, there are often differences in postal reliability, cost, possible carriers, and timeliness.

   1.2 When designing materials (letters, questionnaires, etc.) that will be mailed to the respondent, assess the following:
   1.2.1 Literacy levels among the target population
   1.2.2 Use of languages and/or regional dialects other than the country's official language(s) and any implications for the feasibility of a self-completed questionnaire. Indeed, there are some languages and dialects that do not have a written form.

   1.3 Determine how data entry of returned mail questionnaires will occur. Data entry can occur manually but it is more efficient to use optical or intelligent character recognition software, wherein the computer will read and code response from paper questionnaires.

   1.4 Before mailing out the paper questionnaire, consider sending a well-written advance letter to legitimize the survey and reassure and motivate potential respondents. Most effective is a carefully drafted, simple, short letter (Couper & de Leeuw, 2003; Lynn, Turner, & Smith, 1997; Dillman, 2000).

1.5 Develop a cover letter to include with the paper questionnaire, introducing the research study and explaining the purpose of the survey, instructions on how to complete the instrument, and organization contact information for any questions the respondent might have.

1.6 Develop an instrument appropriate for the mode and target population, keeping in mind that there will be no interviewer present to assist with the survey administration.
  1.6.1 Assess the literacy of the target population and adjust the text for comprehension if necessary.
  1.6.2 Place instructions clearly next to the survey questions to which they correspond.
  1.6.3 Make the layout of the instrument visually appealing and question order easy to follow. Use visual elements (e.g., brightness, color, shape, position on page) in a consistent way to define the desired path through the questionnaire (Jenkins & Dillman, 1997; Groves, et al., 2009).
  1.6.4 Use skip patterns only when absolutely necessary. Include clear instructions for skip patterns and reinforce with visual and graphical cues, such as boldfacing and arrows.
  1.6.5 Limit the number of open-ended questions
  1.6.6 Ask only one question at a time. Combining multiple items into one question places a heavy cognitive burden on respondents and can impact data quality.

1.7 Provide clear instructions for returning the completed survey to the research organization or other point of collection. Adequate postage should be provided on the envelope so as not to incur cost to the respondent.

1.8 Develop a sample management system to process completed paper questionnaires and develop procedures for its execution.

1.9 Institute protocols to protect respondent confidentiality. It is common for research organizations to assign a unique identification number to each sampled household's questionnaire for sample management purposes as questionnaire are mailed back to the office. This ensures that if a paper questionnaire is lost in the mail or is not otherwise returned to the survey organization, the respondent's answers cannot be linked to their identity by a third party.

1.10 Develop a protocol for addressing non-response, including how many attempts to reach respondents by mail and/or other possible methods will be made.

Lessons Learned

1.1 Because a mail survey is self-administered without an interviewer present, it is crucial that the layout and design of the questionnaire elements is clear and easy to follow, and that instructions are visibly marked. Often, the first page of a mail survey contains a lengthy set of instructions which respondents generally skip or do not retain when completing the questionnaire, argue Jenkins and Dillman (1997). They advise the placement of relevant instructions be directly where they need to be.

1.2 A recent mail survey in Siberia, which varied experimental factors across random subgroups of respondents, achieved greatest response rates when official university letterhead was used in correspondence, when there was an incentive offered, and when a larger number, versus a smaller number, of contacts with the respondent were attempted (Avdeyeva & Matland, 2012).

1.3 Expected response rates for mail surveys will differ by country. For a limited set of studies examining cross-national differences in response rates, see Couper and de Leeuw (2003), de Heer (1999), Hox et al. (2002), and Smith (2007).

2. When administering a survey via the web (i.e., the Internet), develop the questionnaire and protocols with consideration that the survey must be straightforward for respondents to self-administer.

Rationale

Internet penetration has been steadily increasing worldwide in recent years. Given the increased costs of interviewer-administered surveys, many researchers are turning to the use of web-based surveys to reach respondents when an adequate sample frame is available. Web surveys should be designed so respondents can easily access and complete the survey.

Procedural Steps

2.1 Assess each study country's technological infrastructure to select software appropriate for use, depending on instruments prevalent in the study country, for the development, distribution, and completion of the web survey.

2.1.1 Assess Internet speed and reliability in the study country and potential impact on ease of web survey use by respondents and design the survey to fit the country's bandwidth limitations.

2.1.2 Determine which web browser(s) fully supports the web-based survey instrument and communicate this to the respondent. Consider including a link to download a specific web browser to facilitate the respondent's participation in the web survey.

2.1.3 Consider that respondents will likely use different devices to access the survey, including desktop computers, laptop computers, tablets, smartphones, and other electronic devices. The web survey should be able to be completed on a web browser, regardless of the type electronic device. See Instrument Technical Design for additional information on preparing style sheets appropriate for multiple devices.

2.1.4 Plan for adequate programming and testing time on multiple devices. For example, software that may be compatible with Android devices may have glitches in iOS (Apple) devices.

2.2 Determine how respondents will be invited to participate in the web survey.

2.2.1 Before disseminating the link to the web-based survey instrument, consider sending a well-written advance letter to legitimize the survey and reassure and motivate potential respondents. Most effective is a carefully drafted, simple, short letter (Couper & de Leeuw, 2003; Lynn et al., 1997; Dillman, 2000).

2.2.2 Mode of invitation will be limited by the respondent contact information available from the sample frame. For example, a web survey using a sampling frame consisting solely of email addresses will not be able to send an invitation via postal mail because of the lack of a mailing address.

2.3 Determine how respondents will gain access to the survey. Dillman (2000) proposes providing a PIN number for limiting access only to people in the sample. Another option is to provide each respondent with a unique Internet link to the survey, which is linked to the respondent's sample id.

2.4 Develop a concise introduction to be presented at the start of the web survey, introducing the research study and explaining the purpose of the survey, instructions on how to complete the survey, and organization contact information for any questions the respondent might have.

2.5 Develop and test the web survey, keeping in mind that there will be no interviewer present to assist with the survey administration.
   - 2.5.1 Assess the literacy of the target population and adjust the text for comprehension if necessary.
   - 2.5.2 The first question should be an item that is likely to be interesting to most respondents and easy to answer.
   - 2.5.3 Place instructions alongside the survey questions to which they correspond.
   - 2.5.4 Make the layout of the instrument visually appealing.
   - 2.5.5 Program any skip patterns used directly into the instrument, relieving the respondent from navigational decisions.
   - 2.5.6 Keep the survey as brief and engaging as possible. The longer the questionnaire and the greater the number of screens, the more likely the respondent will not finish the questionnaire (MacElroy, 2000).
   - 2.5.7 Limit the number of open-ended questions.
   - 2.5.8 Ask only one question at a time. Combining multiple items into one question places a heavy cognitive burden on respondents and can impact data quality.
   - 2.5.9 Make prompts, particularly those asking for the respondent to correct an answer, helpful, polite, and encouraging.
   - 2.5.10 Decide whether respondents can navigate backwards to revisit and/or revise previous survey items and responses.
   - 2.5.11 See Instrument Technical Design for additional guidance on the layout and technical design of the web survey.

2.6 Decide whether respondents will be permitted to complete the questionnaire in more than one session, allowing for the data to be saved in the interim, and program the instrument accordingly.

2.7 Institute protocols to protect respondent confidentiality.
   - 2.7.1 Ensure that electronic transmission of the data from the respondent's computer to the survey firm collecting the data is secure.

2.8 Select an appropriate electronic sample management system and develop procedures for its execution. If an electronic sample management system is used, coordinating centers may play a role in monitoring fieldwork. See Study Design and Organizational Structure for details.

2.9 Determine which paradata will be collected. Paradata from web surveys can be used to enhance respondents' experience or to understand more about the respondents and how they interact with the web survey (Couper, 2008). See Paradata and Other Auxiliary Data for more information and examples.

2.10 Develop a protocol for addressing non-response, including how many attempts to reach respondents by email and/or other possible methods will be made.

Lessons Learned

2.1 Web surveys are often used in subsequent waves of panel surveys following an interviewer-administered baseline study, and can be a practical and cost-effective mode choice. In such cases, the respondent is familiar with the study and strategies to minimize non-response can be executed via phone, mail, and even in-person visits because complete contact information is generally available.

2.2 With adequate design, web surveys can achieve response rates comparable to non-web surveys.

2.2.1 A randomized telephone/web mode experiment in a Swiss election study found that the use of an incentive in a web survey produced response rates comparable to those from the telephone survey which also included incentives. The web survey was much less costly, even accounting for the cost of incentives, than for the telephone survey (Lipps & Pekari, 2013).

2.2.2 However, like 3MC surveys conducted in other modes, web surveys can produce difference response rates across countries. A comparison of data collected through a web survey from Italy, France, Turkey, and the U.S. showed that France had the highest overall refusal rate, but low item nonresponse for those who did participate. Italy and the U.S. had response rates and low item non response. Respondent in Turkey had the lowest contact and response rates, and the highest item nonresponse for sensitive questions (Ackermann, Ecklund, Phillips, & Brulia, 2016).

2.3 Internet censorship occurs at the national level in at least several non-Western countries, such as China and Iran. If planning a survey in a country where censorship occurs, consider the survey topic and technical programming and determine whether the web is an acceptable form of data collection for the particular study country.

2.3.1 Censorship by certain governments can impact the types of questions that are permitted on a web survey questionnaire.

2.3.2 Censorship can impact response rates due to confidentiality and security concerns among respondents.

2.3.3 If the study country engages in censorship, consider the location of the server hosting the survey, and whether the study respondents will be able to access the server in its host country; that is, whether the server website IP address is accessible from the study country.

2.4 Software and website vendors can restrict access by users in other countries. Regardless of any government censorship, verify that respondents in the study country can access the survey.

2.5 Smartphone apps are currently being used for time-use surveys. For example, a research study in the Netherlands is using a smartphone app to collect time use data in combination with auxiliary data. By requiring respondents to install an app, rather than access a website to complete the survey, researchers can guarantee that respondents will visually see the instrument exactly as the researchers intended. The app does not need permanent Internet access as completed survey data is stored and transmitted as Internet access permits (Sonck & Fernee, 2013).

3. When administering a survey using IVR, develop the questionnaire and protocols with consideration that the survey must be straightforward for respondents to self-administer.

Rationale

IVR can be an effective mode for administering a survey to a population where telephone accessibility is adequate and particularly when the survey topic is sensitive. However, as with mail and web surveys, the absence of an interviewer necessitates careful instrument design and field execution.

Procedural Steps

3.1 Determine which IVR software will be used to carry out the survey, including whether the IVR system will accept incoming phone calls from respondents to complete the survey and/or will initiate outgoing telephone calls to respondents to complete the survey.

3.2 Determine how respondents will be invited to participate in the IVR survey. Mode of invitation will be limited by the respondent contact information available from the sample frame.
- 3.2.1 If postal addresses are available, respondents can receive an invitation with a telephone number to call to participate.
- 3.2.2 If email addresses are available, respondents can receive an invitation and telephone via email.
- 3.2.3 If only telephone numbers are available, the invitation to complete the IVR will occur by telephone.

3.3 If an automated dialing system will be used to initiate contact with the respondent, assess any legal restrictions in place that apply to the use of such systems in the study country.

3.4 Develop a concise introduction to be presented at the start of the IVR survey, introducing the research study and explaining the purpose of the survey, instructions on how to complete the survey, and organization contact information for any questions the respondent might have.

3.5 Decide whether to program the IVR system as touchtone, voice input, or a combination of the two.
  3.5.1 When deciding on the programming, consider the target population. Studies in rural India and Botswana found that respondents with less education and lower literacy do better with touchtone, and cited privacy for touchtone preference as well (Kuun, 2010; Patel et al., 2009).
  3.5.2 A study in Pakistan found that a well-designed speech interface was more effective than a touch-tone system for respondents regardless of literacy level (Sherwani et al., 2009).

3.6 Devote sufficient time to the development of a high-quality IVR system to maintain respondent interest and continued cooperation.
  3.6.1 The IVR system must have a high quality recording, as the respondent is likely to break off the survey if quality is poor.
  3.6.2 See Oberle (2008) for a guide to the development of an IVR system and the associated speech characteristics which need consideration.

3.7 Select an appropriate sample management system and develop procedures for its execution.
  3.7.1 If an electronic sample management system is used, coordinating centers may play a role in monitoring fieldwork. See Study Design and Organizational Structure for details.

3.8 Develop a protocol for addressing non-response, including how many attempts to reach respondents by telephone and/or other possible methods will be made.

Lessons Learned

3.1 Consider the voice used for recording.

3.1.1 In a health helpline project in Botswana, researchers employed a well-known local actress for the IVR recording, and users reacted very positively (Kuun, 2010).
3.1.2 Depending on the social context, using an IVR recording of a male for male respondents and of a female for female respondents may elicit more accurate reporting, particularly of sensitive information.

3.2 Plauche, Nallasamy, Pal, Wooters, and Ramachandran (2006) developed an innovative approach to the challenge that dialectical variation and multilingualism poses to speech-driven interfaces for IVR in India, applicable to other settings as well. In their approach, people from specific villages are recorded during interactions, and their speech is semi-automatically integrated into the acoustic models for that village, thus generating the linguistic resources needed for automatic recognition of their speech.

3.3 Consider an alternate mode for first contact to inform respondent of impending IVR survey, such as SMS or other mailing. In a study in rural Uganda, the IVR survey call was preceded by an SMS message 24 hours prior, about the upcoming call. In a pretest, respondents who didn’t receive the text were unable to make sense of the later survey call (Lerer, Ward, & Amarasinghe, 2010).

3.4 A survey of teachers in Uganda resulted in a number of useful considerations when designing an IVR system to improve response rates and data quality (Lerer et al., 2010).
3.4.1 The IVR call began with the immediate information that “This is a recorded call from Project X. You are not talking to a real person.”
3.4.2 The IVR call provided very specific instructions about whether to use keypad or to speak
3.4.3 Respondents were initially confused by the automation of the IVR system. Researchers had better results when using a chime to get respondents’ attention before the automated voice gave instructions.
3.4.4 Leveraging conversational and turn-taking conventions of normal conversation in the IVR system lead to more success than detailed instructions in eliciting desired user behavior.
3.4.5 An IVR system which projected a loud voice, with prompts recorded like the speaker was using a poor cell connection, resulted in a survey that was easier for respondents to follow.
3.4.6 When producing the IVR recording, use slow speech to get slow speech – respondents will emulate the voice, and resulting data will be easier to understand.

3.4.7 The IVR recording included 3 seconds of silence before the recorded speakers says “thank you” and moves onto next question, which was reported as well-received by respondents.

Data Collection References

Ackermann, A., Ecklundh, E., Phillips, B., & Brulia, A. (2016). *Patterns of nonresponse in a multinational web survey*. Paper presented at the American Association for Public Opinion Research, Austin, Texas.

Afrobarometer Survey. Retrieved June 1, 2016, from http://www.afrobarometer.org/sites/default/files/survey_manuals/ab_r6_survey_manual_en.pdf

Ali, M., Deen, J. L., Khatib, A., Enwere, G., von Seidlein, L., Reyburn, R., ... & Saleh, A. A. (2010). Paperless registration during survey enumerations and large oral cholera mass vaccination in Zanzibar, the United Republic of Tanzania. *Bulletin of the World Health Organization*, *88*(7), 556-559.

American Association for Public Opinion Research (AAPOR). (2003). *Interviewer falsification in survey research: Current best methods for prevention, detection and repair of its effects*. Paper presented at Ann Arbor Falsification Summit on Interviewer Falsification. Retrieved March 26, 2010, from http://www.amstat.org/sections/srms/falsification.pdf

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

Asian Barometer. Retrieved July 1, 2010, from http://www.asianbarometer.org/intro/program-overview

Aurelien, G. A. S., West, B. T., & Sakshaug, J. W. (2016). *How big of a problem is analytic error in secondary analyses of survey data?* The 71th Annual Conference of the American Association for Public Opinion Research. Austin, Texas.

Avdeyeva, O. A., & Matland, R. E. (2012). An experimental test of mail surveys as a tool for social inquiry in Russia. *International Journal of Public Opinion Research*, *25*(2), 173-194.

Avilés, W., Ortega, O., Kuan, G., Coloma, J., & Harris, E. (2007). Integration of information technologies in clinical studies in Nicaragua. PLoS Med, 4(10), e291.

Axinn, W. G., & Pearce, L. D. (2006). *Mixed method data collection strategies*. New York, NY: Cambridge University Press.

Axinn, W. G., Chardoul, S., Ghimire, D., Gatny, H., & Barber, J. (2008). *Techniques for creating a continuous record of relationships, contraception and pregnancy: Examples from radically different social and survey research settings*. Paper presented at the international Conference on Survey Methods in Multinational, Multiregional, and Multicultural Contexts (3MC). Berlin, Germany.

Axinn, W. G., Fricke, T. E., & Thornton, A. (1991). The micro-demographic community study approach: Improving survey data by integrating the ethnographic method. *Sociological Methods Research*, *20*(2), 187-217.

Axinn, W.G., Pearce, L.D., & Ghimire, D.J. (1999). Innovations in life history calendar applications. *Social Science Research, 28*, 243-264.

Baker, R. P., Bradburn, N. M., & Johnson, R. A. (1995). Computer-assisted personal interviewing: An experimental evaluation of data quality and cost. *Journal of Official Statistics*, *11*(4), 413-431.

Bamberger, M., Rao, V., & Woolcock, M. (2010). Using mixed methods in monitoring and evaluation: Experiences from international development. *World Bank Policy Research Working Paper Series.*

Beauclair, R., Meng, F., Deprez, N., Temmerman, M., Welte, A., Hens, N., & Delva, W. (2013). Evaluating audio computer assisted self-interviews in urban South African communities: Evidence for good suitability and reduced social desirability bias of a cross-sectional survey on sexual behaviour. *BMC Medical Research Methodology*, *13*(1), 1.

Belli, R. F. (1998). The structure of autobiographical memory and the event history calendar: Potential improvements in the quality of retrospective reports in surveys. *Memory, 6*(4), 383-406.

Berk, M. L., Mathiowetz, N. A., Ward, E. P., & White, A. A. (1987). The effect of prepaid and promised incentives: Results of a controlled experiment. *Journal of Official Statistics, 3*, 449-457.

Bernabe-Ortiz, A., Curioso, W. H., Gonzales, M. A., Evangelista, W., Castagnetto, J. M., Carcamo, C. P., ... & Holmes, K. K. (2008). Handheld computers for self-administered sensitive data collection: A comparative study in Peru. *BMC Medical Informatics and Decision Making*, *8*(1), 1.

Bhatnagar, T., Brown, J., Saravanamurthy, P. S., Kumar, R. M., & Detels, R. (2013). Color-coded audio computer-assisted self-interviews (C-ACASI) for poorly educated men and women in a semi-rural area of south India: “Good, scary and thrilling”. *AIDS and Behavior*, *17*(6), 2260-2268.

Biemer, P. P. (2004). The twelfth Morris Hansen lecture simple response variance: Then and now. *Journal of Official Statistics, 20*, 417-439.

Biemer, P. P., & Forsman, G. (1992). On the quality of reinterview data with application to the current population survey. *Journal of the American Statistical Association, 87*, 915-923.

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

Blom, A. G., de Leeuw, E. D., & Hox, J. (2011). Interviewer effects on nonresponse in the European Social Surveys. *Journal of Official Statistics*, *27*(2), 3–53.

Bohannon, J. (2016). Survey fraud test sparks battle. *Science* 351, no. 6277, 1014-1014.

Bradley, J., Ramesh, B. M., Rajaram, S., Lobo, A., Gurav, K., Isac, S., ... & Alary, M. (2012). The feasibility of using mobile phone technology for sexual behaviour research in a population vulnerable to HIV: A prospective survey with female sex workers in South India. *AIDS Care*, *24*(6), 695-703.

Broom, A. (2005). Using qualitative interviews in CAM research: A guide to study design, data collection, and data analysis. C*omplementary Therapies in Medicine, 13*, 65-73.

Byass, P., Hounton, S., Ouédraogo, M., Somé, H., Diallo, I., Fottrell, E., ... & Meda, N. (2008). Direct data capture using hand-held computers in rural Burkina Faso: Experiences, benefits and lessons learnt. *Tropical Medicine & International Health,* 13(s1), 25-30.

Carlson, R. O. (1958). To talk with kings. *Public Opinion Quarterly*, *22*(3), 224-229.

Chikwanha, A. B. (2005). *Conducting surveys and quality control in Africa: Insights from the Afrobarometer*. Paper presented at the WAPOR/ISSC Conference, Ljubljana, Slovenia.

Choldin, H. M., Kahn, A. M., & Ara, B. H. (1983). Cultural complications in fertility interviewing. In M. Bulmer & D. P. Warwick (Eds.), *Social research in developing countries* (pp. 261-274)*.* New York, NY: John Wiley & Sons.

Cialdini, R. B. (1988). *Influence: Science and practice*. Glenview, IL: Scott, Foresman.

Cohen, G., & Duffy, J. C. (2002). Are nonrespondents to health surveys less healthy than respondents? *Journal of Official Statistics*, *18*, 13-23.

Couper, M., & Groves, R. (2002). Chapter 7: Introductory interactions in telephone surveys and nonresponse, In D. Maynard, H. Houtkoop-Steenstra, N. Schaeffer, & J. van der Zouwen (Eds.), *Standardization and tacit knowledge: Interaction and practice in the survey interview.* New York, NY: John Wiley & Sons.

Couper, M. P. (2000). Usability evaluation of computer-assisted survey instruments. *Social Science Computer Review*, *18*(4), 384-396.

Couper, M. P. (2005). Technology trends in survey data collection. *Social Science Computer Review*, 23(4), 486-501.

Couper, M. P. (2008). *Designing effective Web surveys*. New York, NY: Cambridge University Press.

Couper, M. P., & de Leeuw, E. D. (2003). Nonresponse in cross-cultural and cross-national surveys. In J. A. Harkness, F. J. R. van de Vijver, & P. P. Mohler (Eds.), *Cross-cultural survey methods* (pp. 157-177). Hoboken, NJ: John Wiley & Sons.

Couper, M. P., Holland, L., & Groves, R. M. (1992). Developing systematic procedures for monitoring in a centralized telephone facility. *Journal of Official Statistics*, *8*, 63-76.

de Heer, W. (1999). International response trends: Results of an international survey. *Journal of Official Statistics*, *15*(2), 129-142.

de Leeuw, E. D., & de Heer, W. (2002). Trends in household survey nonresponse: A longitudinal and international comparison. In R. M. Groves, A. D. Dillman, J. Eltinge, & R. Little (Eds.), *Survey nonresponse* (pp. 41-54). New York, NY: John Wiley & Sons.

Demographic and Health Surveys. (2010). Retrieved August 17, 2010, from http://www.measuredhs.com/aboutsurveys/

Demombynes, G., Gubbins, P., & Romeo, A. (2013). Challenges and opportunities of mobile phone-based data collection: Evidence from south Sudan. *World Bank Policy Research Working Paper*, (6321).

Diero, L., Rotich, J. K., Bii, J., Mamlin, B. W., Einterz, R. M., Kalamai, I. Z., & Tierney, W. M. (2006). A computer-based medical record system and personal digital assistants to assess and follow patients with respiratory tract infections visiting a rural Kenyan health centre. *BMC Medical Informatics and Decision Making*, *6*(1), 1.

Dillman, D. (2000). *Mail and internet surveys.* Hoboken, NJ: John Wiley & Sons.

Dillman, D. A., Lesser, V., Mason, R., Carlson, J., Willits, F., Robertson, R., & Burke, B. (2007). Personalization of mail surveys for general public and populations with a group identity: Results from nine studies. *Rural Sociology*, *72*(4), 632-646.

Dillman, D. A., Smyth, J. D., & Christian, L. M. (2014). *Internet, phone, mail, and mixed-mode surveys: The tailored design method*. Hoboken, NJ: John Wiley & Sons.

Dillon, B. (2012). Using mobile phones to collect panel data in developing countries. *Journal of international development*, *24*(4), 518-527.

Escandon, I. N., Searing, H., Goldberg, R., Duran, R., & Monterrey Arce, J. (2008). The use of PDAs to collect baseline survey data: Lessons learned from a pilot project in Bolivia. *Global Public Health*, *3*(1), 93-104.

Forsman, G., & Schreiner, I. (1991). The design and analysis of reinterview: An overview. In P. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, & S. Sudman (Eds.), *Measurement errors in survey*s (pp. 279-301). New York, NY: John Wiley & Sons.

Forster, D., & Snow, B. (1992). Using microcomputers for rapid data collection in developing countries. *Health Policy and Planning*, *7*(1), 67-71.

Frederick, H. (2012). *Cultural aspects of total survey quality: Entrepreneurship research in the kingdom of Tonga*. In 4th Global Entrepreneurship Research Conference Papers. Imperial College Business School, London.

Freedman, D.S., Thornton, A., Camburn, D., Alwin, D., & Young-Demarco, L. (1988). The life history calendar: A technique for collecting retrospective data. *Sociological Methodology, 18*, 37-68.

Fricke, T.E. (2005). Taking culture seriously: Making the social survey ethnographic. In T. S. Weisner (Ed.), *Discovering successful pathways in children's development: Mixed methods in the study of childhood and family life (pp. 185-218).* Chicago, IL: University of Chicago Press.

Gallup, Inc. (2007). *Gallup World Poll research design*. Retrieved May 7, 2016, from http://media.gallup.com/WorldPoll/PDF/WPResearchDesign091007bleeds.pdf

Gallup, Inc. (2015). How does the Gallup World Poll work? Retrieved Jan 30, 2015 from http://www.gallup.com/178667/gallup-world-poll-work.aspx

Gasquet, I., Falissard, B., & Ravaud, P. (2001). Impact of reminders and method of questionnaire distribution on patient response to mail-back satisfaction survey. *Journal of Clinical Epidemiology*, *54*, 1174-1180.

Gengler, J. (2013). *Interviewer effects in the Arab Gulf region: Evidence from Bahrain and Qatar*. Presented at the 2013 Annual Conference of the American Association for Public Opinion Research, Boston, MA.

Ghimire, D, Williams, N., Thornton, T., Young-DeMarco, L, & Bhandari, P. (2013). *Innovation in the study of international migration*. Paper presented at the Population Association of America, New Orleans.

Goldstein, N. J., Griskevicius, V., & Cialdini, R. B. (2012). Reciprocity by proxy: A novel influence strategy for stimulating cooperation. *Administrative Science Quarterly*, 0001839211435904.

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

Groves, R. M., & Couper, M. P. (1998). *Nonresponse in household interview surveys*. New York, NY: John Wiley & Sons.

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*, 439-457.

Groves, R. M., & Mathiowetz, N. A. (1984). Computer assisted telephone interviewing: Effects on interviewers and respondents. *Public Opinion Quarterly*, *48*(1B), 356-369.

Groves, R. M., Cialdini, R. B., & Couper, M. P. (1992). Understanding the decision to participate in a survey. *Public Opinion Quarterly*, *56*(4), 475-495.

Groves, R. M., Dillman, D. A., Eltinge, J. L., & Little, R. J. A. (Eds.), (2002). *Survey nonresponse*. New York, NY: John Wiley & Sons.

Groves, R. M., Fowler, F. J. Jr., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2009). *Survey methodology* (2nd ed.). Hoboken, NJ: John Wiley & Sons.

Gutiérrez, J. P., & Torres-Pereda, P. (2009). Acceptability and reliability of an adolescent risk behavior questionnaire administered with audio and computer support. Revista Panamericana de Salud Pública, 25(5), 418-422.

Hargreaves, J. R., Morison, L. A., Gear, J. S. S., Makubele, M. B., Porter, J. D. H., Busza, J., Watts, C., Kim, J. C., & Pronyk, P. M. (2006). 'Hearing the voices of the poor': Assigning poverty lines on the basis of local perceptions of poverty - A quantitative analysis of qualitative data from participatory wealth rankings in rural South Africa. *World Development, 35*(2), 212-229.

Harkness, J. A. (1999). In pursuit of quality: Issues for cross-national survey research. *International Journal of Social Research Methodology*, *2*, 125-140.

Hewett, P. C., Erulkar, A. S., & Mensch, B. S. (2004a). The feasibility of computer-assisted survey interviewing in Africa: Experience from two rural districts in Kenya. *Social Science Computer Review*, *22*(3), 319-334.

Hewett, P. C., Mensch, B. S., & Erulkar, A. S. (2004b). Consistency in the reporting of sexual behaviour by adolescent girls in Kenya: A comparison of interviewing methods. *Sexually Transmitted Infections*, 80(Suppl 2), 43–48.

House, C. C., & Nicholls, W. L. (1988). Questionnaire design for CATI: Design objectives and methods. In R. M. Groves (Ed.), *Telephone survey methodology* (pp. 421-436). New York, NY: John Wiley & Sons.

Howe, G., & McKay, A. (2007). Combining quantitative and qualitative methods in assessing chronic poverty: The case of Rwanda. *World Development, 35*(2), 197-211.

Howell, D. (2010). *Enhancing quality and comparability in the comparative study of electoral systems.* In J. A. Harkness, M. Braun, B. Edwards, T. P. Johnson, L. E. Lyberg, P. P. Mohler, B-E. Pennell, T. Smith (Eds.) *Survey methods in multicultural, multinational, and multiregional contexts* (Chapter 30). Hoboken, NJ: John Wiley & Sons.

Hox, J., de Leeuw, E., Couper, M. P., Groves, R. M., de Heer, W., Kuusela, V., ... & Belak, E. (2002). The influence of interviewers' attitude and behavior on household survey nonresponse: An international comparison. In R. Groves, D. A. Dillman, J. L. Eltinge, & R. J. A. Little (Eds.), *Survey nonresponse*, 103-120. New York, John Wiley & Sons.

Hughes, K. (2004). *Immigrants' perception of the 'survey': A cross-cultural comparison*. Research working paper, Statistical Research Division, US Census Bureau.

Hughes, S., & Haddaway, S. (2014). *Tablets and smartphones: Adopting new technologies for household surveys in Kenya*. Paper presented at the Twelfth International Workshop on Comparative Survey Design and Implementation at NORC, Bethesda, Maryland, from March 27-29.

Hulme, D. (2007). *Integrating quantitative and qualitative research for country case studies of development.* The work was part of the program of the Economic and Social Research Council, Global Poverty Research Group. (No. GPRG-WPS-063).

Ikeda, K., Yamada, M., Taniguchi, N., Nishizawa, Y., & Steel, G. (2007). *Japan country report: Second wave of Asian barometer survey*. Working Paper Series, 32. Retrieved March 26, 2010, from http://www.asianbarometer.org/publications/3c27729fb68cd1f80d6c7573f5a93280.pdf

Inter-university Consortium for Political and Social Research (ICPSR). (2012). *Data sharing for demographic research*. Retrieved March 9, 2015, from http://www.icpsr.umich.edu/icpsrweb/DSDR/

Inter-university Consortium for Political and Social Research (ICPSR). (2015). *Data management & curation*. Retrieved October 20, 2015, from http://www.icpsr.umich.edu/icpsrweb/content/datamanagement/confidentiality/

Jaspan, H. B., Flisher, A. J., Myer, L., Mathews, C., Seebregts, C., Berwick, J. R., ... & Bekker, L. G. (2007). Brief report: Methods for collecting sexual behaviour information from South African adolescents—A comparison of paper versus personal digital assistant questionnaires. *Journal of Adolescence*, *30*(2), 353-359.

Jaya, Hindin, M. J., & Ahmed, S. (2008). Differences in young people's reports of sexual behaviors according to interview methodology: A randomized trial in India. *American Journal of Public Health*, 98(1), 169–174.

Jenkins, C. R., & and Dillman, D. A. (1997). Chapter 7: Towards a theory of self administered questionnaire design. In L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz, & D. Trewin (Eds.), *Survey measurement and process quality* (pp. 165-196). New York: Wiley-Interscience.

Kankaraš, M., & Moors, G. (2009). Measurement equivalence in solidarity attitudes in Europe: Insights from a multiple-group latent-class factor approach. *International Sociology, 24*(4), 557–579.

Kankaraš, M., & Moors, G. (2010). Researching measurement equivalence in cross-cultural studies. *Psihologija 2010, 43*(2), 121-136.

Kankaraš, M., Moors, G., & Vermunt, J. K. (2010). Testing for measurement invariance with latent class analysis. In E. Davidov, P. Schmidt, & J. Billiet (Eds.), *Cross-cultural analysis: Methods and applications,* (pp. 359–384). New York, NY: Routledge Taylor & Francis Group.

Kankaraš, M., Vermunt, J. K., & Moors, G. (2011). Measurement equivalence of ordinal items: A comparison of factor analytic, item response theory, and latent class approaches. *Sociological Methods & Research 40*, 279-310.

Keller, H. (2007). *Cultures of infancy*. Mahwah, NJ: Lawrence Erlbaum.

Kessler, R. C., Berglund, P., Chiu, W. T., Demler, O., Heeringa, S. G., Hiripi, E., Jin, R., … & Zheng, H. (2004). The US National Comorbidity Survey replication (NCS-R): Design and field procedures. *International Journal of Methods in Psychiatric Research, 13*, 69-92.

Kessler, R. C., Ustun, T. B., & World Health Organization. (2008). *The WHO World Mental Health Surveys: Global perspectives on the epidemiology of mental disorders*. Cambridge; New York: Geneva: Cambridge University Press; published in collaboration with the World Health Organization.

Kish, L. (1949). A procedure for objective respondent selection within the household. *Journal of the American Statistical Association*, *44*, 380-387.

Kissinger, P., Rice, J., Farley, T., Trim, S., Jewitt, K., Margavio, V., & Martin, D. H. (1999). Application of computer-assisted interviews to sexual behavior research. *American Journal of Epidemiology*, *149*(10), 950-954.

Kleiner, B., & Pan, Y. (2006). Cross-cultural communication and the telephone survey interview. *Conducting Cross-National and Cross-Cultural Surveys*, 81.

Koch, A., Blom, A. G., Stoop, I., & Kappelhof, J. (2009). Data collection quality assurance in cross-national surveys: The example of the ESS. *Methoden Daten Analysen*, *3*, 219-247.

Krawczyk, C. S., Gardner, L. I., Wang, J., Sadek, R., Loughlin, A. M., Anderson-Mahoney, P., ... & Antiretroviral Treatment and Access Study Group. (2003). Test-retest reliability of a complex human munodeficiency virus research questionnaire administered by an audio computer-assisted self-interviewing system. *Medical Care*, *41*(7), 853-858.

Krishna, A. (2007). Subjective assessments, participatory methods and poverty dynamics: The Stages-of-Progress method (October 1, 2007). Chronic Poverty Research Centre Working Paper No. 93. Available at SSRN: http://ssrn.com/abstract=1629295

Krysan, M., & Couper, M. P. (2003). Race in the live and the virtual interview: Racial deference, social desirability, and activation effects in attitude surveys. *Social Psychology Quarterly*, 364-383.

Kulka, R. A., & Weeks, M. F. (1988). Towards the development of optimal calling protocols for telephone surveys: A conditional probabilities approach. *Journal of Official Statistics, 4*, 319-332.

Kuriakose, N., & Robbins, M. (2015). Don't get duped: Fraud through duplication in public opinion surveys. *Statistical Journal of the IAOS*.

Kuun, C. (2010). *OpenPhone project piloted in Botswana*. Retrieved on June 18, 2015 from http://www.csir.co.za/enews/2008_july/ic_05.html

Lane, S. J., Heddle, N. M., Arnold, E., & Walker, I. (2006). A review of randomized controlled trials comparing the effectiveness of hand held computers with paper methods for data collection. *BMC Medical Informatics and Decision Making*, *6*(23).

Langhaug, L. F., Sherr, L., & Cowan, F. M. (2010). How to improve the validity of sexual behaviour reporting: Systematic review of questionnaire delivery modes in developing countries. *Tropical Medicine & International Health*, *15*(3), 362-381.

Langhaug, L. F., Cheung, Y. B., Pascoe, S. J., Chirawu, P., Woelk, G., Hayes, R. J., & Cowan, F. M. (2011). How you ask really matters: Randomised comparison of four sexual behaviour questionnaire delivery modes in Zimbabwean youth. *Sexually Transmitted Infections,* 87(2), 165-173.

Lara, D., Strickler, J., Olavarrieta, C. D., & Ellertson, C. (2004). Measuring induced abortion in Mexico: A comparison of four methodologies. *Sociological Methods & Research*, *32*(4), 529-558.

Lau, C., Lombaard, A., Baker, M., Eyerman, J., & Thalij, L. (2016). *The validity of collecting data using short messge service (SMS): Experimental evidence from four African countries*. Paper presented at the American Association for Public Opinion Research, Austin, Texas.

Lavrakas, P. J. (1993). *Telephone survey methods: Sampling, selection, and supervision* (2nd ed.). Thousand Oaks, CA: Sage Publications.

Le, L. C., & Vu, L. T. (2012). Audio computer-assisted self interview compared to traditional interview in an HIV-related behavioral survey in Vietnam. *MEDICC Review*, *14*(4), 26-31.

Lerer, A., Ward, M., & Amarasinghe, S. (2010). Evaluation of IVR data collection UIs for untrained rural users. In *Proceedings of the First ACM Symposium on Computing for Development* (p. 2). ACM.

Li, L. (2011) Retrieved on May 19, 2016, from http://www.lukeexperiments.com/content/index/media/Luke_Li_Eye_Tracking_Methodology.pdf

Lin, I. F., & Schaeffer, N. C. (1995). Using survey participants to estimate the impact of nonparticipation. *Public Opinion Quarterly*, *59*(2), 236-258.

Lipps, O., & Pekari, N. (2013). *Mode and incentive effects in an individual register frame based Swiss election study*. FORS working paper 3-2013, Lausanne.

Living Standard Measurement Study Survey. (1996). *Working paper 126: A manual for planning and implementing the Living Standard Measurement Study Survey*. Retrieved July 10, 2010, from http://www-wds.worldbank.org/external/default/WDSContentServer/WDSP/IB/2000/02/24/000009265_3961219093409/Rendered/PDF/multi_page.pdf

Lyberg, L., Japec, L., & Tongur, C. (2017). Prevailing issues and the future of comparative surveys. In T. P. Johnson, B-E. Pennell, I. A. L. Stoop, & B. Dorer (Eds.), *Advances in comparative survey methods: Multicultural, multinational and multiregional contexts (3MC*). New York, NY: John Wiley & Sons.

Lynn, P., Turner, R., & Smith, P. (1997). The effect of complexity and tone of an advance letter on response to an interview survey. *Survey Methods Centre Newsletter*, *17*, 13-17.

MacElroy, B. (2000). Variables influencing dropout rates in Web-based surveys. *Quirk's Marketing Research Review*, July/August. http://www.quirks.com/articles/a2000/20000711.aspx

Massey, D. S. (1987). The ethnosurvey in theory and practice. *International migration review*, *21*(4), 1498-1522.

Mensch, B. S., Hewett, P. C., & Erulkar, A. S. (2003). The reporting of sensitive behavior by adolescents: A methodological experiment in Kenya. *Demography*, *40*(2), 247-268.

Mensch, B. S., Hewett, P. C., Gregory, R., & Helleringer, S. (2008). Sexual behavior and STI/HIV status among adolescents in rural Malawi: An evaluation of the effect of interview mode on reporting. *Studies in Family Planning*, *39*(4), 321-334.

Mensch, B. S., Hewett, P. C., Abbott, S., Rankin, J., Littlefield, S., Ahmed, K., ... & Mierzwa, S. (2011). Assessing the reporting of adherence and sexual activity in a simulated microbicide trial in South Africa: An interview mode experiment using a placebo gel. *AIDS and Behavior*, *15*(2), 407-421.

Minnis, A. M., Steiner, M. J., Gallo, M. F., Warner, L., Hobbs, M. M., Van der Straten, A., & Padian, N. S. (2009). Biomarker validation of reports of recent sexual activity: Results of a randomized controlled study in Zimbabwe. *American Journal of Epidemiology*, *170*(7), 918-924.

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. M., 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.

Morse, J. M., & Niehaus, L. (2009). *Mixed method design: Principles and procedures.* Walnut Creek, CA: Left Coast Press.

Mudryk, W., Burgess, M. J., & Xiao, P. (1996). Quality control of CATI operations in Statistics Canada. *Proceedings of the Section on Survey Research Methods* (pp.150-159). Alexandria, VA: American Statistical Association.

NIMH Collaborative HIV/STD Prevention Trial Group. (2007). The feasibility of audio computer-assisted self-interviewing in international settings. *Aids*, *21*, S49-S58.

Oberle, F. (2008). Who, why and how often? Key elements for the design of a successful speech application taking account of the target groups. T. Hempel (Ed.), *Usability of speech dialog systems* (pp. 1-39). Berlin: Springer.

Office of Management and Budget. (2006). *Standards and guidelines for statistical surveys*. Washington, DC: Office of Information and Regulatory Affairs, OMB. Retrieved January 10, 2010, from http://www.whitehouse.gov/omb/inforeg/statpolicy/standards_stat_surveys.pdf

Onono, M., Carraher, N., Cohen, R., Bukusi, E., & Turan, J. (2011). Use of personal digital assistants for data collection in a multi-site AIDS stigma study in rural south Nyanza, Kenya. *African Health Sciences, 11*(3), 464–473.

Pan, Y., Scollon, S. W., & Scollon, R. (2002). *Professional communication in international settings*. Malden, MA: Blackwell Publishers.

Patel, N., Agarwal, S., Rajput, N., Nanavati, A., Dave, P., & Parikh, T. S. (2009). A comparative study of speech and dialed input voice interfaces in rural India. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 51-54). ACM.

Patnaik, S., Brunskill, E., & Thies, W. (2009, April). Evaluating the accuracy of data collection on mobile phones: A study of forms, SMS, and voice. In *Information and Communication Technologies and Development (ICTD), 2009 International Conference* (pp. 74-84). IEEE.

Paudel, D., Ahmed, M., Pradhan, A., & Dangol, R. L. (2013). Successful use of tablet personal computers and wireless technologies for the 2011 Nepal Demographic and Health Survey. *Global Health: Science and Practice*, *1*(2), 277-284.

Pavlidou, T. (1994) Contrasting German-Greek politeness and the consequences. *Journal of Pragmatics,21*, 487-511

Pennell, B-E., Mneimneh, Z. N., Bowers, A., Chardoul, S., Wells, J. E., Viana, M. C., … Saiz, G. V. (2008). Implementation of the World Mental Health Surveys. In R. C. Kessler & T. B. Üstün (Eds.), *The WHO World Mental Health Surveys: Global perspectives on the epidemiology of mental disorders* (pp. 33-57). Cambridge: Cambridge University Press.

Phillips, A. E., Gomez, G. B., Boily, M. C., & Garnett, G. P. (2010). A systematic review and meta-analysis of quantitative interviewing tools to investigate self-reported HIV and STI associated behaviours in low-and middle-income countries. *International Journal of Epidemiology*, *39*(6), 1541-1555.

Pickering, K., Thomas, R., & Lynn, P. (2003). *Testing the shadow sample approach for the English House Condition Survey.* London, U.K.: National Center.

Plauche, M., Nallasamy, U., Pal, J., Wooters, C., & Ramachandran, D. (2006). Speech recognition for illiterate access to information and technology. In *Information and Communication Technologies and Development, 2006. ICTD'06. International Conference* (pp. 83-92). IEEE.

Plummer, M. L., Ross, D. A., Wight, D., Changalucha, J., Mshana, G., Wamoyi, J., ... & Hayes, R. J. (2004b). “A bit more truthful”: The validity of adolescent sexual behaviour data collected in rural northern Tanzania using five methods. *Sexually Transmitted Infections*, *80*(suppl 2), ii49-ii56.

Plummer, M. L., Wight, D., Ross, D. A., Balira, R., Anemona, A., Todd, J., Salamba, …Hayes, R.J. (2004a). Asking semi-literate adolescents about sexual behaviour: The validity of assisted self-completion questionnaire (ASCQ) data in rural Tanzania. *Tropical Medicine & International Health 2004*, 9(6), 737–754.

Potdar, R., & Koenig, M. A. (2005). Does audio-CASI improve reports of risky behavior? Evidence from a randomized field trial among young urban men in India. *Studies in Family Planning*, *36*(2), 107-116.

Rathod, S. D., Minnis, A. M., Subbiah, K., & Krishnan, S. (2011). ACASI and face-to-face interviews yield inconsistent estimates of domestic violence among women in India: The Samata Health Study 2005-2009. *Journal of Interpersonal Violence*, *26*(12), 2437-2456.

Sampson, R. J., & Laub, J. H. (1998). Integrating quantitative and qualitative data. In J. Z. Giele & G. H. Elder Jr. (Eds.), *Methods of life course research: Qualitative and quantitative approaches* (pp. 213-230). Thousand Oaks, CA: Sage Publications.

Scholz, E. & Heller, M. (2009). *ISSP Study monitoring 2007* (GESIS Technical Reports, 2009/5). Mannheim: GESIS. Retrieved March 26, 2010, from http://www.gesis.org/fileadmin/upload/forschung/publikationen/gesis_reihen/gesis_methodenberichte/2009/TechnicalReport_09-5.pdf

Seebregts, C. J., Zwarenstein, M., Mathews, C., Fairall, L., Flisher, A. J., Seebregts, C., ... & Klepp, K. I. (2009). Handheld computers for survey and trial data collection in resource-poor settings: Development and evaluation of PDACT, a Palm™ Pilot interviewing system. *International Journal of Medical Informatics*, *78*(11), 721-731.

Sherwani, J., Palijo, S., Mirza, S., Ahmed, T., Ali, N., & Rosenfeld, R. (2009). Speech vs. touch-tone: Telephony interfaces for information access by low literate users. In *Information and Communication Technologies and*

Development (ICTD), 2009 International Conference on (pp. 447-457). IEEE.

Shirima, K., Mukasa, O., Schellenberg, J. A., Manzi, F., John, D., Mushi, A., ... & Schellenberg, D. (2007). The use of personal digital assistants for data entry at the point of collection in a large household survey in southern Tanzania. *Emerging Themes in Epidemiology*, *4*(1), 1.

Simmons, K., Mercer, A., Schwarzer, S., & Kennedy, C. (2016). *Evaluating a new proposal for detecting data falsification in surveys*. Pew Research Center, retrieved March 11, 2016 from http://www.pewresearch.org/2016/02/23/evaluating-a-new-proposal-for-detecting-data-falsification-in-surveys/

Singer, E. (2002). The use of incentives to reduce nonresponse in household surveys. In R. M. Groves, D. A. Dillman, J. L. Eltinge & R. J. A. Little (Eds.), *Survey nonresponse* (pp. 163-178). New York, NY: John Wiley & Sons.

Singer, E., Hoewyk, J. V., Gebler, N., Raghunathan, T., & McGonagle, K. (1999). The effect of incentives on response rates in interviewer-mediated surveys. *Journal of Official Statistics, 15*, 217-230.

Skjåk, K. K., & Harkness, J. A. (2003). Data collection methods. In J. A. Harkness, F. J. R. van de Vijver, & P. P. Mohler (Eds.), *Cross-cultural survey methods* (pp. 179-193). Hoboken, NJ: John Wiley & Sons.

Smith, A., Lyons, A., Pitts, M., Croy, S., Ryall, R., Garland, S., ... & Tay, E. (2009). Assessing knowledge of human papillomavirus and collecting data on sexual behavior: Computer assisted telephone versus face to face interviews. *BMC Public Health*, *9*(1), 1.

Smith, T. W. (2007). Survey non-response procedures in cross-national perspective: The 2005 ISSP non-response survey. *Survey Research Methods, 1*(1), 45–54.

Smith, T. W. (2017). Improving cross-national/cultural comparability using the total survey error paradigm. In T. P. Johnson, B-E. Pennell, I. A. L. Stoop, & B. Dorer (Eds.), *Advances in comparative survey methods: Multicultural, multinational and multiregional contexts (3MC*). New York, NY: John Wiley & Sons.

Sonck, N., & Fernee, H. (2013). *Using smartphones in survey research: A multifunctional tool*. The Hague: The Netherlands Institute for Social Research.

Stoop, I. A. L. (2005). The hunt for the last respondent. *The Hague: Social and Cultural Planning Office of the Netherlands*.

Thriemer, K., Ley, B., Ame, S. M., Puri, M. K., Hashim, R., Chang, N. Y., ... & von Seidlein, L. (2012). Replacing paper data collection forms with electronic data entry in the field: Findings from a study of community-acquired bloodstream infections in Pemba, Zanzibar. *BMC Research Notes*, *5*(1), 113.

Turner, C. F., Ku, L., Rogers, S. M., Lindberg, L. D., Pleck, J. H., & Sonenstein, F. L. (1998). Adolescent sexual behavior, drug use, and violence: Increased reporting with computer survey technology. *Science*, *280*(5365), 867-873.

Turner, C. F., Rogers, S. M., Miller, H. G., Miller, W. C., Gribble, J. N., Chromy, J. R., ... & Zenilman, J. M. (2002). Untreated gonococcal and chlamydial infection in a probability sample of adults. *JAMA*, *287*(6), 726-733.

Twisk, J. W. (2006). *Applied multilevel analysis: A practical guide for medical researchers*. Cambridge, UK: Cambridge University Press.

United Nations. (2005). *Household surveys in developing and transition countries.* New York, NY: United Nations, Department of Economic and Social Affairs. Retrieved March 26, 2010, from http://unstats.un.org/unsd/HHsurveys/pdf/Household_surveys.pdf

van de Vijver, F. J. R., & Chasiotis, A. (2010). Making methods meet: Mixed design in cross-national research. In J. Harkness, B. Edwards, M. Braun, T. Johnson, L. E. Lyberg, P. Mohler, B-E. Pennell, & T. Smith (Eds.), *Survey methods in multicultural, multinational, and multiregional contexts*. Hoboken, NJ: John Wiley & Sons.

van de Vijver, F. J. R. & Leung, K. (1997). *Methods and data analysis for cross-cultural research*. Thousand Oaks, CA: Sage Publications.

van de Wijgert, J., Padian, N., Shiboski, S., & Turner, C. (2000). Is audio computer-assisted self-interviewing a feasible method of surveying in Zimbabwe? *International Journal of Epidemiology*, *29*(5), 885-890.

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*, 321-334.

van der Linden, W. J., & Hambleton, R. K. (Eds.). (2013). *Handbook of modern item response theory*. New York, NY: Springer Science & Business Media.

van Griensven, F., Naorat, S., Kilmarx, P. H., Jeeyapant, S., Manopaiboon, C., Chaikummao, S., & Tappero, J. W. (2006). Palmtop-assisted self-interviewing for the collection of sensitive behavioral data: Randomized trial with drug use urine testing. *American Journal of Epidemiology*, *163*(3), 271-278.

Verger, P., Baruffol, E., & Rotily, M. (2001). Using the telephone for mental health surveys: An analysis of the impact of call rank, non-response and interviewer effect: *Population*, (6), 987-1010.

Wagner, J., & Stoop, I. (2017). Comparing nonresponse and nonresponse biases cross-nationally. In T. P. Johnson, B-E. Pennell, I. A. L. Stoop, & B. Dorer

(Eds.), *Advances in comparative survey methods: Multicultural, multinational and multiregional contexts (3MC).* New York, NY: John Wiley & Sons.

Wan, X., Raymond, H. F., Wen, T., Ding, D., Wang, Q., Shin, S. S., ... & Novotny, T. E. (2013). Acceptability and adoption of handheld computer data collection for public health research in China: A case study. *BMC Medical Informatics and Decision Making*, *13*(1), 1.

West, B. T., Ghimire, D., & Axinn, W. G. (2015). Evaluating a modular design approach to collecting survey data using text messages. In *Survey Research Methods* (Vol. 9, No. 2, p. 111). NIH Public Access.

Further Reading

Bethlehem, J., & Schouten, B. (2016). Nonresponse error: detection and correction. In C. Wolf, D. Joye, T. W. Smith, & Y. Fu (Eds.), *The SAGE handbook of survey methodology*. Thousand Oaks, CA: Sage.

Blom, A. G. (2016). Survey fieldwork. In C. Wolf, D. Joye, T. W. Smith, & Y. Fu (Eds.), *The SAGE Handbook of survey methodology*. Thousand Oaks, CA: Sage.

de Leeuw, E., & Berzelak, J. (2016). Survey mode or survey modes? In C. Wolf, D. Joye, T. W. Smith, & Y. Fu (Eds.), *The SAGE handbook of survey methodology*. Thousand Oaks, CA: Sage.

Fortin-Rittberger, J., Howell, D., Quinlan, S., & Todosijevic, B. (2016). Supplementing cross-national surveys with contextual data. In C. Wolf, D. Joye, T. W. Smith, & Y. Fu (Eds.), *The SAGE handbook of survey methodology*. Thousand Oaks, CA: Sage.

Loosveldt, G., & Joye, D. (2016).Defining and assessing survey climate. In C. Wolf, D. Joye, T. W. Smith, & Y. Fu (Eds.), *The SAGE handbook of survey methodology*. Thousand Oaks, CA: Sage.

Roberts, C. (2016). Response styles in surveys: Understanding their causes and mitigating their impact. In C. Wolf, D. Joye, T. W. Smith, & Y. Fu (Eds.), *The SAGE handbook of survey methodology*. Thousand Oaks, CA: Sage.

Schnell, R. (2016). Record linkage. In C. Wolf, D. Joye, T. W. Smith, & Y. Fu (Eds.), *The SAGE handbook of survey methodology*. Thousand Oaks, CA: Sage.

Stähli, M. E., & Joye, D. (2016). Incentives as a possible measure to increase response rates. In C. Wolf, D. Joye, T. W. Smith, & Y. Fu (Eds.), *The SAGE handbook of survey methodology*. Thousand Oaks, CA: Sage.

Stoop, I. A. L. (2016). Unit nonresponse. In C. Wolf, D. Joye, T. W. Smith, & Y. Fu (Eds.), *The SAGE handbook of survey methodology*. Thousand Oaks, CA: Sage.

Paradata and Other Auxiliary Data

Mengyao Hu, 2016

Introduction

The use of survey auxiliary data (including paradata) to investigate and reduce survey errors has gained tremendous attention in survey research because of the wide-range of information these data provide about the survey data collection process.

Survey Errors

As discussed in Survey Quality, there are frequently methodological, organizational, and operational barriers to ensuring quality, especially in multinational, multicultural, or multiregional surveys, which we refer to as “3MC” surveys. Various errors, such as nonresponse error, measurement error, coverage error, and sampling error, can threaten the final survey estimates. Different approaches can be taken to enhance survey data quality at each stage of the survey lifecycle. In addition to procedures to standardize and improve the survey data collection process (Kreuter, 2013), quantitative methods (e.g., benchmark data and simulation studies) and qualitative methods (e.g., cognitive interviews and focus groups) have also been used to investigate error sources (see Survey Quality for more information).

With the significant development of technology, auxiliary data, such as paradata, have become widely available to researchers, providing additional tools to evaluate and reduce survey error sources.

Auxiliary Data and Paradata

There is no one universally accepted definition of “auxiliary data”; it has often been defined as all data except the survey data itself (Kreuter, 2013; Smith, 2011). Sampling frame data, data resulting from linkage to secondary datasets, and paradata all fall into the category of auxiliary data.

Paradata has been widely discussed and applied during both data collection and analysis. Couper introduced the term “paradata” into the survey research methodology field (Groves & Couper, 1998) and the definition of paradata has vastly expanded since then. As discussed in the 2011 International Nonresponse Workshop (Smith, 2011), two main types of paradata are available. One is process paradata, which is collected during the process of data collection, (e.g., time stamps and keystroke data); the other is related to observational information (e.g., the observed demographics of respondents and observed neighborhood conditions). Smith (2011) notes that some paradata, like interviewers’ observations about respondents’ characteristics, can fall into both categories.

The types of available and commonly used auxiliary data, including paradata, vary with survey mode. See Appendix 1 for a detailed description of different types of paradata and other auxiliary data associated with survey modes. The discussion in this chapter includes paradata and other auxiliary data, which have been used to investigate and reduce survey errors.

Aims of this Chapter

Paradata and other auxiliary data have been collected and well documented in many surveys. For example, the European Social Survey (ESS) closely monitors the survey process, collects various types of paradata using contact forms (Stoop, Matsuo, Koch, & Billiet, 2010), and documents the paradata for each wave of the survey.

Increased access to paradata and other auxiliary data enables researchers to investigate survey error sources in many different dimensions. For example, recordings of the interaction between the interviewer and respondents at the doorstep can help to reveal reasons for survey nonresponse; keystroke data, such as that indicating a change in the recorded answer, can help to inform potential measurement errors; and contact history records for random walk sampling can be used to evaluate coverage error (see guidelines below).

In addition to the investigation of errors, paradata is often used in responsive designs. In this case, researchers continually monitor selected paradata to inform the error-cost tradeoff in real-time, and use this as the basis for altering design features during the course of data collection or for subsequent waves. For example, to reduce survey error, researchers can implement interventions (e.g., providing additional interviewer training) based on paradata-derived indicators during real-time data collection.

Note that in addition to investigating survey errors and monitoring data collection to reduce survey errors, paradata and other auxiliary data can be used for substantive studies. For example, interviewer observational data on graffiti collected in the ESS can be used to study survey error sources (e.g., whether neighborhoods with more graffiti are less likely to respond to surveys) and to investigate substantive questions (e.g., whether it is predictive of resident satisfaction and residents' plans to move). For the purpose of this chapter, we only focus on the investigation and reduction of different survey error sources.

This chapter aims to provide an introduction to the use of paradata for studying and reducing various survey errors. This chapter follows closely Kreuter (2013). For a more comprehensive discussion of the use of paradata and other auxiliary data to investigate and reduce survey errors, see Kreuter (2013).

Guidelines

Goal: Consider different ways to use paradata to study and reduce nonresponse, measurement, coverage, and sampling error, which are discussed in turn below.

A. Nonresponse error

1. Use paradata and other auxiliary data to investigate nonresponse error.

Rationale

The model of the biasing effect of nonresponse includes two components: the response rate and the differences between respondents and nonrespondents. The former refers to the proportion of eligible sample units who complete an interview. The latter refers to the magnitude of the differences between respondents and nonrespondents on measures of interest (e.g., mean differences of a survey estimate). If there is no difference between respondents and nonrespondents, then there is no nonresponse bias, regardless of the size of the response rate. If nonrespondents differ from respondents, the lower the response rate, the higher the bias is likely to be (Groves 2006). The challenge of studying nonresponse bias is that it is difficult to ascertain differences between respondents and nonrespondents because little data is typically available about nonrespondents. These differences can sometimes be informed by paradata.

For example, if specific paradata are available for both respondents and nonrespondents, and can reveal information about the survey outcomes (i.e., completed interviews, refusals, etc.), they may also inform researchers about the likely differences between respondents and nonrespondents (Kreuter & Olson, 2013). This can help researchers to evaluate nonresponse bias. Examples of such paradata include call history data and interviewer observations.

Procedural steps

1.1 Investigate the paradata / auxiliary data available to study nonresponse error. It is likely to vary depending upon the mode of the survey interview.
 - 1.1.1 Interviewer observation and interviewer-household/respondent interactions are only available in interviewer-administered surveys.
 - 1.1.2 On the other hand, call history data can be collected in both interviewer-administered and self-administered surveys. For

example, in web surveys, the call history data will be related to the emails and invitations sent to sample units.

1.2 Understand the different types of paradata / auxiliary data that can be used to study nonresponse error. Several types of paradata / auxiliary data have been used to study the propensity to participate in surveys and nonresponse bias (for a detailed discussion, see Kreuter & Olson (2013)).

1.2.1 Call history data can inform researchers about:
- The date and time of each call made.
- The outcome of each call (noncontact, refusal, interview, ineligibility, etc,). For more information, refer to the American Association for Public Opinion Research (AAPOR) disposition codes for call outcomes (AAPOR, 2016).
- The number of call attempts made.
- The pattern of the call attempts (e.g., time/day of calls).
- The time between call attempts.

1.2.2 Interviewer observations may include:
- Neighborhood safety.
- Whether the sample unit lives in a multi-unit structure.
- Whether the sample unit lives in a locked building.
- Whether there is an intercom system.
- The condition of the housing unit.
- The demographic characteristics of the household.
- Proxy survey variables.

1.2.3 Recordings of the interviewer-householder interaction may capture the information about:
- The doorstep statements.
- Pitch, speech rate, and pauses of interviewers.

1.2.4 GPS data, if available, can track respondents' locations over time. Google Map data can provide general information on the neighborhood and household. Given that there are not many reports of using Google Map data to study nonresponse error, the validity of this method needs further investigation. Note that the availability of such data may depend on the coverage of Google Map Street Views. Some areas, such as China and many countries in Africa, are not well covered. This map shows the areas covered: https://www.google.com/maps/streetview/explore/
Google Map may provide the following information:
- Whether the housing unit is in a multi-unit structure.
- The condition of the housing unit (if it is a single housing unit which can be observed on the map).
- Some social economic characteristics (i.e., racial group) of the neighborhood may be inferred from the number and

types of stores or restaurants in the area (e.g., if there are several Chinese stores or restaurants, it is likely that many Chinese people may live close by).

- Whether there are many abandoned houses with broken windows in the area, indicating a lower level of safety.

1.3 Know how each type of paradata / auxiliary data can inform researchers about nonresponse error and select paradata or other auxiliary data based on the purpose of the investigation.

1.3.1 Call history data can be used to study nonresponse error in several ways:

- To study "Best time to call". "Best" here refers to the call time that yields the highest cooperation rates. For example, previous literature on the best times to call found that for landline surveys, weekday nights and weekends were better than weekday mornings and afternoons (Brick, Allen, Cunningham, & Maklan, 1996; Reimer, Roth, & Montgomery, 2012). For cell phone interviews, weekday afternoons are also a good time to call. Cultural / region differences need to be taken into consideration in evaluating the best time to call. A study conducted in the U.S. Virgin Islands, Guam, Puerto Rico, and the mainland U.S., found that the best time to make contact varied among these regions, likely due to cultural and working time differences (Ravanam, et al.,2015).
- To evaluate the relationship between cost and response rates. Additional call attempts are likely to lead to improved response rates, but also increase the field time and cost of survey operations.
- To monitor call records in real-time or on a daily basis. This can show the relationship between the average number of calls and expected response rates.
- To study nonresponse bias after data collection.

1.3.2 Interviewer observations and Google Map data can be used to study:

- Survey participation, contactability and cooperation (Kreuter & Olson, 2013; Olson, 2013). For example, interviewer observations on neighborhood safety can be used to examine survey participation – neighborhood safety is often associated with contactability and cooperation, and may reveal the reasons for low participation in certain areas (Lepkowski, Mosher, Davis, Groves, & Van Hoewyk, 2010). Researchers can send invitation letters to respondents in unsafe areas to improve their trust of the interviewers.

- Access impediments using observation data on housing units.
- Demographic characteristics of sampled households or individuals, such as age, gender, education, income, race, and ethnicity, can be predictive of survey participation (Olson, 2013). For example, Korinek, Mistiaen, & Ravallion (2005) found low response rates among individuals with high incomes.

1.3.3 Interviewer-householder/respondent interactions can be used to study:
- Survey cooperation. Interaction at the doorstep is found to be highly correlated with survey participation (Groves & Couper, 1998; Lepkowski, et al., 2010). If certain interviewers lack persuasive skills or if they make many mistakes during the doorstep introduction, special trainings can be provided.
- Topic salience of the survey. Refusals due to the lack of interest in the survey topics can be used to study survey cooperation (Groves, Presser, & Dipko, 2004). It also indicates potential risk for nonresponse bias, since in this case respondents are likely to differ from nonrespondents.

1.4 Study the relationship between paradata and nonresponse error. For example, see whether the paradata indicators are predictive of survey participation and cooperation. Some paradata, such as the interviewer/respondent interaction data, can also be used to diagnose nonresponse bias (e.g., in the evaluation of topic salience).

Lessons learned

1.1 Using paradata on interviewer and respondent interactions, Conrad, et al., (2013) studied the relationship between survey participation and the interviewer's speech behaviors. They found that an interviewer's pitch can influence survey participation. In addition, when interviewers were moderately disfluent, respondents were more likely to agree to participate in surveys.

1.2 Maitland, Casas-Cordero, and Kreuter (2009) studied the relationship between paradata and cooperation and contactability of households in the National Health Interview Survey (NHIS) sample. They found that refusals due to health related reasons were associated with important survey variables. This suggests that topic salience-related paradata can be used for studying nonresponse bias.

1.3 Blom (2011) studied nonresponse across different countries using paradata including call-record data from the European Social Survey

(ESS), data on interviewer attitudes, and doorstep behavior. They found that countries differ in their contacting and cooperation processes, which can in part be explained by interviewer effects, such as contacting strategies and doorstep behavior.

2. **Use paradata and other auxiliary data to reduce nonresponse in a responsive design framework for quality control purposes in data collection.**

Rationale

To increase response rates and / or to achieve a more balanced sample, researchers often use paradata to monitor data collection in real time, and provide interventions, often in subsequent waves, accordingly (paradata-driven responsive design). For example, researchers can use paradata and statistical algorithms to optimize calling strategies (e.g., finding the best time to call or determining how many call attempts to make) for subsequent waves of data collection.

Procedural steps

Groves and Heeringa (2006) listed the following four steps for responsive design using paradata-derived indicators:

2.1 Before starting the survey, pre-identify a set of design features that may affect nonresponse error, such as unit and item nonresponse.

2.2 Identify a set of indicators of the nonresponse error and monitor those indicators in the initial phases of data collection.
- 2.2.1 Review and identify key performance indicators (KPIs) derived from paradata (such as interviews per hour, daily completion rate, or average interview duration). See Jans, Sirkis, and Morgan (2013) for a review of the summarized KPIs.
- 2.2.2 Select the appropriate indicators.
- 2.2.3 Monitor the indicators in the initial phase of data collection on a daily or weekly basis.

2.3 Implement interventions by altering the active features of the survey, either immediately or in subsequent phases based on cost/error tradeoff decision rules.
- 2.3.1 Different types of management interventions can be implemented. The 2006–2010 National Survey of Family Growth (NSFG) implemented three different types of management interventions, as described in Wagner, et al., (2012):

- Case prioritization, which aimed at “checking whether the central office could influence field outcomes by requesting that particular cases be prioritized by the interviewers.” The prioritized cases were flagged in the sample management system, and interviewers were asked to put more effort into these cases.
- Screener week, which refers to “shifting the emphasis of field work in such a way that eligible persons (and proxy indicators of nonresponse bias for those persons) would be identified as early as possible.” In this process, with the attempt to reduce nonresponse in the screener interviews, increased efforts were made to reach previously not-contacted screener sample. In addition, demographic information collected in the screening interviews can help with sample balance.
- Sample balance, as described in Wagner, et al. (2012) is sought to “minimize the risk of nonresponse bias by endeavoring to have the set of respondents match the characteristics of the original sample (including nonresponders) along key dimensions, such as race, ethnicity, sex, and age.” In NSFG, variation in subgroup response rates was chosen as a proxy indicator for nonresponse bias. To bring the composition of the interviewed cases closer to the true population, researchers in NSFG prioritized cases from subgroups that were responding at lower rates.

2.4 Combine data from the separate design phases into a single estimator.

2.4.1 After the first three steps, data from all phases are combined to produce final survey estimates. Proper weighting procedures and instructions for variance estimations need to be provided accordingly.

Lessons learned

2.1 Researchers found that demographic factors in NSFG are predictive of key survey variables. Therefore, differences in response rate across demographic groups can be used as indicators for potential nonresponse bias. Assuming that nonresponders are missing at random (i.e., nonresponders and responders within each subgroup do not differ with respect to the survey variables being collected), equal response rates across different demographic groups will minimize the size of difference between respondents and nonrespondents. Paradata such as interviewer observations can be used as proxy indicators in a responsive design to reduce nonresponse bias (Wagner, et al., 2012)

2.2 Researchers can use interviewer-generated paradata to make judgments about the likelihood that individual sample cases will become respondents. By building predictive response propensity models using paradata, it is possible to estimate the probability that the next call on a sample case will produce an interview or not at daily level. In the NSFG, the collection of paradata and other auxiliary data began at the listing stage of the sample and ended at the last call on the last case (Groves & Heeringa, 2006).

2.3 When specific subgroups of active sample cases in the 2006-2010 U.S. NSFG were found lagging on key process indicators, interventions were implemented for quality control. For example, if the response rates of older male Hispanics lagged, interviewers could target cases in this specific subgroup, flagging them in the sample management system. Priority calls could then be assigned to these cases (Kirgis & Lepkowski, 2013).

2.4 ESS uses information extracted from various auxiliary data, such as the call records, to provide feedback to fieldwork organizations for the next round, and to analyze nonresponse. The contact forms allow them to calculate response rate across countries and compare field efforts across countries. As mentioned by Stoop, Matsuo, 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 analyze nonresponse bias. However, real-time intervention remains a challenge for ESS since many countries still use paper-and-pencil questionnaires (Stoop, Billiet, Koch, & Fitzgerald, 2010).

3. Use paradata and other auxiliary data to study nonresponse to within-survey requests.

Rationale

In addition to the traditional interviewing techniques of asking respondents questions based on a questionnaire, surveys often collect data which does not follow the question and answer format (Wagner, 2013). Examples include collecting biomeasures (such as height, weight, and blood pressure), seeking access to administrative records (such as individual-level identification numbers like driver’s license or insurance record information), and asking respondents to mail back a questionnaire left behind after an in-person interview (Sakshaug, 2013). Such requests often include asking for additional permissions and mode-switching (e.g., from a face-to-face to a mail survey, as in the leave-behind questionnaire), which is likely to produce nonresponse (Sakshaug, 2013).

Respondents' refusal to within-survey requests produce a second layer of nonresponse, which may potentially bias the survey estimates. Paradata can be used to investigate the nonresponse error, and, as discussed above, can be used for prediction and possibly to prevent and reduce nonresponse. For a more detailed review of this topic, see Sakshaug (2013).

Procedural steps

3.1 Use paradata and other auxiliary data to investigate respondents' consent to link survey administrative records, if applicable.

3.1.1 As mentioned by Sakshaug (2013), two hypothesized reasons for such refusal are: privacy / confidentiality concerns and general resistance toward the survey interview. Use Interviewer observations related to both privacy/confidentiality concerns and general resistance toward the survey interview to understand nonresponse. The Health and Retirement Study has collected interviewer observations on both, as described below (Sakshaug, 2013).

- Questions regarding privacy concerns: *"During the interview, how often did the respondent express concern about whether his/her answers would be kept confidential? (never, seldom, or often)"*
- Questions regarding general resistance toward the survey interview: *"How was respondent's cooperation during the interview? (excellent, good, fair, poor)"*

3.1.2 Use call history information, such as *"ever refused to participate in the survey*," "*total number of call attempts for a completed interview",* and *"whether the respondent was a nonrespondent in previous waves*," to predict general resistance.

3.1.3 Create a paradata-based index, and use the index to predict respondent likelihood to consent to link survey administrative records.

3.1.4 Provide interventions in subsequent waves of the survey:

- When paradata indicates privacy/confidentiality concerns, flag those cases in the system, and ask interviewers to provide more justification about why the administrative records are needed and how confidentiality of the data is ensured.
- When paradata indicates the likelihood of general resistance, provide interventions such as asking the interviewer to emphasize that this will be a less

burdensome process. Also, using a shorter questionnaire can help.

3.2 Use paradata and other auxiliary data to study respondents' consent to collect biomeasures.

3.2.1 Paradata such as interviewer observations (e.g., whether the respondent asked how long the interview would last), call record information (e.g., the number of contact attempts in each wave), and time stamps can be used to predict the likelihood respondents will consent to biomeasure collection.

3.2.2 Model biomeasure consent using paradata, and evaluate whether they are predictive of respondents' likelihood to consent.

3.2.3 For future waves / surveys, provide interventions for those with a low likelihood to consent:

- Identify respondents with a low likelihood to consent using selected paradata. For example, those who were nonrespondents in previous waves or who required many calls before participating may have a lower likelihood of consenting.
- Provide interventions such as (1) increased incentives; (2) justifications of the importance of biomeasure data collection.

3.3 Use paradata and other auxiliary data to study respondents' consent to the switch of data collection modes.

3.3.1 The following paradata can be used to predict the likelihood of respondents refusing or dropping out during the mode-switch.

- Item nonresponse rate in the initial mode.
- If contact information, such as email address or cell-phone number, is needed for the mode-switch, whether the essential contact information was provided.
- The amount of time that elapsed during the interview using the first mode.

3.3.2 Model refusals and dropouts in the mode-switch process using paradata, and evaluate whether they are predictive of respondents' likelihood to consent and perform the mode-switch.

3.3.3 For future waves, researchers can pre-identify and intervene with persons unlikely to take part in the mode switch:

- Use paradata, such as item nonresponse in the initial mode, to pre-identify respondents with a low likelihood to take part in the mode switch.
- Provide interventions such as (1) incentivizing persons flagged with a low likelihood to participate, (2) if response is critical, precluding the switch to a self-administered

mode and resuming the interview in the initial interviewer-administered mode. Note that the measurement advantages of self-administration would be lost if applying this intervention.

Lessons learned

3.1 Sakshaug, Couper, Ofstedal, and Weir (2012) utilized different types of paradata (primarily interviewer observations and call record information) to study consent to the linkage to survey administrative records. A paradata-derived index was constructed and found to be strongly predictive of the consent outcome.

3.2 Sakshaug, Couper, and Ofstedal (2010) used data from the 2006 HRS to analyze the relationship between paradata and consent to biomeasures. The paradata items used to predict consent included interviewer observations, call record information, and time stamps. Overall, holding respondent demographic and health characteristics, and interviewer attributes constant, the interviewer observations and call record history data were strongly predictive of consent.

3.3 Sakshaug and Kreuter (2011) used paradata from a telephone survey of University of Maryland alumni to predict mode switch response. In the main study, alumni were contacted by telephone and administered a brief screening interview. Eligible respondents were then randomly assigned to one of three mode groups: interactive voice recognition (IVR), Web, or continuation of interview via the telephone. They found that the paradata were predictive of respondents' participation in the mode-switch. For example, item missing data in the screening interview was found predictive of mode switch dropout for the web group. Whether or not an email address was provided was also predictive of mode switch dropout.

4. Use paradata and other auxiliary data for nonresponse adjustment.

Rationale

One strategy to minimize the effect of nonresponse error is to use nonresponse adjustment weights in the post-processing of the survey data. Traditional variables used to construct nonresponse adjustment weights are those available on informative sampling frames. For example, for a survey on students at a school, the age, gender and grade information are known. In a panel survey, information from previous waves can also be used to create nonresponse adjustment weights. In many surveys, an informative frame may be unavailable, and variables on the frame may have small effects on survey estimates (Maitland, et al., 2009).

Most recently, researchers have used auxiliary data (e.g., variables collected about the sample households such as interviewer observations or Google Map data) in the weighting process for nonresponse adjustment. The procedural steps and examples are described below. For a more comprehensive review on the use of paradata for nonresponse adjustment, see Olson (2013).

Procedural steps

4.1 Understand the characteristics of an ideal auxiliary variable for sample-based nonresponse adjustment. Four characteristics are summarized in Olson (2013).
- 4.1.1 Non-missing values of the variable must be available for both respondents and nonrespondents.
- 4.1.2 The variable should be measured completely and without error for all sampled units.
- 4.1.3 This variable should be strongly associated with important survey variables of interest.
- 4.1.4 This variable should also be a strong predictor of survey participation, thus reducing nonresponse bias in an adjusted mean.

4.2 Identify paradata that are available for both respondents and nonrespondents.

4.3 Select paradata / auxiliary data -derived variables which can predict survey participation and which are associated with the survey variables of interest (Olson, 2013). Examples of such variables can be:
- 4.3.1 The sample unit's neighborhood.
- 4.3.2 The sample unit's housing unit.
- 4.3.3 Persons in the sampled housing unit.
- 4.3.4 Call record information collected as part of the sample management system.
- 4.3.5 Observations recorded by interviewers about their interaction with the sampled household at each contact.

4.4 Develop unit nonresponse adjustment weights using selected paradata / auxiliary data for nonresponse adjustments.

Lessons learned

4.1 Olson (2013) found that the challenges of using paradata for nonresponse adjustment depend on how well the paradata meet the four characteristics of auxiliary variables mentioned in Guideline 4.1. There are several reasons for this. First, the paradata itself may be subject to item nonresponse, posing many unknown questions to researchers. For example, researchers need to decide whether to impute these item missing data on paradata or not; what imputation methods to use; how to create weights if imputation is done. Second, no single variable can be a strong predictor for both participation and the survey variable of interest. Therefore, multiple paradata in addition to other auxiliary variables are often used when creating weights. Third, measurement error in paradata itself needs further research. For example, interviewer observations on neighborhood, household unit, or interactions with respondents and other household members can be subject to interviewer effects.

4.2 As mentioned above, in practice, it is challenging to identify variables that relate with both propensity to respond and the survey outcome variables. Previous literature found that paradata, as studied in several surveys, can have a very low correlation with the survey outcome variable (Kreuter, et al., 2010). Wagner, Valliant, Hubbard, and Jiang (2014) examine the use of the level-of-effort paradata (e.g., number of calls, ever refused) for nonresponse adjustment in HRS data. The model comparison results reveal that the level-of-effort paradata are predictive of the probability of response. However, they are not predictive of key survey outcomes. Therefore, they are excluded from the model (Wagner, et al., 2014). Biemer, Chen, and Wang (2013) also evaluated the use of level-of-effort paradata on nonresponse adjustment, and their model fails to remove nonresponse bias. Their paper argues that this may be attributable to errors associated with the paradata (Biemer, et al., 2013).

4.3 Kreuter, Lemay, and Casas-cordero (2007) examined whether interviewer observation data from the ESS could be useful for nonresponse adjustments in three selected countries – Greece, Portugal and Poland. They compared differences in point estimates for data weighted only with selection weights and those weighted with selection weights and a nonresponse adjustment weight based on interviewer observation. They found that the interviewer observations affected the survey point estimates only when there was a correlation between the survey variables and the key survey statistics, and there were small changes in point estimates. Weak correlations were found between response propensity, survey outcome and interviewer observations, which, as mentioned by the authors, may be due to the high interviewer variability.

B. Measurement error

5. Use paradata and other auxiliary data, such as audit files produced by computer-assisted interviewing (CAI) software, to monitor

interviewer performance and identify those interviewers who could benefit from more training.

Rationale

Non-standardized interviewer behavior (such as interviewing too quickly or too slowly, not reading the questions as worded, or using improper probes) may introduce measurement errors in surveys. Paradata and other auxiliary data, such as time stamp and behavior-coding data, can help to monitor such behavior over time. If some interviewers are not following instructions, interventions can be applied to reduce the measurement error attributable to these interviewers.

Procedural steps

5.1 Identify a set of indicators of measurement error (e.g., very short interview time), which may be affected by interviewers. More detailed discussions on interviewer performance indicators can be found in Durand (2005), Laflamme and St-jean (2011), and West and Groves (2013). Examples include:

5.1.1 Time stamps. Interviews that are unusually short or unusually long suggest potential measurement error. The source may be the respondent. Respondents may rush through the interview or provide many “Don’t know” responses -- or respondents of low cognitive ability may take more time understanding and answering questions. However, interviewers can also be the source of measurement error indicated by short or long interview times. For example, interviewers may intentionally skip certain questions or read very fast -- or, they may even falsify interviews.

5.1.2 Behavior coding. Behavior coding, developed from audio recordings of the interview, can detect whether interviewers follow standardized interviewing instructions. Examples include whether they are reading the question exactly as worded; whether they are reading at a proper speed, not too fast or too slow; and whether they are providing appropriate probing.

5.1.3 Questionnaire routing. Routings may be inconsistent with the instrument introductions. Or an interviewer may purposely lead the respondents into a specific skip pattern in order to lessen the interview time by skipping follow-up questions. If in an interpenetrated sample design (see Interviewer Recruitment, Selection, and Training) where respondents are randomly assigned to interviewers, such falsifications may be detected by comparing the questionnaire routings followed by a specific interviewer with those followed by other interviewers.

5.1.4 Keystrokes. This type of paradata can reveal whether interviewers press certain function keys, whether they change their answers, and how they navigate. Behavior such as constantly changing answers may indicate measurement error.

5.1.5 How many times error or help windows were displayed. If the error or help windows display very often for a specific interviewer, it may indicate that this interviewer is not familiar with the instrument, and may introduce measurement errors.

5.1.6 Paradata about interviewers themselves, such as age, gender, pay grade, skills, experience, and personality traits. Such demographic information can be obtained from separate data collection exercises (Turner, 2013).

5.2 Monitor the selected indicators from the beginning of data collection both at an aggregate level and individual interviewer level. Based on the indicators, researchers can not only see trends (e.g., whether interviewers spend more or less time per completed interview as time goes on), but can also evaluate the performance of each interviewer (West & Groves, 2013).

5.3 Identify interviewers who deviate from standardized interviewing. If irregularities or falsified data are discovered for a specific interviewer, he or she should be flagged, and appropriate interventions conducted.

5.4 Implement interventions, such as providing more training, monitoring future outcomes, and looking for improvement by those interviewers.

Lessons learned

5.1 West and Groves (2013) propose and evaluate a calculated interviewer performance indicator which can be used in all interviewer-administered survey modes. This indicator gives each interviewer a score reflecting his or her performance. It gives more weight to successful interviews on difficult cases versus relatively easy cases. Their paper reports that “calling-history paradata are the strongest predictors of obtaining interviews in both face-to-face and telephone interviews”.

5.2 One example of using paradata and other auxiliary data to monitor interviewer behavior is the Saudi National Mental Health Survey, where interviewers are evaluated in real-time based on various indicators (question field time, failed verifications, long pauses, etc.)

Interviewers are ranked based on these indicators and the "worst" three are flagged and examined in detail (Mneimneh, Pennell, Lin, & Kelley, 2014).

5.3 The China Mental Health Survey uses auxiliary data including paradata to monitor the survey process. Interviews are flagged if (1) the response time for a certain number of variables is below a predetermined threshold; (2) the adjusted interview length is below a predetermined threshold; (3) the number or “Don’t Know” or “Refused” responses are above a predetermined threshold; or (4) interviews are completed between 11pm and 7am. Interviewers are also flagged for further investigation if the number of interviews completed is significantly higher than other interviewers, or if the distribution of selected variables deviates significantly from other interviewers (Sun & Meng, 2014).

5.4 In the Panel Study of Income Dynamics (PSID), audit trail data (ADTs) produced by the Blaise software, including interview length, average question count per interview, entry of question-level notes, use of question-level help, and backups within the interview, were used to monitor interviewer behavior (Kirgis et al., forthcoming).

5.5 To develop a “replacement mechanism for traditional interviewer evaluation techniques”, the NSFG explored the use of paradata in interviewer monitoring and evaluation. This concept was later adopted by the Health and Retirement Study, with case-level information being added and mean / median indicators being monitored. The paradata/auxiliary data-related indicators used in the evaluation process included field time, error escapes, suppressions, jumps, backups, “don’t know” and “refused” responses, help key use, and remarks entered. Three indices were constructed using these nine indicators: “too fast,” “many error checks,” and “many ‘don’t know’ and ‘refused’”. Interventions made, based on the evaluation results, included extra practice interviews, re-training on proper interviewing techniques, increasing the number of verification interviews, and group re-training and investigation at case level (Kirgis et al., 2015).

5.6 In the Ghana Socioeconomic Panel Study, keystroke data from a computer-assisted personal interview (CAPI) was used to monitor and evaluate interviewers’ performances. Given the complexity of the survey instrument, a questionnaire “dashboard” was designed to show the status of all the questionnaire sections and all the respondents within the household. This allowed interviewers to jump to any section or block of questions in the questionnaire quickly and to switch respondents easily. Keystroke data was used to evaluate interviewers' navigation patterns, and researchers also identified interviewers with a higher number of mid-section or mid-block exits—i.e., those who "jumped around too much". The increased movement within the questionnaire was also found to be related with increased interview length and cost. To reduce survey time and cost, proper interventions can be provided to the identified interviewers (Lin, Mneimneh, & Kwaiser, 2015).

6. **Use paradata and other auxiliary data to investigate measurement error in web surveys, so that real-time intervention can be applied to reduce potential measurement error.**

Rationale

Paradata and other auxiliary data in web surveys can identify the device used to complete the survey (e.g., laptop, smartphone or computer), and can inform the entire process of filling out the questionnaire (e.g., how respondents navigate and the time spent on each question). Such navigation and time stamp paradata can be used to monitor respondents' behavior in the data collection process. If the paradata indicates that some respondents are employing satisficing behaviors, special prompts can be provided to those respondents. For a more comprehensive review of the use of paradata in web surveys, see Callegaro (2013).

Procedural steps

6.1 Understand how web survey paradata / auxiliary data are collected. As described by Callegaro (2013), Heerwegh (2003), and Heerwegh (2011), paradata and other auxiliary data can be collected on the server side and/or the client side. Time stamp data, which can be collected at the page level, is one example of server-side paradata. Client-side paradata are collected at the respondent device level, and include mouse clicks and changing answers. Callegaro (2013) points out that JavaScript is essential for collecting client-side paradata. Without JavaScript, respondents can finish the survey, but the paradata cannot be collected.

6.2 Understand the types of paradata available in web surveys. Callegaro (2013) identifies two major types of web survey paradata: (1) paradata that indicate the device respondents are using (device-type paradata) and (2) paradata that reveal the navigation process respondents use (navigation paradata).
6.2.1 Device-type paradata include:
- The browser used (user agent string).
- The operating system.
- The language of the operating system.

- The screen resolution.
- The browser window size.
- Whether the browser has JavaScript.
- Adobe Flash, or other active scripting-enabled devices.
- The IP address of the device used to fill out the survey.
- The GPS coordinates at the time of beginning the interview and at the time of completion.
- Cookies (text files placed on the visitor's local computer to store information about that computer and the page visited).

6.2.2 Questionnaire navigation paradata include:

- Authentication procedures.
- Mouse clicks.
- Mouse position per question/screen.
- Change of answers.
- Typing and keystrokes.
- Order of answering.
- Movements across the questionnaire (forward/backward), scrolling.
- Number of appearances of prompts and error messages.
- Detection of current window used.
- Whether the survey was stopped and resumed at a later time.
- Clicks on non-question links (e.g., hyperlinks, frequently asked questions (FAQs), and help options).
- Last question answered before stopping the survey.
- Time spent per question/screen.

6.3 Understand the privacy and ethical issues involved in collecting web survey paradata. Some paradata can be used to identify respondents, such as IP addresses or email addresses. Researchers need to carefully protect such data (Callegaro, 2013).

6.4 Become familiar with the software used to collect web survey paradata, such as Client Side Paradata (CSP) by Dirk Heerwegh and User-ActionTracer (UAT) (Stieger & Reips, 2010).

6.5 Identify the type of operating system and the device respondents are using. This will inform researchers about measurement error, usability, and comparability issues related to different devices.

6.6 Use questionnaire navigation paradata to better understand the process respondents are using to answer the web survey. With detailed navigation paradata, researchers can even reconstruct the survey-taking experience (Callegaro, 2013).

6.7 Use navigation paradata to evaluate the quality of the web survey. For example, respondents who rush through the survey may be more likely to provide poor quality data. Quality indexes can be calculated based on available navigation paradata. For instance, in a study aimed at improving establishment surveys, Haraldsen (2005) created a quality index based on client-side paradata. The index included the number of prompts, error messages, and data validation messages.

6.8 Use paradata and other auxiliary data to identify satisficing behaviors or obstacles respondents have in web surveys so that real-time intervention can be applied to reduce potential measurement error. For example, if respondents take a long time to answer a question, given the possibility that they may be having difficulties understanding the question, a definition or clarification prompt can be provided. Similarly, if the paradata indicates that respondents are speeding through the questionnaire, a prompt can be designed to ask them to take their time and read the questions carefully (see lessons learned).

Lessons learned

6.1 Paradata can be used to guide interventions for quality control in web surveys. To improve data quality, Conrad, Couper, Tourangeau, Galesic, and Yan (2009) implemented an experiment based on the speed respondents took to answer each question in a web survey If a respondent provided an answer is less than the typical minimum reading time, the respondent was prompted with a message encouraging him or her to put more thought into the question. The study found that prompting increased completion time but reduced straightlining in grid questions, which is usually viewed as an indicator of satisficing (Conrad, et al., 2009).

7. Use paradata and other auxiliary data to investigate measurement error in non-web surveys.

Rationale

Paradata and other auxiliary data, such as behavior coding data, can help researchers better understand interviewer behavior and response characteristics which, in turn, can be related to measurement error (Yan & Olson, 2013). For example, researchers have found that doubts and/or less confident expressions in surveys, such as “I don’t know,” or “maybe”, are predictive of less accurate responses (Draisma & Dijkstra, 2004; Dykema, Lepkowsk, & Blixt, 1997; Yan & Olson, 2013). For a detailed review of this topic, see Yan and Olson (2013).

Procedural steps

7.1 Choose appropriate respondent paradata / auxiliary data as indicators of measurement error. Examples might include:

7.1.1 Response times. These can indicate potential problems in survey questions, and can indirectly reveal respondent uncertainty. Many literatures have evaluated the relationship between response time and measurement error (e.g.,Bassili & Fletcher, 1991).

7.1.2 Keystrokes and mouse clicks. These can be used to study respondents' navigation, change of responses, the response process when respondents use different designs (e.g., drop-down boxes versus radio buttons), and the use of clarification features (Yan & Olson, 2013).

7.1.3 Behavior codes. Researchers can use behavior codes to investigate the question-answering process and to study respondents' uncertainty about their answers, which can be related to measurement error (Schaeffer & Dykema, 2011). Indicators of respondent uncertainty include:

- Verbal expressions of doubt and uncertainty, such as "I am not sure" (Draisma & Dijkstra, 2004; Dykema, et al., 1997; Schaeffer & Dykema, 2011; Yan & Olson, 2013).
- Nonverbal expressions of doubt and uncertainty, such as hesitation, raised eyebrows, or averted gaze (Conrad, Schober, & Dijkstra, 2004).
- Paralinguistic and verbal cues, such as "Uhm…". (Conrad, Schober, & Dijkstra, 2008; Draisma & Dijkstra, 2004).
- Answering too quickly (Callegaro, Yang, Bhola, Dillman, & Chin, 2009; Kaminska, McCutcheon, & Billiet, 2011; Malhotra, 2008; Yan & Olson, 2013).
- Answering too slowly (Bassili & Fletcher, 1991; Bassili & Krosnick, 2000; Heerwegh, 2003; Yan & Olson, 2013).
- Changing responses (Bassili & Fletcher, 1991; Yan & Olson, 2013).

7.1.4 Interviewer evaluations. Interviewers are often asked to evaluate respondents' level of cooperation at the end of the interview. This is one of the most straightforward ways of analyzing the relationship between paradata / auxiliary data and measurement error.

7.2 Choose appropriate interviewer paradata / auxiliary data as indicators of measurement error. Examples might include:

7.2.1 Behavior codes. Researchers can use behavior codes to identify interviewer behavior deviating from standardized interviewing, which can be related to measurement error (Schaeffer & Dykema, 2011).

7.2.2 Interviewer vocal characteristics. These data, such as speech rate and pitch, are usually measured from audio recordings.

7.2.3 Interviewers' demographic characteristics and attitudes about the survey process as well as the specific project's substantive attitudes, all of which can contribute to measurement error (Mneimneh, de Jong, Cibelli-Hibben, & Moaddel, 2015).

7.3 Use paradata, especially keystroke data, to replicate respondents' navigation processes and to investigate usability issues in CAI systems. For example, it can reveal whether a special function key has been used by respondents, and whether respondents are having difficulty typing the answers to open-ended questions.

7.3.1 To reduce measurement error, use paradata and other auxiliary data to improve survey questions. "Various types of paradata such as question timings, keystroke files and audio recordings can provide an indication of respondent difficulty in answering survey questions" (Yan & Olson, 2013). Researchers can use paradata to detect survey questions with potential problems in the pretest, and then to revise and improve the questionnaire.

7.4 Use paradata to improve the survey response process during data collection (real-time responsive design). For example, researchers can use paradata to detect situations where respondents may be having difficulty answering the questions (e.g., they take a long time to respond). Systems can provide tailored clarifications for those respondents (Conrad, Schober, & Coiner, 2007).

7.5 Set standards for how these paradata / auxiliary data can be collected. See Olson and Parkhurst (2013) for methods of collecting paradata for the purpose of measurement error investigation.

7.6 Analyze paradata / auxiliary data, based on the research question. As described in Yan and Olson (2013), the analysis of paradata needs to consider the following factors:

7.6.1 Determine units of analysis. Sometimes, the paradata obtained are nested in nature. For example, "response times, mouse clicks, keystrokes, verbal behaviors, and vocal characteristics are recorded for each action taken for each question item for each respondent, nested within each interviewer for a given survey" (Yan & Olson, 2013). Different systems of aggregation can be used to organize the data. See Yan and Olson (2013) for detailed examples. Note that there is no single ideal way to choose the unit of analysis. Decisions must be made based on the specific research question.

7.6.2 Manage the data. The management of data can be different for each type of paradata (Yan & Olson, 2013).

- Response time. If response latencies (the time spent until the occurrence of an event) are of interest, researchers need to calculate "the differences in time from the beginning of an event to the end of the event" (Yan & Olson, 2013). As proposed by Yan and Olson (2013), four factors must be considered when analyzing response time:
  (1) The validity of the data, which largely depends on the survey mode: In web surveys, researchers must decide whether server-side or client-side paradata are more valid. For example, response time collected at the server-side includes upload and download time, which is generally longer than client-side data (also see Guideline 9.1.1).
  (2) The presence of outliers, which may distort the results if kept for analysis: Outliers can be defined in many ways. The most common way is based on the number of standard deviations the data point is from the mean length of time. Usually, once identified, outliers are excluded from analysis. An alternative method is to impute or use other cases to replace them.
  (3) Whether the distribution is normal or skewed: If the latter, transformation of the data can be done to deal with the skewed distribution.
  (4) Baseline adjustment: People have different cognitive abilities and may differ in their speed of talking. Thus, it is natural that some respondents simply answer survey questions more quickly than others (Yan & Tourangeau, 2008). In situations where such differences are not of research interest, it can add "noise", and increase measurement error to the response time data. Yan and Olson (2013) state, "to account for these differences, researchers can subtract a 'baseline' measurement calculated separately for each respondent from the average of response timings to multiple items external to (and usually administered prior to) the questions of interest from the response timing data for the questions of interest."
- Keystroke and mouse clicks. Management of these two types of paradata depends on the level and unit of analysis. Researchers must decide whether to analyze them at an action, question, section, or survey level. Unlike other types of paradata, they are collected as dichotomies -- "Yes (1) or No (0)". Yan and Olson (2013) recommend that researchers first evaluate whether there is enough

particular keystroke data to analyze statistically. If events are rare, similar events can be combined for analysis.

- Behavior codes. Data can be obtained in different ways. The most detailed analysis is based on transcriptions of interviews; alternately, coders can listen to the interviews and code the data. As a first step, researchers must decide whether to code from transcriptions or by listening to recordings. Second, detailed and comprehensive coding instructions are needed to guide the coders. Usually, at least two coders are required to ensure the reliability of the coding process. After coding is complete, the data must be examined and unreliable codes removed from the analysis.
- Vocal characteristics can be obtained and processed using software like Praat (http://www.fon.hum.uva.nl/praat/).
- Interviewer evaluation data can often be analyzed directly. The most common data management issues related to interviewer evaluations is item nonresponse. Methods such as multiple imputations can be considered to deal with this issue.

7.6.3 Use statistical analysis to answer the research questions. Based on the research questions, various methods of analysis are available. For example, response latencies can be analyzed by comparing mean time spent to answer specific questions. Statistical models such as survival models or logistic regression models can be employed to evaluate response latencies and other paradata such as keystroke data. In addition, based on the research questions, paradata can be used as either independent variables or dependent variables in the analysis (Yan & Olson, 2013).

Lessons learned

7.1 Draisma and Dijkstra (2004) studied the relationship between response time and response error, collecting response time via audio recordings. They found that the longer a respondent spent answering a question, the less likely the respondent was to give a correct answer.

7.2 Yan and Tourangeau (2008) studied factors influencing response time. They found that many things can affect response time, including both item-level characteristics (e.g., the length and difficulty of a question) and respondent characteristics (e.g., age, education, and experience with the internet).

7.3 Response time can be studied, not only at the individual question level, but also at the entire questionnaire level. Malholtra (2008) evaluated the relationship between measurement error and the time respondents spent completing a questionnaire. They reported that more satisficing behavior was found for low-education respondents in the fastest group.

7.4 Using various paradata, Liebe, Glenk, Oehlmann, and Meyerhoff (2015) evaluated whether the use of mobile devices (tablets and smartphones) affect survey data quality in web surveys. They found that mobile device users "spent more time than desktop/laptop users to answer the survey". A longer interview is observed when a smaller-screen device is used. Acquiescence is more likely to occur when the screen size is large.

C. Coverage and sampling error

8. Use paradata and other auxiliary data to investigate and reduce coverage and sampling errors.

Rationale

Coverage error is related to the quality of the sampling frame from which respondents are selected. Various auxiliary data including paradata can be used to evaluate the quality of sampling frames. One way to study coverage error is to use flag files in the U.S. Postal Service's Delivery Sequence File, where addresses can be flagged as vacant, institutional, or seasonal (Eckman, 2013). Another common auxiliary data is geocoding data, often used to construct the frame when there is no match between the postal geographies and the census geographies (e.g., the U.S.). The types of paradata / auxiliary data available for coverage error study depend on the frames used, and, sometimes, the sampling procedures. In random route sampling, which combines frame construction and sampling into one step, paradata, such as contact history, cannot only be used to study coverage error, but also to study sampling errors. See Eckman (2013) for a detailed discussion.

Procedural steps

8.1 Clearly define the research questions regarding the investigation of sampling error.

8.2 Based on the research question, select paradata or other auxiliary data that can indicate coverage error or sampling error. Given the sampling frame, the types of paradata / auxiliary data available may differ. As described by Eckman (2013), examples include:

8.2.1 Postal Delivery Databases. In some countries, centralized postal registers are available and can be used as sampling frames for housing units. Various studies have evaluated the quality of such frames (e.g., Wilson and Elliot (1987)). Sometimes the address is only a mail box, which works for mail surveys but not face-to-face surveys. In the United States, the U.S. Postal Service's Delivery Sequence File, which contains auxiliary data such as flags in the file, can be used to construct the frames. The flags indicate whether the address is vacant, a dormitory, or seasonal housing. When geocoding is used to construct sampling frames, it can also generate paradata related to coverage error. For example, the software of geocoding can report how precisely an address is geocoded

8.2.2 Housing Unit Listing. Housing unit listing is often done by interviewers when no postal address information is available. Interviewers are sent to selected areas to list housing units in order to create the frame for future sample selection. Paradata and other auxiliary data are closely related to the process of listing. In dependent listing, interviewers are provided with a map with an initial list of addresses, referred as an input list. Interviewers are asked to delete the inappropriate units and to confirm the correct units on the list (Eckman, 2013). Paradata or other auxiliary data, such as whether a housing unit is from the input list or added by the interviewer, can be collected for analysis. Interviewer observations on the quality of the list can also be collected in the listing process, indicating possible coverage errors. Much more paradata can be collected when technology is used in the listing process. For example, if a laptop or smart phone device is provided for listing, various paradata, like time spent for listing, keystroke data, and even GPS data can be collected.

8.2.3 Random Route Sampling. This procedure combines frame construction and sampling. Similar to housing unit listing, if technology is used, time spent for listing and sampling, keystroke data, and GPS data can be collected for studying coverage error and sampling error.

8.2.4 Missed Unit Procedures. To prevent undercoverage in household surveys, procedures like half-open interval are used to find and select units missing from the sample frame. Flags of these added cases can be collected to evaluate coverage error.

8.2.5 Telephone frames. Telephone frames can suffer from overcoverage when randomly selected phone numbers are ineligible units (e.g., business telephone numbers). Undercoverage can occur when some people have no phone numbers. Casady and Lepkowski (1993) propose a list-assisted methodology where "banks of 100 consecutive numbers are assigned a score reflecting how many numbers in that bank also appear in the directory of listed phone numbers" (Casady & Lepkowski, 1993; Eckman, 2013). Auxiliary data from the fame construction process, such as the bank-level score, can be used to study coverage error.

8.2.6 Household rosters. This step of sampling is essential when a survey unit is an individual, but the frames available are at household unit level. Various methods can be used for within-household selection, such as next birthday, oldest male/youngest female, or a random selection based on a full roster of all members of the household (Eckman, 2013). Paradata collected in the roster process, such as those that record the behavior of the interviewers, can be used to detect errors related with household rosters.

8.2.7 Population Registers. An alternative method of selecting respondents is to directly use population registers. Few countries have population registers, but for those that do, they are an appealing way for survey researchers to draw a sample. In Sweden, registers include the time that each record is updated, which can be used as an indicator of precision in the records. In some countries, the registers are not centralized. For example, in Germany, each community has its own registers. In this case, interactions with the organizations at each community can be collected to study coverage error (Eckman, 2013).

8.2.8 Subpopulation frames. In some cases, the target population is not the entire population (e.g., a survey of adult females only). Usually, a screening interview is required to filter out ineligible cases when no register data is available. In this case, paradata indicating the process of the screening interviews can be used to study coverage error.

8.2.9 Web surveys. Despite their increasing popularity, web surveys suffer from many coverage errors. Commonly used frames for web surveys include mail, telephone, and in-person interviews. To help those who do not know how to use computers or how to access a website, researchers sometimes provide training programs. Auxiliary data indicating which cases are provided with training can be used to study whether including such cases can reduce undercoverage bias (Eckman, 2013).

8.3 Set standards for how these paradata or auxiliary data can be collected. The collection methods may differ based on the survey mode and the types of frames used. For example, some auxiliary data can be obtained from organizations that provide the frames (e.g., the organizations at each community in Germany who keep the registers). Some other auxiliary data including paradata, such as computer-generated paradata, need to be obtained using specific software.

8.4 Collect the paradata or auxiliary data and analyze them.

8.5 Use statistical methods to evaluate coverage or sampling error using paradata or other auxiliary data. Similar to using paradata to investigate measurement errors, statistical models such as survival models or logistic regression models can be employed to evaluate the relationship between paradata (or other auxiliary data) and coverage and sampling error. In addition, based on research questions, paradata or other auxiliary data can be used as either independent variables or dependent variables in the analysis.

Lessons learned

8.1 Shore, Montaquila, and Hsu (2010) compared the eligibility of cases added by interviewers via a missed unit procedure to those on the original frame. They found that the units added through the missed unit procedure were more likely to be vacant units than cases already on the frame.

8.2 Alt, Bien, and Krebs (1991) used contact history records and compared cases selected through random walks and through a population register. They found that fewer calls were needed in the random walk sample to complete an interview. It is likely that interviewers select those cases which are easier to reach, which may lead to coverage and sampling error.

8.3 Dahlhammer (2009) used keystroke files to study the process that interviewers use to take rosters. This study found that sometimes interviewers went back after the respondent selections were finished (e.g., to change the number of household members). The author interprets this as interviewers trying to select a more cooperative household member. Those who are excluded from the rosters are undercovered on the frames of household members.

D. Quality of Paradata and Other Auxiliary data

9. Develop a quality control framework for paradata and other auxiliary data.

Rationale

Data quality is known have a critical impact on the final survey statistics in research. It is not surprising that results can be biased when the variables used in analysis are subject to measurement error. This is also true with paradata and other auxiliary data. Poor quality paradata / auxiliary data can lead to biased estimates in post-survey analysis, such as biased nonresponse adjustments (West & Sinibaldi, 2013). In the data collection process, if paradata or other auxiliary data prone to measurement error is used in a responsive design to help researchers make intervention decisions, it is likely that the interventions will not be as useful as expected, survey cost will be increased instead of decreased, and survey data quality will be diminished (West & Sinibaldi, 2013). Therefore, developing a quality control framework for paradata / auxiliary data collection and analysis is of critical importance to researchers who would like to make good use of such information.

Procedural steps

9.1 Review the characteristics and collection procedures of different types of paradata / auxiliary data, and understand the nature of the paradata / auxiliary data quality.

9.1.1 Computer-generated paradata

- Very few peer-review studies have examined the quality of computer generated paradata. In general, it is thought they are of good quality.
- Although technical issues are rare (West & Sinibaldi, 2013), they do happen and could lead to failure to collect automatically generated paradata (Lenzner, Kaczmirek, & Lenzner, 2010).
- As mentioned in Section 6, paradata can be collected on the server side and/or the client side. Note that the quality collected from each side may be different. Take the response time as an example. West and Sinibaldi (2013) point out that, unlike client-side paradata, “server-side response time measures are captured by the server sending the web survey to the respondent’s computer and collecting the data, and therefore include uploading and downloading times.” Thus, for the same questions, the response time data collected through server-side will likely

be longer than that collected from client-side. This is consistent with findings from Yan and Tourangeau (2008).

- Paradata involving the interaction between respondents and the computer may be prone to measurement error. For example, as mentioned in West and Sinibaldi (2013), problems with response timings are more common in computer-assisted telephone interviews (CATI) using an automatic voice sensor , where the sensors may be “tricked” by respondents asking questions or engaging in other vocal behaviors that do not represent answering a question, such as thinking aloud. Such response time paradata can be prone to measurement error.

9.1.2 Interviewer-generated paradata

- In computer-assisted interviews, paradata is usually directly entered into the computer system. In paper and pencil surveys, however, interviewer-generated paradata is recorded on paper, and then entered into a computerized dataset. Editing error can be introduced during this process.
- Interviewer-recorded paradata can be subject to errors. For example, previous literature found that interviewers are likely to underreport call attempts in CAPI data collection, such as failing to report “drive-by” visits (Biemer, et al., 2013; West & Sinibaldi, 2013).

9.2 For computer-generated paradata, develop procedures to ensure all programming works as designed and to prevent the occurrence of technical issues. Develop clear protocols on how to construct paradata-derived indicators,

9.3 For interviewer-generated paradata, develop clear protocols regarding how to record paradata, and protocols for coders on how to code interviewer-generated paradata in the dataset. In the training process, provide clear instructions to interviewers and coders on the collection and coding of paradata.

9.4 Develop quality examination procedures for different types of paradata. West and Sinibaldi (2013) have introduced different methods of examining the quality of paradata, such as using “ ‘gold standard’ validation data …, as well as indirect indicators of the quality, including reliability, ease of collection, and missing data issues.” For more information, refer to West and Sinibaldi (2013).

9.5 If any paradata is found highly subject to measurement error, provide interventions for future waves / surveys to improve data quality.

Lessons learned

9.1 In a study comparing interviewer observations on housing status with respondent self-reports, “accuracy rates for interviewer observations ranged from 46% for privately rented dwellings to 89% for owner-occupied dwellings.” This implies non-negligible measurement error with interviewer observations (Pickering, Thomas, & Lynn, 2003; West & Sinibaldi, 2013).

E. Documentation

10. Document how paradata and other auxiliary data are collected and the steps used to construct the paradata / auxiliary data-based indicators.

Rationale

The collection and construction of paradata and other auxiliary data are not usually documented in surveys. Documentation will help give secondary data users a clear picture about the data collection and variable construction process, which will help them to use or analyze paradata / auxiliary data in a more efficient way.

Procedural steps

10.1 Document how each type of paradata / auxiliary data is collected, such as whether time stamp data is collected through the server side or the client side.

10.2 If any indicator is constructed based on paradata, such as response time calculated from time stamp data, document clearly how those variables are constructed.

10.3 Document the coding procedure, if there is any, for interviewer-generated paradata, such as open-ended descriptions of the neighborhood.

10.4 If paradata is used in a responsive design, document clearly how paradata is used to monitor and inform intervention in the data collection process.

10.5 Document clearly how the paradata provided can be linked to the main survey data. Provide instructions on how to use the paradata-derived indicators. Ensure that respondent confidentiality is maintained when any paradata files are available for use.

Lessons learned

10.1 The National Health Interview Survey provides a detailed documentation for the paradata collected (NHIS, 2014).

10.2 ESS provides a well-documented introduction of the contact form data they release for many countries (European Social Survey, 2014).

Appendix 1. Examples of uses of paradata and other auxiliary data to investigate measurement, nonresponse, coverage, and sampling error

Means of Collection	Type of Paradata	Mode of data collection*	Measurement error	Nonresponse error	Coverage and/or Sampling Error
Audio recording of the interview	Behavior coding	Face-to-face, Telephone	Studying interviewer and respondent interactions	Effect of interviewers introduction on response rate	
	Vocal characteristics	Face-to-face, Telephone	Studying interviewer and respondent interactions, respondents' uncertainty	Effects of interviewers' doorstep speech on nonresponse error	
Interviewer presence	Interviewer observation and/or evaluations	Face-to-face, Telephone	Provides insights into what is going wrong in a survey)	Neighborhood observations	Variables related to frame Quality, such as "locked buildings or communities"
Location / area of the respondents' home known	Google Map data	Face-to-face or Mail: Actual address can be used for coding Google Map data Telephone: Area code of the telephone number can be used for coding Google Map data		Neighborhood observations through Google Maps	Neighborhood observations through Google Maps
GPS devices or GPS-enabled cell phones	GPS data	Face-to-face, Telephone (using a mobile telephone), web surveys		Association between respondents living habits like working time at office and nonresponse	Can be used to study coverage error in random route sampling
Computerized questionnaires	Time stamps	Face-to-face (computer-assisted personal interviews (CAPI)), Telephone (computer-assisted telephone interviews (CATI)), Web surveys	Effects of time to complete the survey on survey data quality	Time stamps in previous wave can be used to predict the likelihood respondents will consent to subsequent wave / studies	Time interviewers spent for listing and sampling

	Keystroke	Face-to-face (CAPI), Telephone (CATI), Web surveys	Number and frequency of changing answers)	Keystroke and mouse clicks in previous wave can be potentially used to predict the likelihood respondents will consent to subsequent wave / studies	Keystroke and mouse clicks data at interviewer side – whether they go back and changed the within household selection procedure
	Mouse Clicks	Face-to-face (CAPI), Telephone (CATI), Web surveys	Can be used to study respondents' navigation, change of responses, the response process when different designs are used)		
	Information on the operating system, web browser, screen resolution	Web surveys	Potential measurement error related different operating systems)	Effects of web browser and screen resolutions on nonresponse error	
Lab settings	Eye-tracking		Can be used to study satisficing behavior, such as whether respondents read all response categories		

References

Alt, C., Bien, W., & Krebs, D. (1991). Wie zuverlässig ist die Verwirklichung von Stichprobenverfahren? Random route versus Einwohnermeldeamtsstichprobe. Retrieved from http://www.ssoar.info/ssoar/handle/document/20981

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

Bassili, J. N., & Fletcher, J. F. (1991). Response-time measurement in survey research a method for CATI and a new look at nonattitudes. *Public Opinion Quarterly, 55*(3), 331-346.

Bassili, J. N., & Krosnick, J. A. (2000). Do strength-related attitude properties determine susceptibility to response effects? New evidence from response latency, attitude extremity, and aggregate indices. *Political Psychology, 21*(1), 107–132.

Biemer, P. P., Chen, P., & Wang, K. (2013). Using level-of-effort paradata in non-response adjustments with application to field surveys. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 176*(1), 147–168.

Blom, A. G., de Leeuw, E. D., & Hox, J. (2011). Interviewer effects on nonresponse in the European Social Surveys. *Journal of Official Statistics*, *27*(2), 3–53.

Brick, J. M., Allen, B., Cunningham, P., & Maklan, D. (1996). Outcomes of a calling protocol in a telephone survey. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 142–249.

Callegaro, M. (2013). Paradata In Web surveys. In F. Kreuter (Ed.), *Improving surveys with paradata* (pp. 263-282). Hoboken, NJ: John Wiley & Sons.

Callegaro, M., Yang, Y., Bhola, D. S., Dillman, D. A., & Chin, T. Y. (2009). Response latency as an indicator of optimizing in online questionnaires. *Bulletin de Méthodologie Sociologique, 103*(1), 5–25.

Casady, R. J., & Lepkowski, J. M. (1993). Stratified telephone survey designs. *Survey Methodology, 19*(1), 103–113.

Conrad, F. G., Schober, M., & Dijkstra, W. (2004). Non-verbal cues of respondents' need for clarification in survey interviews. *ASA. Survey Methods Research*, pp. 4934–4950.

Conrad, F. G., Schober, M. F., & Coiner, T. (2007). Bringing features of human dialogue to Web surveys. *Applied Cognitive Psychology, 21*(2), 165–187.

Conrad, F. G., Schober, M., & Dijkstra, W. (2008). Cues of communication difficulty in telephone interviews. In J. M. Lepkowski, C. Tucker, J. M. Brick, E. D. de Leeuw, L. Japec, P. J. Lavrakas, … R. L. Sangster (Eds.), *Advances in telephone survey methodology* (pp. 212–230). Hoboken, NJ: John Wiley & Sons.

Conrad, F., Couper, M., Tourangeau, R., Galesic, M., & Yan, T. (2009). *Interactive feedback can improve accuracy of responses in Web surveys*. Paper presented at the European Survey Research Association (ESRA) Conference,Warsaw, Poland.

Conrad, F. G., Broome, J. S., Benkí, J. R., Kreuter, F., Groves, R. M., Vannette, D., & McClain, C. (2013). Interviewer speech and the success of survey invitations. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 176(1), 191-210.

Dahlhammer, J. (2009). *Using paradata to assess and monitor data quality in the National Health Interview Survey (NHIS)*. Presented at the Statistics Canada/U.S. Census Bureau Methodological Interchange. Ottawa, Ontario, Canada.

Draisma, S., & Dijkstra, W. (2004). Response latency and (para)linguistic expressions as indicators of response error. In S. Presser, M. P. Couper, J. T. Lessler, E. Martin, J. M. Rothgeb, & E. Singer (Eds.), *Methods for testing and evaluating survey questionnaires* (pp. 131–147). Hoboken, NJ: John Wiley & Sons.

Durand, C. (2005). Measuring interviewer performance in telephone surveys. *Quality and Quantity, 39*(6), 763–778.

Dykema, J., Lepkowsk, J. M., & Blixt, S. (1997). The effect of interviewer and respondent behavior on data quality: Analysis of interaction coding in a validation study. In L. E. Lyberg, P. P. Biemer, M. Collins, E. D. de Leeuw, C. Dippo, N. Schwarz, & D. Trewin (Eds.), *Survey measurement and process quality* (pp. 287-310). Hoboken, NJ: John Wiley & Sons.

Eckman, S. (2013). Paradata for coverage research. In F. Kreuter (Ed.), *Improving surveys with paradata* (pp. 97–120). Hoboken, NJ: John Wiley & Sons.

European Social Survey. (2014). *Sampling for the European Social Survey Round VII: Principles and requirements*. Retrieved May 5, 2016, from http://www.europeansocialsurvey.org/docs/round7/methods/ESS7_sampling_guidelines.pdf

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

Groves, R. M., & Couper, M. P. (1998). *Nonresponse in household interview surveys*. New York, NY: John Wiley & Sons.

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., Presser, S., & Dipko, S. (2004). The role of topic interest in survey participation decisions. *Public Opinion Quarterly, 68*(1), 2–31.

Haraldsen, G. (2005). *Using client side paradata as process quality indicators in Web surveys - WebSM*. ESF Workshop on Internet survey methodology, Dubrovnik. Retrieved August 17, 2015 http://www.websm.org/db/12/11428/rec/

Heerwegh, D. (2003). Explaining response latencies and changing answers using client-side paradata from a Web survey. *Social Science Computer Review, 21*(3), 360–373.

Heerwegh, D. (2011). Internet survey paradata. In M. Das, P. Ester, & L. Kaczmirek (Eds.), *Social and behavioral research and the internet. Advances in applied methods and research strategies* (pp. 325–348). London: Routledge.

Jans, M., Sirkis, R., & Morgan, D. (2013). Managing data quality indicators with paradata-based statistical quality control tools: The keys to survey performance. In F. Kreuter (Ed.), *Improving surveys with paradata* (pp. 191-229). Hoboken, NJ: John Wiley & Sons.

Kaminska, O., McCutcheon, A. L., & Billiet, J. (2011). Satisficing among reluctant respondents in a cross-national context. *Public Opinion Quarterly, 74*(5), 956–984.

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 F. Kreuter (Ed.), *Improving surveys with paradata* (pp. 121-144). Hoboken, NJ: John Wiley & Sons, Inc.

Kirgis, N., Mneimneh, Z., Sun, Y., Lin, Y., & Ndiaye, S. K. (2015) *Using paradata to monitor interviewer behavior and reduce survey error*. A presentation given at 2015 International Total Survey Error Conference, Baltimore, MD, USA.

Korinek, A., Mistiaen, J. A., & Ravallion, M. (2005). Survey nonresponse and the distribution of income. *The Journal of Economic Inequality, 4*(1)33–55.

Kreuter, F. (2013). Introduction. In F. Kreuter (Ed.), *Improving surveys with paradata* (pp. 1-9). Hoboken, NJ: John Wiley & Sons.

Kreuter, F., & Olson, K. (2013). *Paradata for nonresponse error investigation*. University of Nebraska - Lincoln, Sociology Department, Faculty Publications. Paper 220.

Kreuter, F., Lemay, M. L., & Casas-cordero, C. (2007). Using proxy measures of survey outcomes in post-survey adjustments: Examples from the European Social Survey (ESS). *Section on Survey Research Methods* (1983), 3142–3149.

Kreuter, F., Olson, K., Wagner, J., Yan, T., Ezzati-Rice, T. M., Casas-Cordero, C., … Raghunathan, T. E. (2010). Using proxy measures and other correlates of survey outcomes to adjust for non-response: Examples from multiple surveys. J*ournal of the Royal Statistical Society: Series A (Statistics in Society), 173*(2), 389–407.

Laflamme, F., & St-jean, H. (2011). Proposed indicators to assess interviewer performance in CATI survey. *Proceedings of the Joint Statistical Meetings,* 118–28.

Lenzner, T., Kaczmirek, L., & Lenzner, A. (2010). Cognitive burden of survey questions and response times: A psycholinguistic experiment. *Applied Cognitive Psychology, 24*(7):1003–1020.

Lepkowski, J. M., Mosher, W. D., Davis, K. E., Groves, R. M., & Van Hoewyk, J.. (2010). The 2006-2010 National Survey of Family Growth: Sample design and analysis of a continuous survey. *Vital and Health Statistics. Series 2, Data evaluation and methods research, 150,* 1–36. Retrieved April 10, 2015 (http://europepmc.org/abstract/med/20928970).

Li, L. (2011) Retrieved on May 19, 2016, from http://www.lukeexperiments.com/content/index/media/Luke_Li_Eye_Tracking_Methodology.pdfLiebe, U.,Glenk, K., Oehlmann, M., & Meyerhoff, J. (2015). Does the use of mobile devices (tablets and smartphones) affect survey quality and choice behaviour in Web surveys? *Journal of Choice Modelling, 14,* 17–31.

Lin, Y., Mneimneh, Z. N., & Kwaiser, K. (2015). *Using paradata to monitoring interviewers' behavior: Case studies from national household surveys in Kingdom of Saudi Arabia and Ghana*. The 6th Conference of the European Survey Research Association (ESRA) Conference.

Maitland, A., Casas-Cordero, C., & Kreuter, F.. (2009). An evaluation of nonresponse bias using paradata from a health survey. *Proceedings of the Section on Survey Research Methods of the American Statistical Association,* 370–378.

Malhotra, N. (2008). Completion time and response order effects in Web surveys. *Public Opinion Quarterly, 72*(5), 914–934. R

Mneimneh, Z. N., Pennell, B., Lin, Y., & Kelley, J. (2014). *Using paradata to monitor interviewers' behavior: A case study from a national survey in the Kingdom of Saudi Arabia*. Comparative Survey Design and Implementation (CSDI) Conference.

Mneimneh, Z., De Jong, J., Cibelli-Hibben, K., & Moaddel, M. (2015). *Interviewer effects: Gender, Islamic hijab, and respondents' sociopolitical and cultural attitudes in a nationally representative survey in Tunisia*. Paper presented at the American Association for Public Opinion Research in Hollywood, Florida, May 16-17.

NHIS. (2014). 2014 NHIS Paradata File. Retrieved on May 14, 2016, from http://www.cdc.gov/nchs/nhis/2014paradata.htm

Olson, K. (2013). Paradata for nonresponse adjustment. *The Annals of the American Academy of Political and Social Science, 645*(1), 142–170.

Olson, K., & Parkhurst, B. (2013). Collecting paradata for measurement error evaluations. In F. Kreuter (Ed.), *Improving surveys with paradata* (pp. 43-72). Hoboken, NJ: John Wiley & Sons.

Pickering, K., Thomas, R., & Lynn, P. (2003). *Testing the shadow sample approach for the English House Condition Survey.* London, U.K.: National Center.

Ravanam, M. S., Bauroth, S., Ward, C., Seither, R., Nyaku, M. K., Hill, H. A., & Elam-Evans, L. D. (2015). *Operational challenges for national data collection that includes U.S. territories*. International Field Directors and Technologies Conference.

Reimer, B., Roth, V., & Montgomery, R. (2012). Optimizing call patterns for landline and cell phone surveys. *Proceedings of the Section on Survey Research Methods. American Statistical Association*, 4648–4660.

Sakshaug, J. W. (2013). Using paradata to study response to within-survey requests. In F. Kreuter (Ed.), *Improving surveys with paradata* (pp. 171–190). Hoboken, NJ: John Wiley & Sons.

Sakshaug, J. W., & Kreuter, F. (2011). Using paradata and other auxiliary data to examine mode switch nonresponse in a "recruit-and-switch" telephone survey. *Journal of Official Statistics*, *27*(2), 339–357.

Sakshaug, J. W., Couper, M. P., & Ofstedal, M. B. (2010). Characteristics of physical measurement consent in a population-based survey of older adults. *Medical Care*, *48*(1), 64–71.

Sakshaug, J. W., Couper, M. P., Ofstedal, M. B., & Weir, D. R. (2012). Linking survey and administrative records: Mechanisms of consent. *Sociological Methods & Research*, *41*(4), 535–569.

Schaeffer, N. C., & Dykema, J. (2011). Response 1 to Fowler's chapter: Coding the behavior of interviewers and respondents to evaluate survey questions. In J. Madans, K. A. Miller, & G. Willis, *Question evaluation methods: Contributing to the science of data quality (*pp. 23–39). Hoboken, NJ: John Wiley & Sons.

Shore, S. P., Montaquila, J. M., & Hsu, V. (2010). The presence and characteristics of households at addresses obtained by traditional field listing and from USPS-based lists. *Section on Survey Research Methods, ASA* (pp. 1709–1718).

Smith, T. W. (2011). The report of the international workshop on using multi-level data from sample frames, auxiliary databases, paradata and related sources to detect and adjust for nonresponse bias in surveys. *International Journal of Public Opinion Research, 23*(3), 389-402.

Stieger, S., & Reips, U. (2010). What are participants doing while filling in an online questionnaire: A paradata collection tool and an empirical study. *Computers in Human Behavior, 26*(6), 1488–1495.

Stoop, I., Billiet, J., Koch, A., & Fitzgerald, R. (2010). *Improving survey response: Lessons learned from the European Social Survey*. John Wiley & Sons.

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.

Sun, Y., & Meng, X. (2014). *Using response time for each question in quality control on China Mental Health Survey (CMHS)*. Comparative Survey Design and Implementation (CSDI) Conference.

Turner, M. (2013). *New perspectives on interviewer-related error in surveys: Application of survey paradata*. University of Southampton, Social Sciences, Doctoral Thesis.

Wagner, J. (2013). Using paradata-driven models to improve contact rates in telephone and face-to-face surveys. In F. Kreuter (Ed.), *Improving surveys with paradata* (pp. 145-170). Hoboken, NJ: John Wiley & Sons.

Wagner, J., West, B. T., Kirgis, N., Lepkowski, J. M., Axinn, W. G., & Ndiaye, S. K. (2012). Use of paradata in a responsive design framework to manage a field data collection. *Journal of Official Statistics, 28*(4), 477–499.

Wagner, J., Valliant, R., Hubbard, F., & Jiang, L. (2014). Level-of-effort paradata and nonresponse adjustment models for a national face-to-face survey. *Journal of Survey Statistics and Methodology, 2*(4), 410–432.

West, B. T., & Groves, R. M. (2013). A propensity-adjusted interviewer performance indicator. *Public Opinion Quarterly, 77*(1), 352–374.

West, B. T., & Sinibaldi, J. (2013). The quality of paradata : A literature review. In F. Kreuter (Ed.), *Improving surveys with paradata* (pp. 339-359). Hoboken, NJ: John Wiley & Sons.

Wilson, P. R., & Elliot, D. J. (1987). An evaluation of the postcode address file as a sampling frame and its use within OPCS. *Journal of the Royal Statistical Society. Series A (General), 150*(3), 230-240.

Yan, T., & Olson, K. (2013). Analyzing paradata to investigate measurement error. In F. Kreuter (Ed.), *Improving surveys with paradata* (pp. 73-95). Hoboken, NJ: John Wiley & Sons.

Yan, T., & Tourangeau, R. (2008). Fast times and easy questions: The effects of age, experience and question complexity on Web survey response times. *Applied Cognitive Psychology, 22*(1), 51–68.

Data Harmonization

Peter Granda and Emily Blasczyk, 201 6

Introduction

Harmonization refers to all efforts that standardize inputs and outputs in multinational, multicultural, or multiregional surveys, which we refer to as “3MC” surveys.

Harmonization is a generic term for procedures used predominantly in official statistics that aim at achieving, or at least improving, the comparability of different surveys and measures collected. The term is closely related to that of standardization (see Sample Design and Questionnaire Design). Harmonizing procedures may be applied in any part of the survey lifecycle, such as study design, choice of indicators, question wording, translation, adaptation, questionnaire design, sampling, data collection, data coding, data editing, or documentation. The need to harmonize arises for all 3MC surveys. This is particularly true if the goal is to combine the data into a single integrated dataset.

There are two general approaches for harmonizing data: input harmonization and output harmonization:

1. Input harmonization aims to achieve standardized measurement processes and methods in all national or regional populations included in the 3MC survey. Comparability can be realized through standardization of definitions, indicators, classifications, training, and technical requirements.

2. Output harmonization begins with different national or regional measurements, possibly derived from non-standardized measurement processes. These measurements are “mapped” into a unified measurement scheme. Thus, only the statistical outputs are specified, the individual countries/regions may decide how to collect and process the data necessary to achieve the desired outputs. (Granda, Wolf, & Hadorn, 2010; Lyberg & Stukel, 2010; Multinational Time Use Study [MTUS], 2014; Doiron, Raina, L’Heureux, & Fortier, 2012). It is also possible to incorporate output harmonization in the original planning to produce datasets for 3MC research as the Luxembourg Income Study has done for many years with both individual and household level data collected from many countries since 1983 (http://www.lisdatacenter.org/).

Guidelines

Goal: To ensure that survey and statistical research teams follow accepted standards when creating harmonized data and documentation files, and use a harmonization strategy that best fits their basic source materials and the objectives they wish to achieve.

1. **Decide what type of harmonization strategy to employ, taking into account that many harmonization efforts will require some combination of strategies.**

   Consider "input" harmonization when the survey process is centrally coordinated.

   Rationale

   "Input" harmonization, usually applied in a multi-national context, seeks to impose strict standards and protocols from the beginning for the whole survey process by which each national survey applies the same survey procedures and a common questionnaire (see Sample Design and Translation). Also known as "prospective," this strategy is meant to assure a high degree of comparability. Some adaptations may occur for individual data collection sites, but the goal is to maintain comparability (Doiron et al., 2012).

   Procedural steps

   1.1 Provide detailed specifications, protocols, and procedures for all aspects of the survey process. The different specifications (Data Protocol, Sampling, Translation, etc.) of the European Social Survey (ESS) and the Demographic and Health Surveys (DHS) Toolkit are good examples (ESS, 2010; The Demographic and Health Surveys Toolkit, 2014).

   1.2 Decide what items to standardize.

   1.3 Consider if variations may be necessary to account for site-specific interests. This can either be due to site specific research foci or resource limitations (Doiron et al., 2012).

   Consider "output" harmonization, also known as retrospective harmonization, when the survey collection process is largely determined at the level of individual countries or cultures and there is minimal or no agreement on standardization.

   Rationale

   This type of harmonization is implemented through two main strategies: "ex-ante" and "ex-post." In practice, a study may utilize both strategies.

Ex-ante refers to measurements designed to be comparable and harmonized in data processing. When comparability has been considered during survey planning, the understanding of concepts, common goals, and specific targets can be established for the data collection process. The precise wording of the survey items may vary but the items seek to capture the same concept (see Questionnaire Design and Adaptation).

The second variant is an ex-post strategy, by which statistical or survey data are deemed inferentially equivalent and made comparable after the fact through a conversion process (Fortier et al., 2011a). The items to be harmonized were not designed to be comparable, but are assessed and edited to achieve commonality. An ex-post strategy can be used in situations where existing repositories will be exploited for comparative research or where intensive early planning is not possible because of financial or policy constraints.

Procedural steps

1.4 Use an ex-ante strategy whenever possible. This enhances comparability since harmonization is addressed at the planning stage of each national data collection, as well as at the end of the process when creating harmonized data files.

1.5 Implement an appropriate planning process.

1.6 Use an ex-post strategy only if no consideration regarding harmonization has been given by data collectors at the start of data collection(s), but researchers later believe (e.g., because of common concepts or similar questions across surveys) that a harmonized data file can be produced through a conversion process to create comparable variables or statistics. The Integrated Public Use Microdata Series, International (IPUMS-I) and The Integrated Fertility Survey Series (IFSS) are two such examples (www.international.ipums.org); (IFSS, 2014).

    1.6.1 For any ex-post plan, ensure that data access, intellectual property, and any other ethical or legal issues are resolved for all intended source studies prior to beginning harmonization with the source in question (Fortier et al., 2011a). Even if study investigators have their data publicly available, it is advisable to obtain permission from them if planning to harmonize their data with other datasets. An individual study's data use agreement may not apply, and a formal request to the respective research ethics or data access committees may be necessary (Doiron et al., 2012).

1.7 Record all decisions about the "conversion" process systematically. One option is to use two separate databases to record all work: a production database which stores the original and harmonized materials, and a user's database which provides the analysts access to the overall process.

1.8 Make provisions so that all data conversions can be traced back to the original data.

1.9 For any output harmonization technique, adopt a detailed "data processing plan" that includes descriptions of how the producer(s) of the harmonized data deal with the following:
   1.9.1 Differences in study design, such as panel or cross-sectional design, and/or in mode of data collection.
   1.9.2 Differences across studies with regard to what is measured (e.g., definitions of study population, concepts, variables).
   1.9.3 Differences in how to measure (e.g., scale of measurement, wording and routing of questions, respondents asked).
   1.9.4 Differences in how estimates are generated (imputation, weighting, or nonresponse adjustments).
   1.9.5 Procedures used to create and define harmonized variables, including any harmonized weights calculated.

Lessons learned

1.1 Input harmonization involves adherence to appropriately standardized methodologies throughout the survey lifecycle. For example, the ESS seeks to collect data every other year, uses face-to-face interviews, aims to collect high-precision data, applies detailed sampling and fieldwork protocols, uses standardized translation protocols in all participating countries, aims to achieve standardized response rates, adopts consistent coding procedures, and creates and distributes well-documented datasets in a timely fashion. All of these procedures require greater organizational capabilities and resources throughout the planning and data collection stages. The results are transparent, high quality, and can produce more valuable public-use data files at the end.

1.2 Not all comparative research will be able to follow the same procedures, so it is important to decide which methods are best, given the actual resources, survey process structure, and the intended level of precision. In addition, the creation of such common standards and their implementation at the local level requires considerable expertise. This also may not be available in all 3MC contexts. The Generations and Gender Programme is a large, longitudinal 3MC survey that studies relationships between parents and children and also between partners. It is conducted using both a paper-and-pencil instrument (PAPI) as well as computer-assisted Interviewing (CAPI) and seeks to follow consistent harmonization practices. While much harmonization work occurs centrally, individual country teams are urged to follow certain procedures to improve comparability. This method requires considerable coordination among components of the survey teams at all levels (Kveder & Galico, 2014).

1.3 Flexibility can be designed. Research sites in different countries may not be able to follow the same procedures, so it is important to decide which methods can be adapted and define procedures for adapting given the actual resources, survey process structure, and the intended level of precision. For example, the Malaria Indicator Survey is an optional component that can be conducted with or without biomeasure collection (www.dhsprogram.com). The creation of such common standards and their implementation at the local level requires considerable expertise. This also may not available in all 3MC contexts.

1.4 In a working paper, Roland Gunther describes in detail the harmonization efforts surrounding the European Community Household Panel (ECHP) (Gunther, 2003). This survey was designed from the beginning to use input harmonization, with its design of uniform questionnaires as well as detailed definitions, rules, procedures, and models to make comparability across nations easier, and is exemplary of the use of input harmonization. After the first phase of the project, a few countries decided to cease collecting national samples for the ECHP, and instead to conduct their own national surveys, resulting in the need to do ex-post harmonization. Those doing the harmonization work learned that this kind of ex-post harmonization was resource-intensive and required staff experienced in both the original source and target formats of the ECHP framework. They also had to know in detail how their national questionnaires differed. Common problems included concepts heavily affected by national contexts, as well as differences in scales of measurement, variable coding schemes, and definitions of concepts. Solutions to such problems were often found through ad hoc decisions about recoding, combining, or collapsing variables, and almost never through estimation techniques.

1.5 These harmonization strategies are almost never applied exclusively on any single statistical or survey data collection. Depending on specific cultural and national characteristics, data producers should consider strategies that will enable them to collect their data in the most efficient manner. In some situations, they may want to combine strategies. For example, data producers may start with an input harmonization plan, but should be prepared to do some ex-post output harmonization to account for differences across cultures. For example, the Demographic and Health Surveys has standardized instruments but also provides a Standard Recode Manual (DHS, 2015).

1.6 Health researchers, in particular, emphasize the importance of ex-post output harmonization. Because of the volume of datasets generated by national governments and individual investigators which affect public health policies, the desire to pool cases cross-nationally to increase sampled sizes is highly desired. To insure comparability investigators involved in this process developed a very systematic approach to harmonization and encouraged its use throughout relevant research communities.

1.7 Output harmonization projects also generate copious amounts of metadata describing the source variables, target variables, and the entire harmonization process. This new metadata provides researchers with opportunities to analyze this information and create additional linkages. For example, individual variables can be grouped into substantive categories or concepts to enhance the analytical power of a new, harmonized dataset.

2. **When deciding which variables to harmonize, create an initial plan and define clear objectives about what you want to achieve. The plan should include making all data conversions reversible.**

Rationale

Creating a harmonization plan from the beginning of the project allows data producers to document all of their decisions at the time they are made. In case errors occur or are identified by users at a later time, all data conversions should be reversible.

Procedural steps

2.1 Before fieldwork, consult with experts or an advisory committee on a systematic design process, and with methodology groups to investigate comparability issues. If pre-fieldwork coordination is not possible, form an advisory committee of researchers knowledgeable about the subject matter at the beginning of the harmonization process, if possible, and consult with them regularly.

2.2 Show the advisory group results of the harmonization process at different points in the process to allow for possible changes in rules used to create new variables.

2.3 Consider establishing a testing group of users knowledgeable about the subject matter separate from the harmonization process, who provide feedback on the analytic usefulness of the data before they are released publicly.

2.4 Implement a systematic conversion creation process with appropriate quality controls.

2.5 Identify and become familiar with software tools that facilitate a comparison of variables from different surveys, in order to determine if and how these could be harmonized. Such tools often work from a common database that stores all the information about each variable.

2.6 Establish partnerships with producers of harmonization tools. This may be more beneficial than creating new tools, which often requires costly programming efforts.

2.7 Where software tools are unavailable or impractical, use manual comparisons in making harmonization decisions and consult with substantive and methodological experts in doing so.

2.8 Identify and become familiar with interactive documentation tools that allow for proper and transparent documentation. For example, the DataSHaPER (http://www.p3g.org/biobank-toolkit/datashaper) and Opal (https://www.maelstrom-research.org/what-we-offer/open-source-software) are tools designed to harmonize epidemiological data.

Lessons learned

2.1 Realize that not all concepts measured in the survey process are equally amenable to harmonization efforts. For example, cross-national harmonization of the number of births and marriages is a far easier task than comparisons of divorce rates where local laws, customs, and data collection methods may differ substantially. Other concepts, such as international population migration, may not, due to a lack of precise definition and great variety in measurement criteria, lend themselves to harmonization at all, or only at the most basic level. Three characteristics that could influence harmonization potential are (i) the relative importance to the research intending to use the harmonized items, (ii) the individual the item targets (for example the participant, the participant's family members), and (iii) the period of time to which the variable refers. (Fortier et al., 2011b).

2.2 Good decision-making about the harmonization process will benefit from the use of software tools, as well as input from a diverse group of survey researchers who can offer advice on various procedures and techniques to use when producing harmonized files. The ISSP Data Wizard (German Social Science Infrastructure Services, 2010) was used by the International Social Survey Programme (ISSP). It was one of the first tools developed to support procedures that were previously performed manually to harmonize data at the cross-national level. The tool offered rule-based checks, automation of partial steps, and the visualization of certain conditions, to make the harmonization process more efficient, easier, and less susceptible to mistakes.

2.3 The European Values Study (EVS) formed a number of work groups, both before and after fieldwork. The aim on the one hand was to set standards at an early stage, and on the other to consolidate and merge data which had been cleaned by participating national survey teams. This project produced an integrated source questionnaire and a set of equivalency tables to assist secondary researchers. The project web site makes all of this information easily accessible (EVS, 2014). These processes and products provide critical information to secondary users of these data.

2.4 The DataSchema and Harmonization Platform for Epidemiological Research (DataSHaPER) is one potential tool for output harmonization. Fortier's 2011 paper showed that using the DataSHaPER across 53 studies, 64% of "essential" constructs from those selected could be harmonized completely or "partially." This estimate used the most conservative criteria and evaluation of harmonization potential would likely improve this statistic (Fortier et al., 2011b). A newer version of this tool is Opal (https://www.maelstrom-research.org/what-we-offer/open-source-software).

3. **Focus on both the variable and survey levels in the harmonization process.**

Rationale

Harmonization efforts usually concentrate on comparing and integrating information involving specific variables across data files. However, it is equally important to consider the overall characteristics of the surveys that make them good candidates for harmonization, and to report the decisions involving this process to end users.

Procedural steps

3.1 Recognize the different aspects involved in converting source variables, which might include variable concepts or scales of measurement, into target variables. The concept of citizenship, for example, presents significant challenges to researchers who want to investigate this topic (Minkel, 2004).

3.2 Describe similarities and differences between the source variables and the target variables, including discussion of universe statements, question wording, coding schemes, and missing data definitions. There may be an unavoidable loss of information resulting from harmonization, such as if a variable that was continuous is being harmonized with a categorical variable (Fortier et al., 2011b).

3.3 Consider file-level attributes when creating the harmonized data file, including how survey weights, imputation procedures, variance estimation, and key substantive and demographic concepts will change in the process.

3.4 Pay particular attention to sampling designs and data collection methods in making assessments about the degree of comparability between different surveys. See Survey Quality for a discussion of how quality profiles can be developed and used to assess comparability in a 3MC survey.

Lessons learned

3.1 Data producers must recognize the degrees of individual item or variable persistency when creating questionnaires and collecting data. Item persistency over time is very important in generating harmonized data files. There are considerable differences, for example, between an "absolute" persistent variable, such as "country of birth," and a less persistent variable, such as "country of citizenship." The concept might mean different things in different countries, is subject to change, and could be reported validly for multiple countries by some respondents (Minkel, 2004).

3.2 Quota sampling destroys comparability. (Häder & Gabler, 2003; Heeringa & O'Muircheartaigh, 2010). Harmonization will not make data from quota sampling comparable with data gathered via probability sampling. The ISSP is an example of a 3MC survey program that abolished quota sampling.

3.3 The European Social Survey (ESS) provides detailed insight into weighting issues and makes this information available. See the ESS data site for each survey round for the latest version.

3.4 The Collaborative Psychiatric Epidemiology Surveys (2014) created a harmonized data file from three comparable surveys on mental health. Data producers created a pooled weight for the harmonized file, based on race/ancestry groupings and on the geographic domains of the sampling frames of each individual survey. Understanding the specific characteristics of each input file was an essential part of creating a harmonized output file (Heeringa & Berglund, 2007). All of this information was provided to users in a comprehensive explanation of the original and harmonized weights.

4. **Develop criteria for measuring the quality of the harmonization process. This includes testing it with users knowledgeable about the characteristics of the underlying surveys, the meaning of source variables, and the transformation of source variables into target variables.**

Rationale

Researchers may analyze harmonized files in new and unexpected ways. It is crucial to provide them sufficient information about the concepts and definitions presented, and the assumptions underlying the decisions made in their construction.

Procedural steps

4.1 Devise procedures to judge the quality of the harmonized outputs based on such quality criteria as consistency, completeness, and comparability.

4.1.1 Consistency can be judged by comparing the results from multiple independent efforts of harmonizing a variable; completeness is assessed based on the degree to which the original information is preserved in the harmonized data; and comparability is the degree to which the harmonized outputs can accurately report important social or economic concepts over time or between countries or cultures.

4.1.2 The Statistical Office of the European Communities (EUROSTAT) proposed the following set of quality criteria when reporting statistics which also apply to harmonization outputs (Database of Integrated Statistical Activities, 2014; Joint UNECE/EUROSTAT Work Session on Statistical Data Confidentiality, 2009):

- Relevance of the statistical concepts.

- Accuracy of the estimates.
- Topicality and timeliness of the dissemination of results.
- Accessibility and clarity of the information.
- Comparability of the statistical data.
- Coherence.
- Completeness.

4.1.3 Strictly speaking, these traits apply to official statistical data. However, many of them would apply equally to academically produced survey data, particularly those regarding the comparability of social, economic, and demographic concepts in a 3MC context, and the accuracy of estimates.

4.2 Be prepared to modify and update harmonized datasets after public release, based on comments from the research community, if errors are uncovered, or if certain variables need further explanation.

4.3 Prepare presentations at social science research conferences that describe the harmonization process to potential users.

Lessons learned

4.1 The usefulness of well-harmonized data is clearly recognized by many international organizations. The United Nations Economic and Social Council recognized the importance of harmonizing environmental data collection activities in order to produce comparable indicators on the environment and its relationship to the economy. They determined to bring the System of Environmental-Economic Accounts (SEEA) to an international statistical standard. The SEEA now provides the first international standard for environmental-economic accounting (United Nations Economic and Social Council, 2005; see also http://unstats.un.org/unsd/envaccounting/seea.asp).

5. Provide the widest range possible of data and documentation products about the entire harmonization process.

Rationale

Regardless of whether utilizing input or output harmonization as a strategy, all aspects of the survey planning, collection, and dissemination process should be considered when producing harmonized data files or creating accompanying documentation. Users should have access not only to the harmonized end result, but also to detailed information about all steps taken by the producers, as well as source materials, in order for them to fully understand what decisions were made during the entire process.

Procedural steps

5.1 Define the elements of the harmonization process and start documenting it from the beginning in order to ensure that all decisions are captured even before a definite plan to produce a public-use data file exists.

5.2 To the greatest extent possible, document each target variable with information from all source variables, transformation algorithms, and any deviations from the intended harmonized approach.

5.3 If possible, provide users with access to the original data files used in producing the harmonized file. If direct access to original data is not permissible due to confidentiality concerns, implement procedures to assist users in proper check-backs or re-transformations. Also consider implementing some form of restricted-use data agreement to allow access under controlled conditions.

5.4 Prioritize providing users with the code or syntax used in creating new variables for the harmonized file.

5.5 Provide users with as complete as possible documentation, including crosswalks, which describe all the relationships between variables in individual data files with their counterparts in the harmonized file. An interactive, web-based documentation tool is often the best way to present such documentation.
   5.5.1 Include original questionnaires and information about the data collection process whenever possible.

5.6 Report on as many of the following elements of the data lifecycle as it applies to the particular harmonization process:
   Study Design and Operational Structure:
      5.6.1 Project planning.
   Sample Design, Questionnaire Design and Instrument Technical Design:
      5.6.2 Sampling frame.
      5.6.3 Sample size.
      5.6.4 Sample design (See Instrument Technical Design, Questionnaire Design, Sample Design).
      5.6.5 Duration of the field period.
      5.6.6 Instrument construction and design.
   Adaptation of Survey Instruments and Translation:
      5.6.7 Translation and adaptation (See Translation).
   Data Collection:
      5.6.8 Mode(s) of interview.
      5.6.9 Respondent follow-up if panel survey.

5.6.10 Data collection methods (See Data Collection: Face-to-Face Surveys, Data Collection: Telephone Surveys, Data Collection: Self-Administered Surveys, and Survey Quality).

Data Processing and Statistical Adjustment:

5.6.11 Editing.
5.6.12 Item nonresponse.
5.6.13 Unit nonresponse.
5.6.14 Any special treatment given to demographic and country-specific variables.
5.6.15 Sample weights.
5.6.16 Variance estimation.
5.6.17 Data production, including both planned and ad-hoc decisions implemented during variable conversion.
5.6.18 Documentation production.

Data Dissemination:

5.6.19 Dissemination (See Data Dissemination).

This list is based on documentation provided in the Integrated Health Interview Series (IHIS). The IHIS is an effort to provide an assortment of variables from the core household and person level files from the National Center for Health Statistics' seminal data collection effort on the health conditions for the US population from 1969 to the present. It provides extensive user notes and FAQ pages to describe how their harmonization project coped with several of these components (Integrated Health Interview Series, 2014).

5.7 Consider archiving the original and harmonized data with a trusted data archive to ensure continued availability of all data and documentation files and long-term preservation. See Data Dissemination for additional discussion regarding archiving.

Lessons learned

5.1 The Eurobarometer Survey Series, in operation since 1973, now includes several dozen cross-sectional surveys, all of which have been harmonized into single cross-national files before being made available to researchers. These surveys are released initially with basic information about each study and the characteristics of all variables, and are then further processed by the social science data archives, led by German Social Sciences Infrastructure Services, to include variable frequencies, more complete documentation, and online analysis services for researchers (Eurobarometer Survey Series, 2014). Such partnerships between data producer and social science data archives encourage long-term preservation, enhance access, and make it possible to continually improve services to the research community.

5.2 Some harmonization projects have gone to great lengths to describe their procedures in specific detail. For example, the Multinational Time Use Study (MTUS) has a User Guide and a comprehensive description of its coding procedures used in creating its harmonized data file (MTUS, 2014). Similarly, the Generations and Gender Programme (GGP) of the United Nations Economic Commission for Europe Population Activities Unit (UNECE-PAU) provides reports and guidelines about how the organization implements its harmonization decisions (Kveder & Galico, 2014). These projects provide transparency to both creators and users of these data and serve as an example for others to follow.

References

Collaborative Psychiatric Epidemiology Surveys. (2014). Retrieved November 04, 2014, from http://www.icpsr.umich.edu/CPES

The Data Schema and Harmonization Platform for Epidemiological Research (DataSHaPER). (2014). Retrieved November 05, 2014 from http://www.p3g.org/biobank-toolkit/datashaper

Database of Integrated Statistical Activities. (2014). *Database of Integrated Statistical Activities activities by statistical topic or domain: 4.1 Metadata (Eurostat)*. Retrieved November 04, 2014, from http://www1.unece.org/stat/platform/display/DISA/4.1+Metadata+(Eurostat)

The Demographic and Health Surveys Program (DHS). (2015). *Survey process*. Retrieved June 14, 2016, from http://dhsprogram.com/What-We-Do/Survey-Process.cfm

The Demographic and Health Surveys Toolkit. (2014). Accessed November 04, 2014, from https://www.k4health.org/toolkits/dhs

Doiron, D., Raina, P., L'Heureux, F., & Fortier, I. (2012). Facilitating collaborative research: Implementing a platform supporting data harmonization and pooling. *Norsk epidemiologi*, 21(2).

Eurobarometer Survey Series. (2014). Retrieved November 05, 2014, from http://www.gesis.org/eurobarometer-data-service/survey-series/

European Social Survey (ESS). (2010). Retrieved May 03, 2010, from http://www.europeansocialsurvey.org

European Values Study (EVS). (2014). Retrieved November 05, 2014, from http://www.europeanvaluesstudy.eu/

Fortier, I., Doiron, D., Burton, P., & Raina, P. (2011a). Invited commentary: consolidating data harmonization—how to obtain quality and applicability? *American Journal of Epidemiology*, 174(3), 261-264.

Fortier, I., Doiron, D., Little, J., Ferretti, V., L'Heureux, F., Stolk, R. P., & Burton, P. R. (2011b). Is rigorous retrospective harmonization possible? Application of the DataSHaPER approach across 53 large studies. *International Journal of Epidemiology*, 40(5), 1314-1328.

German Social Science Infrastructure Services (GESIS). (2010). *ISSP DataWizard*. Retrieved November 05, 2014, from http://www.gesis.org/issp/issp-members-area/module-standard-setups/module-standard-setups-history/

Granda, P., Wolf, C., & Hadorn, R. (2010). Harmonizing survey data. In J. A. Harkness, M. Braun, B. Edwards, T. Johnson, L. E. Lyberg, P. Ph. Mohler, B-E., Pennell & T. W. Smith (Eds.), *Survey methods in multinational, multicultural and multiregional contexts* (pp. 315-332). Hoboken, NJ: John Wiley & Sons.

Gunther, R. (2003). *Report on compiled information of the change from input harmonization to ex-post harmonization in national samples of the European Community Household Panel – Implications on data quality* (Working Paper #19). Retrieved November 05, 2014, from https://www.destatis.de/DE/Methoden/Methodenpapiere/Chintex/ResearchResults/Downloads/WorkingPaper19.pdf?__blob=publicationFile

Häder, S., & Gabler, S. (2003). Sampling and estimation. In J. A. Harkness, F. J. R. van de Vijver & P. M. Mohler (Eds.), *Cross-cultural survey methods* (pp. 117-136). Hoboken, NJ: John Wiley & Sons.

Heeringa, S., & Berglund, P. A. (2007). Integrated weights and sampling error codes for design based analysis. *National Institutes of Mental Health Collaborative Psychiatric Epidemiology Survey Program Data Set. User Guide*.

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. P. Mohler, … T. W. Smith (Eds.), *Survey methods in multicultural, multinational, and multiregional contexts* (pp. 251–267). Hoboken, New Jersey: John Wiley & Sons.

Integrated Fertility Survey Series (IFSS). (2014). Retrieved November 05, 2014, from http://www.icpsr.umich.edu/icpsrweb/IFSS/

Integrated Health Interview Series (IHIS). (2014). Retrieved November 05, 2014, from http://www.ihis.us/ihis/

Integrated Public Use Microdata Series, International (IPUMS) (2015). Retrieved July 28, 2015 from https://international.ipums.org/international/

Joint UNECE/EUROSTAT Work Session on Statistical Data Confidentiality. (2009). Retrieved November 05, 2014, from http://www.unece.org/stats/documents/2009.12.confidentiality.htm

Kveder, A. & Galico A. (2014). *Guidelines for cleaning and harmonization of generations and gender survey data*. Retrieved November 04, 2014, from http://www.unece.org/fileadmin/DAM/pau/_docs/ggp/GGP_2008_DCHGuide.pdf

Luxembourg Income Study. (2015). Retrieved October 7, 2015 from http://www.lisdatacenter.org/

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. Johnson, L. E. Lyberg, P. Ph. Mohler, B-E., Pennell & T. W. Smith (Eds.), *Survey methods in multinational, multicultural and multiregional contexts* (pp. 227-249). Hoboken, NJ: John Wiley & Sons.

Minkel, H. (2004). *Report on data conversion methodology of the change from input harmonization to ex-post harmonization in national samples of the European Community Household Panel – Implications on data quality* (Working Paper #20). Retrieved November 05, 2014, from https://www.destatis.de/DE/Methoden/Methodenpapiere/Chintex/ResearchResults/Downloads/WorkingPaper20.pdf?__blob=publicationFile

Multinational Time Use Study (MTUS). (2014). Retrieved November 05, 2014, from http://www.timeuse.org/

System of Environmental-Economic Accounting (SEEA). (2014) Retrieved November 05, 2014 from http://unstats.un.org/unsd/envaccounting/seea.asp

United Nations Economic and Social Council. (2005) *Environmental-economic accounting* (E/CN.3/2005/15). Retrieved November 05, 2014, from http://unstats.un.org/unsd/statcom/doc05/2005-15e.pdf

Further Reading

Bauer, G., Jungblut, J., Muller, W., Pollak, R., Weiss, F., & Wirth, H. (2006). *Issues in the comparative measurement of the supervisory function.* Unpublished manuscript. Retrieved May 23, 2008, from http://www.mzes.uni-mannheim.de/publications/papers/Supervisor_Function.pdf

Behr, D, J. D., Schneider, S., & Wolf, C. (2016). Harmonizing questions and data in comparative surveys. In C. Wolf, D. Joye, T. W. Smith, & Y. Fu (Eds.), *The SAGE handbook of survey methodology*. Thousand Oaks, CA: Sage.

Bilgen, I., & Scholz, E. (2007). *Cross-national harmonisation of socio-demographic variables in the International Social Survey Programme (ISSP)*. Anaheim, CA: American Association of Public Opinion Research.

Burkhauser, R. V., & Lillard, D. R. (2005). The contribution and potential of data harmonization for cross-national comparative research. *Journal of Comparative Policy Analysis, 7*(4), 313-330.

Carlson, R. O. (1958). To talk with kings. *Public Opinion Quarterly, 22*(3), 224-229.

Desrosieres, A. (2000). Measurement and its uses: Harmonization and quality in social statistics. *International Statistical Review/Revue Internationale de Statistique, 68*(2), 173-187.

Ehling, M. (2003). Harmonising data in official statistics. In J. H. P. Hoffmeyer-Zlotnik & C. Wolf (Eds.), *Advances in cross-national comparison: A European working book for demographic and socio-economic variables* (pp. 17-31). New York, NY: Kluwer Academic / Plenum Publishers.

Esteve, A., & Sobek, M. (2003). Challenges and methods of international census harmonization. *Historical Methods, 36*(2), 66-79.

Gandek, B., Alacoque, J., V, U., Andrew-Hobbs, M., & Davis, K. (2003). Translating the short-form headache impact test (HIT-6) in 27 countries: Methodological and conceptual issues. *Quality of Life Research, 12*, 975-979.

Gil Alonso, F. (2006). *Toward a European statistics system: Sources of harmonized data for population and households in Europe.* Paper presented at the International Data Session of the EAPS European Population Conference.

Hantrais, L., & Mangen, S. (Eds). (1996). *Cross-national research methods in the social sciences*. New York, NY: Pinter.

Harkness, J. A. (Ed.) (1998). *Cross-cultural survey equivalence*. Mannheim, Germany: ZUMA.

Hoffmeyer-Zlotnik, J. H. P. (2004). *Data harmonization*. Roundtable at a conference for the Network of Economic and Social Infrastructures, Luxembourg.

Jan, W. (2009). Establishing Equivalence. In T. Landman & N. Robinson (Eds.), *The SAGE handbook of comparative politics* (pp. 84-100). London: Sage Publications.

Kallas, J., & Apostolos, L. (2008). A documentation model for comparative research based on harmonization strategies. *IASSIST Quarterly* (Winter-Summer), 12-24. Retrieved May 02, 2016, from http://www.iassistdata.org/sites/default/files/iq/iqvol321_4kallas.pdf

Kennett, P. A. (2001). *Comparative social policy: Theory and research*. Buckingham: Open University Press.

Korner, T., & Meyer, I. (2005). Harmonising socio-demographic information in household surveys of official statistics: Experiences from the Federal Statistical Office Germany. In J. H. P. Hoffmeyer-Zlotnik, & J. A. Harkness (Eds.), *Methodological aspects in cross-national research.* (ZUMA-Nachrichten Spezial 11) (pp. 149-162). Mannheim, Germany: ZUMA.

Niero, M., Martin, M., Finger, T., Lucas, R., Mear, I., Wild, D., et al. (2002). A new approach to multicultural item generation in the development of two obesity-specific measures: The obesity and weight loss quality of life (OWLQOL) questionnaire and the weight related symptom measure (WRSM). *Clinical Therapeutics, 24*(4), 690-700.

Olenski, J. (2003). *SSDIS: Global standard for harmonization of social statistics.* Unpublished manuscript. Retrieved May 23, 2008, from http://unstats.un.org/UNSD/demographic/meetings/egm/Socialstat_0503/docs/no_10.pdf

Przeworski, A., & Teune, H. (1966). Equivalence in cross-national research. *Public Opinion Quarterly, 30*(4), 551-568.

U.K. Office for National Statistics (ONS). (2016) *Primary set of harmonised concepts and question*. Retrieved May 23, 2016, from, http://webarchive.nationalarchives.gov.uk/20160105160709/http://www.ons.gov.uk/ons/guide-method/harmonisation/primary-set-of-harmonised-concepts-and-questions/index.html

Data Processing and Statistical Adjustment

Rachel A. Orlowski, Frost Hubbard, Emily Blasczyk, and Dan Zahs, 2016

Introduction

The following guidelines detail ways in which the data collected within each country or culture in multinational, multicultural, or multiregional surveys, which we refer to as a "3MC" surveys, must be processed (i.e., coded, captured, and edited). Although these processing steps tend to be sequential, they may also have an iterative flow. Regarding the survey lifecycle more generally, data processing does not have to wait until all the data have been collected; some of these processing steps can, and possibly should, be taken prior to or concurrent with data collection. The flow involved in processing the survey data may also differ between paper-and-pencil (PAPI) and computer-assisted (CAPI) questionnaires. In computer-assisted surveys, capturing the data, performing edit checks, and building data files should, at least partially, occur automatically while the data are being collected. In doing so, some effort may be eliminated. The data processing effort should be considered when determining the mode of data collection, as well as the costs associated with that decision. See Study Design and Organizational Structure, Instrument Technical Design, Data Collection: General Considerations, Data Collection: Face-to-Face Surveys, Data Collection: Telephone Surveys, and Data Collection: Self-Administered Surveys for more details.

After processing, the data from each country can be harmonized with those from other countries (see Data Harmonization). The calculation of outcome rates and statistical adjustments (i.e., missing value imputation, survey weight creation, and variance estimation) can be performed, as described in these guidelines. Finally, the data should be disseminated as an integrated, cross-cultural dataset (see Data Dissemination). Substantive analyses can be performed on the disseminated dataset (See Statistical Analysis).

Processing and adjustment activities often are not given adequate attention. This is unfortunate because costly errors can still occur after the data have been collected. Just as interviewers may introduce measurement error, data processing operators (e.g., coders, keyers) may potentially introduce processing error, sometimes systematically (Biemer & Lyberg, (2003). Often only a few errors are responsible for the majority of changes in the estimates (Statistics Canada, 2009). To lessen effort and minimize error, checks should be performed throughout the field period, while the respondent is still available, rather than waiting until the end of data collection. The burden of programming and checking should not be underestimated (Statistics Canada, 2009).

These guidelines are broken down into Data Processing Steps (Guidelines 1 through 3) and Statistical Adjustment Steps (Guidelines 4 thru 7). The Quality and Documentation guidelines (Guidelines 8 and 9) are applicable to both. Please note that this chapter assumes the reader has a basic understanding of statistics and has experience in survey data management and analysis. Please refer to Further Reading or an introductory statistics textbook if a statistics refresher is needed.

Guidelines

Data Processing

Goal: To code and capture data from their raw state to an edited data file that can be (1) used within the survey organization for quality assessment of the survey implementation and (2) harmonized with other countries' data files in preparation for statistical adjustment, dissemination, and eventually substantive research.

1. **Use coding to classify survey responses into categories with associated numeric values.**

   Rationale

   To statistically analyze survey responses, they must be transformed into numeric form; this is done by coding. Coding is both a summarization and translation process (Groves et al., 2009a). All responses to a particular survey item need to be summarized into a discrete number of categories. When the survey item is closed-ended (such as the response options in a "Strongly Agree—Agree—Neither Agree nor Disagree—Disagree—Strongly Disagree" scale), the number of categories is explicitly defined—five categories for a five-point scale. Any close-ended questions ideally will have precoded response options—that is their numeric codes will have been defined prior to the start of data collection. (Following coding, further transformation may occur to the coded data, such as reordering scales or collapsing categories with low cell counts for dissemination in order to protect respondent confidentiality.) When the survey item is open-ended, the number of categories is not obvious and should be determined via coding to the analytic purpose of that survey item. Coding is a translation process because the responses must be mapped to categories and nonnumeric category descriptions must be mapped to numeric values. It is possible to analyze non-numeric categorical data, but numeric codes are preferable because most statistical software is designed for numeric values.

   Many code structures, also known as code frames, are defined during questionnaire and instrument development, (see Instrument Technical Design); upon collecting the data, they are revisited and possibly revised. However, codes cannot be fully defined before data collection for some items, and, for example, some open-ended questions may be entirely coded after data collection, or may have their code structures revised during data collection to account for answers that do not fit into the existing code frame. Data quality, in these situations, depends partly upon the interviewer recording all of the information provided by the respondent and partly upon the coder's ability to distinguish among coding categories and to assign the appropriate numeric value.

It should be noted that studies may have data sources other than questionnaires which require coding (Groves et al., 2009a). Such sources could include visual images/recordings, audio recordings, and samples of physical materials and biomeasures (e.g., dust, soil, salvia, blood).

Procedural steps

The creation of code frames for open-ended questions in some areas follows the same principles as the creation of close-ended questions. However, there are some important differences between the two processes. Below the guidelines are divided into 1) items that apply to both open and closed-ended questions, 2) to open-ended only, and 3) to closed-ended only. It is important to note that there are forms of questions that fall between closed and open-ended questions (e.g., numerical open-ended questions, such as "how many times did you do X", or a question that has closed ended responses with an "other-specify" option).

For both closed and open-ended questions:

1.1 Whenever possible and appropriate, take advantage of established coding schemes (Elias, 1997). This is true for both open-ended and closed-ended questions, though the development of open-ended code frames is often further refined and adapted for the particular research questions of the study. See Guideline 1.8 for more details.

1.2 Design each code frame to have the following attributes (Groves et al., 2009a):
- 1.2.1 Unique numeric values and text labels. No number or text label should be used twice.
- 1.2.2 A code number for each possible response category (remember to include code numbers for item-missing data—e.g., "Don't Know," "Refused," and "Not Applicable").
- 1.2.3 Mutually exclusive response categories for each variable (e.g., "Full-time", "Part-time", "Self-employed" are not mutually exclusive).
- 1.2.4 The appropriate number of categories to meet the analytic purpose (see Questionnaire Design).

1.2.5 When using hierarchical code structures, have the first character represent the main coding category with subsequent characters representing subcategories (Biemer & Lyberg, 2003). For example, the International Standard Classification of Occupations (ISCO) code is structured as 4 digits, with left to right as Major group, Sub-major group, Minor group, and Unit group. The occupation "data entry clerk" is 4132. Major group = Clerical support workers (4), Sub-major Group = General and keyboard clerks (41), Minor group = Keyboard operators (413), Unit group = Data entry clerk (4132) (ISCO, 2007).

1.3 Determine which variables should have codes that are standardized across countries and which could have country-specific codes. This decision needs to be communicated between the coordinating center and survey organizations. Decide how these codes will be reconciled once the data are harmonized. See also Data Harmonization.

1.4 Document codes in a data dictionary. There should be a data dictionary entry for each survey item (see Instrument Technical Design for examples of a data dictionary entry). Each entry should contain the following information:

1.4.1 Variable ID, name, and label.
1.4.2 Data format.
1.4.3 Response options and associated code numbers.
1.4.4 Universe statements.
1.4.5 Interviewer and respondent instructions.

1.5 Building upon the data dictionary, develop a codebook which summarizes how the survey responses are associated with all of the data. The codebook includes metadata on the survey items, such as the question text and raw frequency of responses. This document can be used to facilitate quality control (Biemer & Lyberg, 2003).

1.6 Test the instrument prior to data collection/data entry to catch any missing or improperly specified data. Test the instrument at data entry as well as reviewing the data produced. Sometimes a data entry application will accept a value, but the data are not stored properly. Look for:

1.6.1 Missing categories.
1.6.2 Incorrect value limits (e.g. variable on weight in pounds only accepts values 1000 or below).
1.6.3 Improperly specified data structure such as
- Character vs. numeric field consistency
- Field size (e.g. name field only holds 15 characters and names collected are longer than 15 characters)

1.6.4 Entirely null variables, indicating instrument logic is omitting the question.

For open-ended questions:

The following example from the telephone study Survey of Consumers (SCA) 2012 will be used to illustrate concepts pertaining to open-ended coding:

A2. We are interested in how people are getting along financially these days. Would you say that you (and your family living there) are better off or worse off financially than you were a year ago?
1. BETTER NOW
3. SAME
5. WORSE NOW
8. DK
9. NA
A2a. Why do you say so? (Are there any other reasons?)

The open-ended responses to A2a were coded into numeric categories representing reasons the respondent felt better or worse off. See Appendix A for full code frame of this example question.

1.7 There are standard code frames that are used internationally to create comparable data. These should be used by a survey where relevant. For example, for occupation coding there is the International Standard Classification of Occupations (ISCO).

1.8 The creation of new code frames for open-ended questions is a challenging and important part of data processing. It is said that "coding is analysis" (Saldaña, 2012). The concepts and analytic items for coding open-ended data are established by previous research, defined by the research goal, and discovered by coding the data. While code frames from previous studies may be used as a base, it is important to approach coding text without bias.
1.8.1 It is common to use pretest data to establish a code frame for the rest of the study. However, it is rare that a pretest will have enough responses to develop a fully robust frame. Often further modifications may be necessary.
1.8.2 Some studies have a separate, later release for open-ended coded data to allow for the extra time needed for processing.

1.9 Open-ended responses are converted to quantitative data by assessing the presence and/or frequency of words and phrases. (Busch et al., 2012) These words or phrases are selected because they represent concepts that are of interest to the researcher.

1.9.1 For example, the open-ended response to the example A2A *"I'm better off because I got a raise this year."* would be coded to "10. Better pay" (Survey of Consumers, 2012).

1.9.2 It is important to consider the context of the entire response as there are ways context can affect how to code a response.

- In the example A2A, the response "higher interest rates" is a code for both the "better off" (code number 18) and "worse off" (code number 55) reasons. See Appendix A for the full code frame (Survey of Consumers, 2012). In some contexts, higher interest rates would benefit a respondent, such as for investments, but in another context, higher interest rates might mean that the respondent will owe more on his or her loans. The entire response must be read to understand if the respondent sees higher interest rates as a benefit or detriment.
- A respondent doesn't have an answer prepared in advance. They are thinking through their answer as they respond and may discount or revise previous statements as they answer. In the above example the respondent may have an answer to A2A such as *"Well, gas prices have gone down and that has helped with the cost of driving to work, but on the other hand my landlord raised the rent and my wife's hours got cut at her job so overall we're worse off."* In this example, the respondent has discarded their "better off" reason and decided they are "worse off." This is less prevalent in a written open-ended response, but it can still occur there as well.

1.9.3 Multiple words or phrases may be coded under the same code. In the example, the SCA would code responses mentioning *"raise in wages or salary on present job, promotions, higher commissions, change to higher paying job (include Armed Forces induction or discharge) (Any family member who gets a raise is coded 10); increased tips, bonuses"* to "10. Better Pay" (Survey of Consumers, 2012).

1.9.4 At the same time, one open-ended response may have multiple codes assigned to it. For example, the A2A response *"My wife started working when our child started kindergarten. Also, my grandmother passed away in May and I received some money as inheritance which helped us."* could be assigned both codes "12. More work, hence more income" and "13. Increased contributions from outside FU." If coding a response for multiple items, the data may be structured similar to how a closed-ended "select all that apply" question would be. See guideline 1.16.1 for more information on how data of this type are often structured.

1.9.5 Different disciplines may create different, but equally valid code frames. (Saldaña, 2012) For example, in the text *"There's just no place in this country for illegal immigrants. Round them up and send those criminals back to where they came from,"* a researcher interested in public policy may create the code "immigration issues" for this response, while another researcher interested in racial issues might create the code "xenophobia."

1.10 A good code frame starts with a good survey question. A poor survey question will result in responses that are unclear, confusing or off-topic. When writing an open-ended question, it is important to consider:

1.10.1 Are you asking a question the respondent will understand and know the answer to?

1.10.2 Does the question need to be open-ended? If the purpose of the question is to capture specific categories of interest, then an open-ended format may not be necessary.

- For example, one study may be interested in tracking major purchases and would ask about each item separately, "1. Do you own a boat, yes or no?", "2. Do you own a second home, yes or no?" etc. Another study, researching people's plans for a major purchase may want to have it open-ended in order to capture items the researchers hadn't considered. In the first example, the researchers are interested in learning how many people own boats and second homes and in the second example the researchers are interested in learning what items people want, which may be a boat or a second home.

1.10.3 See Questionnaire Design for more details on writing open-ended questions.

1.11 Ultimately, each of the coded items should themselves represent overall concepts that are of research interest. For example, a study (as cited in Saldaña (2012)) on British Muslim girls conducted by Basit in 2003, coded interview data into 23 major categories that clustered into 6 themes. One major theme was "identity," its sub-categories were "ethnicity", "language", and "religion." The relationship between these concepts can also be analyzed through relational analysis (Busch et al., 2012).

1.12 The process of creating the code frame should be iterative. Every time a response is coded, it should be compared with all those responses that have already been assigned that code. This ensures consistent coding and allows for refinement of the codes. This is known as "constant comparison" (Taylor & Gibbs, 2010).

1.12.1 This entire process should itself be repeated to refine and improve the code frame. In the second (or third, or more) cycle, categories may be dropped, combined, or relabeled (Saldaña, 2012).

1.13 For interviewer administered surveys, once a code frame is established, decide if the responses will be field coded by the interviewer or by a trained coder after the case is complete.
  1.13.1 These techniques can be combined: answers can be field coded and later verified by a trained coder. This can cut down on the cost of having an entirely separate and additional coding process.
  1.13.2 If the coding is complex or has many categories it is best to use a trained coder who can take the time to properly code the responses. It is important that field coding not interrupt the "flow" of the interview.

1.14 Consider providing users with both coded data and the raw (but de-identified) open-ended responses so they may conduct their own content analysis.

For close-ended questions:

1.15 Use consistent codes across survey items (Groves et al., 2009a). For example:
  1.15.1 A "Strongly Agree—Agree—Neither Agree nor Disagree—Disagree—Strongly Disagree" scale would always have the values ranging from "1 = Strongly Agree" to "5 = Strongly Disagree".
  1.15.2 A "Yes—No" item would always have the values 1 = Yes and 5 = No (see Instrument Technical Design for an explanation of this coding convention).
  1.15.3 Item-missing data from refusal would always have the values of 9 (or if two-digit code numbers, the values of 99—etc.).

1.16 Be aware how data structure varies across survey software.
  1.16.1 "Select all that apply" questions can come in variety of formats. Some software produces a variable for each category and data contains a binary 'yes/no', indicating whether or not the item was selected; while other software produces a variable for the total number of responses, with the first variable containing the value of the first item mentioned, the second variable containing the value of the seconditem mentioned, and so on. For example:

Question:
Which of the following items do you own? Select all that apply.
1. Laptop
2. Cell phone
3. Tablet

Each category has a variable. Data indicates 1=Selected, 0=Not selected.

ID	CATEGORY_1 (laptop)	CATEGORY_2 (cell phone)	CATEGORY_3 (tablet)
1000	0	1	0
2000	1	1	0
3000	1	1	1

Each selection has a variable. Data indicates what survey item was selected first, second, third.

ID	SELECTION_1	SELECTION_2	SELECTION_3
1000	2=Cell phone		
2000	1=Laptop	2=Cell phone	
3000	1=Laptop	3=Tablet	2=Cell phone

1.16.2 Repeating question groups, used for asking a block of questions that repeat for distinct events/items also have a variety of formats. Some software produces a wide file with repeating columns for each group, while others produce a row for each event/item. For example:

Questions:
A1. Could you estimate the date of your [most/next most] recent hospitalization?
A2. What was the most immediate reason that led to your visit on [DATE]?
1. Chest pain
2. Shortness of breath/difficulty breathing
3. Physical injury (sprain, break, bleeding)
4. Other

Data structure is wide, repeating columns for each group:

id	numvisits	date_1	reason_1	date_2	reason_2	date_3	reason_3
1000	2	3/15/2015	1	12/3/2015	2		
2000	1	5/17/2015	3				
3000	3	6/21/2015	2	8/13/2015	2	11/7/2015	2

Data structure is long, repeating rows for each event/item:

id	visitnum	date	reason
1000	1	3/15/2015	1

1000	2	12/3/2015	2
2000	1	5/17/2015	3
3000	1	6/21/2015	2
3000	2	8/13/2015	2
3000	3	11/7/2015	2

1.16.3 Data may need to be transformed to meet the analytic purpose.

Lessons Learned

1.1 Data are often recoded and transformed in post-processing. It is important to budget this time and expense into the study.

2. Decide how coding and data capture will be conducted and monitored.

Rationale

The methods used to create coded data will vary depending on several factors. One of the major factors that determine coding is the mode of data collection. All surveys require coding to classify responses. However, a paper instrument requires a separate process (data capture) to convert the physical survey into a digital data file, whereas a computerized instrument may only need open-ended responses coded.

When using a paper-and pencil-questionnaire (PAPI), it is important to capture all data provided, even when skip patterns are not followed correctly. Develop a protocol to handle errors when editing the data (see Guideline 3 below).

It is also important to capture information other than the survey data, such as the information from the coversheet, for each sample element, household observations, and interview details (e.g., date, time, and length of the interview). These data will aid in monitoring, evaluating, and potentially improving the data collection process. There are alternatives to manual keying, such as optical character recognition commonly known as “scanning” (ICR), mark character recognition (MCR), voice recognition entry (VRE), and touchtone data entry (TDE).

The resources available will often dictate how data capture will be conducted. The data from all countries may be keyed at a single location (typically the coordinating center), or it may be conducted by each country individually and combined afterward (Biemer & Lyberg, 2003).

The decisions for how coding will be monitored are also affected by these factors. Some method of monitoring is important to ensure data quality. Even computerized questionnaires require monitoring for errors.

Procedural steps

2.1 Determine how data capture will occur. This may vary across countries depending upon their respective amount of funding, resource availability, infrastructure constraints, and cultural feasibility. When country-specific adaptations are necessary, it is important to establish a data capture monitoring system that ensures comparability across countries.

2.2 Design the coding harmonization strategies needed for the data to achieve comparability across countries. For more information, see Data Harmonization.

2.3 Design the data entry software to maintain the question order and measurement units of the paper survey. In the case of mixed mode studies, it may also be necessary to reconcile differences between the data captured via the two modes. The primary goal should be to make data entry and simple and logical process, but consistency between the two modes is also important.

- 2.3.1 For paper surveys, decided whether or not to program the software to allow the keyer to ignore errors made in filling out the form (e.g. when the skip pattern was not correctly followed). The decision depends on whether or not it is of interest to capture these errors.
- 2.3.2 Consider distributing a data entry shell to all study site countries using PAPI in a 3MC survey to facilitate data harmonization.

2.4 Decide if coding is to be centralized or decentralized. Depending on resource availability, as well as the data being collected, consider centralized coding versus decentralized coding. Centralized coding occurs at one location, typically the coordinating organization. Decentralized coding applies to situations where each individual country conducts its own coding prior to data being combined, as well as situations where coders from one organization work in multiple locations, such as their own homes. Keep in mind that:

- 2.4.1 Supervisory control is easier with centralized coding. This often results in higher inter-coder reliability (see Appendix B).
- 2.4.2 Centralized coding typically involves fewer coders, with each coder having a larger workload. The larger workload can result in a higher coder design effect (see Appendix C). Training is key to reducing this effect.

2.4.3 Decentralized coding often occurs when administrative data, such as hospital records, are collected and combined into a single data source. Different hospitals and clinics may have variation in their coding procedures. It is important to consider the caliber of the various sources of data, and it should be recognized that some recoding of such data may be required (Jordan, Porcheret, & Croft, 2004).

2.5 Properly train coders on the study's coding design, and periodically assess their abilities. This ensures that coders have equivalent coding abilities and that coding is consistent, which reduces coder design effect.

2.6 Endeavor to control manual coding by using independent verification instead of dependent verification. (Biemer & Lyberg, 2003)

2.6.1 In independent verification two coders code all responses separately. Discrepancies are handled with a computer or an adjudicator (Biemer & Lyberg, 2003).

2.6.2 Independent verification is more costly than dependent verification, but is more reliable.

2.6.3 Independent verification reduces the likelihood of under-detection of errors (O'Regan, Lynch, Odell, 1988).

2.6.4 Independent verification also reduces coding errors
- The likelihood of two or three coders independently assigning the same erroneous code is small.
- Independent verification is not foolproof, especially if the coders are not properly trained or monitored.

2.6.5 In dependent verification, the first coder codes responses and a second coder verifies the responses and makes changes to any code they deem erroneous, meaning the verifier has access to the initial outcome and revises any detected errors.

2.6.6 A survey can use both independent and dependent verification to offset cost. Consider using independent verification for key items that are difficult to code (such as occupation coding) and dependent verification for other items that are more straightforward, such as a "strongly-agree" to "strongly disagree" scale.

2.6.7 Strive to verify 100% of the data entry (see Federal Committee on Statistical Methodology (1983) and Groves et al. (2009a)).

2.6.8 Look for the following keyer errors (Wurdeman, 1993):
- Wrong column/field.
- Corrected/modified (misspelled) responses.
- Be especially cautious about correctly coding the first character of hierarchical code structures because errors at the higher levels are usually more serious.

- For example, the International Standard Classification of Occupations (ISCO) code is structured as 4 digits, with left to right as Major group, Sub-major group, Minor group, and Unit group. The occupation "data entry clerk" is 4132, whereas 5132 is the occupation code for "bartenders." (ISCO, 2007)

2.7 Consider automated alternatives to key entry, including (Biemer & Lyberg, 2003):
   2.7.1 Optical character recognition (OCR) to read machine-generated characters.
   2.7.2 Intelligent character recognition (ICR), commonly known as scanning, to interpret handwriting.
   2.7.3 Mark character recognition (MCR) to detect markings (i.e., bubbles).
   2.7.4 Voice recognition entry (VRE) to automatically transcribe oral responses.
   2.7.5 Touchtone data entry (TDE) to interpret numbers pressed on a telephone keypad.

2.8 When using automated coding systems:
   2.8.1 Decide between using exact matching, which results in less error but also fewer assignments, or inexact matching, which has the opposite outcome.
   2.8.2 Check for any responses that are left uncoded, and manually code them.
   2.8.3 Frequently recalibrate and configure scanning equipment to minimize the frequency of with which the software misreads information (e.g., with OCR).
   2.8.4 Store the code structure as a dictionary database with alternative descriptions, so a realistic response pattern can be handled.

2.9 Evaluate the coding process.
   2.9.1 For manual keying: Collect and monitor paradata on coding and verification, such as error rates, at the variable, code number, and coder level.
   2.9.2 For automated coding: Collect paradata on the scanning operation, such as rejects and substitutes, by character and by machine.
   2.9.3 Assess the reliability of coding.
   - A common way to calculate reliability of a code is to compute the inter-coder reliability, or Cohen's kappa (i.e., a statistical measure that accounts for chance). Kappa is most informative when there are a small number of coding categories (see Appendix B for the formula for kappa).

- If the reliability is less than what is specified as acceptable, provide additional coder training and consider revising the coding frame.
- Consider revising the code if the original code is not reliable.

2.10 Flag any concerns from keyers or errors from the automated system for expert review at a later time, during data editing (see Guideline 3 below). Errors should not hinder the performance of the keyers or halt automated coding (United Nations, 2005).

Lessons learned

2.1 Although using a comprehensive data dictionary for automated coding generally results in less manual coding, expanding the dictionary does not always mean more accuracy (Biemer & Lyberg, 2003). Additions to a data dictionary or coding reference file can lessen the automated coding software's ability to exactly match and assign code numbers to the responses, resulting in more manual coding. The Canadian Census of Population and Housing in 1991 updated their reference file not only to add items, but also to remove phrases that were generating errors (Tourigny & Moloney, 1991).

2.2 With automatic coding, consider the effort made in revising the codes in relation to the automation gained. The data dictionary for one of the Swedish household expenditure surveys was updated 17 times, increasing in size from 1459 to 4230 descriptions. The third update (containing 1760 descriptions) allowed 67% of the data to be automatically coded while later versions of the data dictionary could only code up to 73% of the responses—a gain of only 6% after 14 additional updates.

2.3 Those with prior experience coding survey data may not always be the best people to code data on a particular survey. Substantive knowledge may also be necessary when selecting coders, depending on the complexity of the survey items. For example, the World Mental Health Survey employs coders who are psychologists or psychiatrists in order to diagnose verbatim responses.

2.4 Coding errors are not trivial; they can systematically alter results, and damage accuracy of estimates.

2.5 A computerized instrument does not prevent data errors. For example, if the instrument has incorrect skip logic or has improper specification to columns, data will be lost or truncated.

2.6 Many established 3MC surveys are partly or wholly paper-and-pencil based, making data entry necessary. While studies vary somewhat in the details, typically each participating country is responsible for entering and cleaning its own data, a supervisor or data manager checks questionnaires before data entry occurs, and some percentage of questionnaires are double-entered. Whatever protocol is used, it is important to fully document the data entry process. The following are examples of data entry strategies for studies that were partially or entirely paper and pencil:

2.6.1 Round 6 of the Afrobarometer Survey used a paper-and-pencil instrument. Each participating country was responsible for entering, checking and cleaning its own data. The project utilizes a data-entry template which outline the variable names and data types required. However, each country may have its own questions or codes. The data were also reviewed by the core partner data managers and the Afrobarometer data manager. Data cross-checks were performed on a regular basis. Either rolling data entry or batch data entry was employed at the discretion of the data manager. A minimum of 25% of all questionnaires was double-entered (Afrobarometer Survey, 2014).

2.6.2 In the Asian Barometer, another pencil-and-paper survey, quality checks are implemented at every stage and data cleaning involves checks for illegal and logically inconsistent values. A minimum of twenty percent of the data are entered twice by independent teams (see http://www.asianbarometer.org/survey/survey-methods).

2.6.3 Round 5 of the European Social Survey (ESS) was administered as either a pencil-and-paper or a computer assisted survey, depending upon the country's resources. National coordinators were responsible for entering and cleaning their own data and documenting their cleaning procedures before submitting the data to the ESS Archive. Files were further scrutinized for content and consistency once uploaded to the ESS Archive (ESS, 2015).

2.6.4 The Living Standard Measurement Study Survey (LSMS) is also pencil-and-paper and each participating country is responsible for its own data editing and cleaning. Data entry operators enter the data into a specially designed program after each of the two rounds of the LSMS. Each country uses computers with specially designed software to check for accuracy, consistency, and missing data. Further data cleaning is performed by the data manager (LSMS, 1996).

2.6.5 The World Mental Health Survey can be administered as either a pencil-and-paper or a computer assisted survey, depending upon the country's resources. Data from pencil-and-paper versions of the interview are entered manually with a data-entry program designed by the WMH Data Collection Coordination Centre. Computer assisted versions, by nature, are automated. Guidelines require all completed pencil-and-paper interviews to be edited for legibility, missing data, and reporting standards by specially trained editors. In the majority of participating countries, follow-ups are done on questionnaires with errors. Independent double entry is recommended, but keying-acceptance sampling (ranging from 10% to 20%) is allowed and used by the majority of the participating countries to evaluate keying errors. Standard coding schemes and procedures are given to all participating countries. Ten percent double coding is required. Clean datasets, checked for common errors, such as, blank or missing variables, out-of-range responses, and consistency checks, are required from all participating countries (Kessler, Üstün, & World Health Organization, 2008).

2.7 Data entry software ranges from simple spreadsheets to sophisticated applications with built-in edit checks. When possible, a standardized set of tools should be used across countries to meet quality standards. Consider the use of publically available software if cost is a concern. For instance, the US Census Bureau has a data entry application, Census and Survey Processing System (CSPro, 2010) that is available without cost. CSPro is a software package for entering, editing, tabulating, and disseminating census or any survey data. CSPro was the recommended data entry program for the Afrobarometer Round 6.

2.8 Sophisticated data entry software will help the staff keying the data (for example, by accounting for skip patterns in the questionnaire). Having this level of sophistication will likely reduce entry errors but will likely cost substantially more to program and to test properly.

2.9 Often, the same individual(s) creates many of the entry errors (often on the same variables). By limiting the number of individuals who perform data entry, it is easier to isolate potential problems and to offer appropriate follow-up training.

3. Edit the data to check for errors throughout the survey lifecycle.

Rationale

Cleaning the data (e.g., correcting errors) is the primary purpose for editing, but editing can also provide information about data quality (e.g., exposing where interviewers or respondents may have difficulty performing their roles) and about improvements to future surveys (e.g., revealing where a particular design decision may be an error source) (Biemer & Lyberg, 2003).

Editing can be defined as two phases: 1) identification, followed by 2) correction. Editing can occur at various points in the survey lifecycle (Biemer & Lyberg, 2003). Incorporating editing procedures prior to and during data collection is a better allocation of resources than only after data collection. For example, in computer-assisted surveys, the application can notify the interviewers (or respondents, if self-administered) of inconsistent or implausible responses. This gives interviewers/respondents a chance to review, clarify, or correct their answers. Prior to data capture, survey organizations can manually look for obvious errors, such as skipped questions or extraneous marks on a form. Then, during data capture, editing software can be used to check for errors at both the variable and case level.

Procedural steps

3.1 Program computer-assisted applications to aid in the editing process during both data collection and data processing tasks. For example, in a computer-assisted personal interview (CAPI) instrument, an age value of 233 would prompt the interviewer to confirm the value and then reenter it as perhaps 23 or 33. It may also be coded to missing if a reasonable estimate cannot be made. See Instrument Technical Design for further discussion of instrument programming.
   3.1.1 Limit programming computer-assisted data capture applications to the most important edits so as not to increase the length of the survey or to disrupt the interview/data entry (Groves et al., 2009a).
   3.1.2 Decide if the edit check is a soft check or hard check. A soft check asks for the value to be confirmed but lets the survey progress with the original value. A hard check does not allow the survey to progress until an acceptable value is entered. A survey will often have both soft and hard checks. Limit the number of hard checks to only crucial items.
   3.1.3 If the interviewer/keyer chooses to retain the original value after the edit check, program the application to allow for a comment to be written about that decision. These comments can prevent erroneous editing.

3.2 Create editing decision rules both during and after data collection (see Biemer & Lyberg, 2003; Groves et al., 2009a; Office of Management and Budget, 2006; Statistics Canada, 2009, and United Nations, 2005). Rules can include:
   3.2.1 Developing systematic protocol to resolve:

- Wild values (e.g., out-of-range responses, unspecified response categories, etc.)
- Implausible outliers (e.g., extremely high or low values)
- Imbalance values (e.g., subcategories that do not sum to the aggregate)
- Inconsistent values (e.g., parents' ages that are not reasonably higher than their children's, males that report pregnancies, etc.).
- Entirely blank variables

3.2.2 For paper-and-pencil instruments in particular, deciding how to resolve (Wurdeman, 1993):
- Single-response variables with many response values.
- Illegible responses.
- Erasures.
- Markings outside the response check box.
- Crossed out (but still legible) responses.
- Added response categories (e.g., "None," "Not Applicable," "Refused," etc.).
- Incorrect skip patterns.

3.2.3 Comparing the current data to data from prior waves (or to data from related respondents), when applicable.
3.2.4 Verifying the correct number of digits for numeric variables.
3.2.5 Setting a minimum number of items filled to be considered a complete interview (including item-missing data on key variables).
3.2.6 Confirming the proper flow of skip patterns.
3.2.7 Flagging omitted or duplicated records.
3.2.8 Ensuring a unique identification number for every sample element, as well as a unique identification number for each interviewer.

3.3 Establish decision rules as to whether the potential errors should be accepted as correct, changed to another value, or flagged for further investigation (Biemer & Lyberg, 2003).
3.3.1 Follow up on the suspicious values only if they could seriously affect the estimates, weighing the costs and logistics of recontacting the respondent (Statistics Canada, 2009).

3.4 Editing software may not be efficient in small surveys, but it is critical in large surveys (Biemer & Lyberg, 2003).

3.5 Create a flag that indicates a change has been made to the collected data, and keep an unedited dataset in addition to the corrected dataset (Office of Management and Budget, 2006). The latter will help decide whether the editing process adds value. If unedited data are not kept it is truly impossible to establish whether or not improvements have been made.

3.6 Assess a random sample of each interviewer's completed questionnaires by examining the captured data. Review the use of skip patterns and the frequency of item-missing data to see if any interviewers need additional training on navigating the instrument or probing for complete answers.

3.7 Consider using logical imputation, when appropriate.

3.7.1 Logical imputation is the process of eliminating item-missing data by reviewing data the respondent provided in prior waves or in other items within the same questionnaire and then adding the logical value.

3.7.2 For example, if a series of questions regarding the number of drinks of beer, wine, and hard alcohol consumed in the past week all have values but the final question in the series regarding the sum of drinks consumed in the past week is blank, then the total number of drinks can be logically imputed by adding the values from the individual beer, wine, and hard alcohol items.

3.7.3 Note that this is not a statistical technique; values are deduced through reasoning. Be aware of the danger of creating systematic error by using such logic.

3.8 Collect paradata on the editing process, so it can gradually improve and become less costly (see examples in Guideline 8) (see also Biemer & Lyberg (2003) and Granquist & Kovar (1997)).

Lessons learned

3.1 Overediting may delay the release of the dataset, reduce its relevance to users, and be extremely expensive (see Biemer & Lyberg (2003) and Granquist & Kovar (1997)). A lot of editing is not cost-effective. Make selective editing decisions based on the importance of the sampling element or variable, the severity of the error, the costs of further investigation, and the effects of changes in the final estimates. Often, the level of detail required for any variable(s) depends strongly on the funding sources and the purpose of the estimates. These considerations should be balanced with the other needs of the study. The time and money saved by implementing selective editing can be redirected to other processing steps or other tasks of the survey lifecycle.

3.2 Editing must be a well-organized process; if it is not, on-going changes to the data may actually reduce their quality (Fellegi & Holt, 1976). Identify fields involved in the most failed edits and repair them first.

Statistical Adjustment

Goal: To improve estimates of target population parameters based on sample survey data.

4. Use disposition codes and calculate outcome rates based on established, cited survey research standards.

Rationale

Response rates are one indication of survey quality and can also be used to adjust survey estimates to help correct for nonresponse bias. Therefore, reporting response rates and other outcome rates based on an established survey research standard is an important part of dissemination and publication (see Data Dissemination for additional discussion). Additionally, outcome rates often serve as indicators of a survey organization's general performance.

Procedural steps

4.1 Have the coordinating center provide a list of specific disposition codes and a clear description of how to code and classify all sample elements during the field period (using temporary disposition codes) and at the end of the field period (using final disposition codes). These disposition codes will allow the standardization of outcome rate calculations across countries.

4.1.1 Generally, disposition codes identify elements as a completed interview or non-interview. Non-interviews are further subdivided depending upon whether the sample element is eligible or ineligible to participate in the study. For surveys where sample elements are people, ineligible non-interviews might include the respondent being deceased, the housing unit being unoccupied, or the respondent having emigrated outside of the boundaries of the study area. Eligible non-interviews include refusal to participate, noncontacts, and others as defined by the study.

4.1.2 Disposition codes are mutually exclusive, and, while each sample element may be assigned a number of different temporary disposition codes across the field period, ultimately it will be assigned *only one* final disposition code.

4.2 Based on an established survey research standard, assign all sample elements into mutually exclusive and exhaustive categories and calculate response rates.

4.2.1 Assigning elements into predetermined final categories makes it possible to recalculate each country's response rate in a standard way for comparison across countries, as appropriate.

4.2.2 The World Association for Public Opinion Research/American Association for Public Opinion Research (WAPOR/AAPOR) provides one example of an established survey research standard (AAPOR, 2016).

- According to WAPOR/AAPOR's "Standard Definitions of Final Dispositions of Case Codes and Outcome Rates for Surveys," there are four main response rate components: Interviews, Non-interviews-Eligible, Non-interviews-Unknown Eligibility, and Non-interviews-Ineligible.
- WAPOR/AAPOR defines six separate response rates (RR1-RR6) (AAPOR, 2016).
  - Response rates ending in odd numbers (i.e., RR1, RR3, and RR5) do not consider partially-completed interviews to be interviews. Response rates ending in even numbers (i.e., RR2, RR4, and RR6) consider partially-completed interviews to be interviews.
  - RR1 and RR2 assume that all sample elements of unknown eligibility are eligible.
  - RR3 and RR4 estimate the percentage of elements of unknown eligibility that are actually eligible.
  - RR5 and RR6 assume that all elements of unknown eligibility are ineligible.
- Appendices D-G in Data Collection: General Considerations contain a description of disposition codes and templates for calculating response rates from the AAPOR.

4.3 Based on an established survey research standard, calculate other important outcome rates such as contact, cooperation, or refusal rates.

4.3.1 There are many different industry standards available. WAPOR/AAPOR's outcome rate calculations are an example of one such standard (AAPOR, 2016). Another has been developed by Statistics Canada (Singh, Hidiroglou, Gambino, & Kovačević, 2001).

Lessons learned

4.1 Ensure that each disposition code is clearly described and reviewed during each participating country's interviewer training. Countries may not be familiar with the specified disposition codes or the response rate terminologies. As another check, consider obtaining contact attempt records from each country early in the data collection period in order to ensure that all countries are correctly identifying different outcomes and understand the difference between temporary and final disposition codes. Implement all disposition codes according to the study requirements.

4.2 Standardize the specific disposition codes as much as possible across all participating countries. However, recognize that some special, country-specific disposition codes may need to be created to adequately describe the situation. For example, since best practice suggests allowing the sample design to differ across countries, different disposition codes regarding ineligible elements may need to be created for certain countries.

5. Develop survey weights for each interviewed element on the sampling frame.

Rationale

Depending upon the quality of the sampling frame, the sample design, and patterns of unit nonresponse, the distribution among groups of observations in a survey dataset may be much different from the distribution in the survey population. To help correct for these differences, sampling statisticians create weights to reduce the sampling bias of the estimates and to compensate for noncoverage and unit nonresponse. An overall survey weight for each interviewed element typically contains three adjustments: 1) a base weight to adjust for unequal probabilities of selection (w_{base}); 2) an adjustment for sample nonresponse (adj_{nr}); and 3) a poststratification adjustment (adj_{ps}) for the difference between the weighted sample distribution and population distribution on variables that are considered to be related to key outcomes. If all three adjustments are needed, the overall weight is the product of these three adjustments:

$$w = w_{base} * adj_{nr} * adj_{ps}$$

However, it is not always necessary to create all three weight adjustments when creating an overall survey weight. Create the adjustments only as needed. For example, if all elements had equal probabilities of selection, a base weight would not be necessary. The overall survey weight would then be the product of any nonresponse adjustment and any poststratification adjustment (Groves et al., 2009a).

Presently, the field of survey research lacks a methodology that can help develop weights for other major survey errors, such as processing and measurement error. At this time, evaluation methods are used instead of development and application of weights.

Procedural steps

5.1 If necessary, calculate the base weight for each element.
5.1.1 Each element's base weight is the inverse of the probability of the selection of the specified element across all stages of selection. If necessary, calculate the nonresponse adjustment for each element.

5.2 There are many ways to calculate nonresponse adjustments. This guideline will only explain one method, which uses observed response rates within selected subgroups. This method is easier to calculate than others but assumes that all members within a specific subgroup have the same propensity of responding. For information on other nonresponse adjustment methods, see Bethlehem (2002), Särndal and Lundström (2005), and Wagner and Stoop (forthcoming).
5.2.1 Compute response rates for mutually exclusive and exhaustive subgroups in the sample that are related to the statistic of interest.
5.2.2 The inverse of a subgroup's response rate is the nonresponse weight for each eligible, sampled element in the subgroup.

5.3 If necessary, calculate the poststratification adjustment.
5.3.1 Multiply $w_{base} * adj_{nr}$ to obtain a weight that adjusts for both unequal selection probabilities and sample nonresponse for each eligible element.
5.3.2 Using this weight, calculate a weighted sample distribution for certain variables related to the statistics of interest where the population distribution is known (e.g., race and sex). See Kalton & Kasprzyk (1986) for a method of computing poststratification weights when the population distribution is unknown for certain subgroups (e.g., using raking or iterative proportional fitting).
5.3.3 In 3MC surveys, make sure that the official statistics used by each participating country to estimate the population distribution have the same level of accuracy. If that is not the case, seek corrections or alternatives.
5.3.4 Divide the known population count or proportion in each poststratum by the weighted sample count or proportion to compute adj_{ps}.
- For example: According to 2007 estimates from Statistics South Africa, women comprised 52.2% of the total population residing in the Eastern Cape Province. Imagine the weighted estimate of the proportion of women in the Eastern Cape from a small local survey after nonresponse adjustments was 54.8%. The poststratification adjustment,

adj_{ps}, for female respondents in the Eastern Cape would be .522/.548 = .953.

5.3.5 Note that missing values for any variable needed for poststratification adjustments should be imputed (see Guideline 6 for information on imputation).

5.4 Multiply the needed weight adjustments together to determine an overall weight for each element on the data file.

5.5 If necessary, trim the weights to reduce sampling variance.

5.5.1 Survey statisticians trim weights by limiting the range of the weights to specified upper and lower bounds (e.g., using no less than the 10th percentile and no more than the 90th percentile of the original weight distribution).

5.5.2 Trimming of weights produces a reduction in sampling variance but might increase the mean square error (Biemer & Christ, 2008).

5.6 If necessary, consider other weight components (besides the base weight, nonresponse adjustment, and poststratification adjustment).

5.6.1 There may be weight components other than the three described in this guideline. Other possible weight components are country-specific adjustments and weights that account for differential probability of selection for certain questionnaire sections.

5.7 Apply the final weight to each record when calculating the statistic of interest.

5.7.1 Weights can be scaled for different analytical purposes. One common technique is to scale the weights so that they sum to the total size of the population.

5.8 Understand the advantages and disadvantages of weighting.

5.8.1 Weighting can reduce coverage bias, nonresponse bias, and sampling bias at the country or study level, depending on whether the weights were designed to reflect the population of a specific country or the entire study.

5.8.2 Caveats:

- Weighting can increase sampling variance. See Appendix D for a rudimentary measure of the increase in sampling variance due to weighting.
- When forming nonresponse adjustment classes, it is assumed that respondents and nonrespondents in the same adjustment class are similar. This is a relatively strong assumption.

- If the accuracy of the official statistics used to create poststratification adjustments differs by country, comparability across countries can be hampered (Gabler & Häder, 1997). In addition, if the poststratification adjustments do not dramatically impact the survey estimates, consider not using the adjustment.

Lessons learned

5.1 Ensure that all participating countries thoroughly document their sampling procedures and selection probabilities at every stage of selection. Countries that do not routinely employ survey weights or use complex survey designs may not be accustomed to recording and maintaining this information. Without this information, it can be very difficult to recreate base weights once data collection is complete.

6. Consider using single or multiple imputation to compensate for item-missing data.

Rationale

Item-missing data are common in social science research data. Imputation is often used to address this problem. The aim of imputation is to reduce the bias in the estimate of the statistic of interest caused by item-missing data and to produce a rectangular dataset without gaps from the missing data that can be analyzed by standard software.

The two main methods of imputation—single and multiple imputation—are described in this guideline (Kalton, 1983b; Marker, Judkins, & Winglee, 2002).

Single Imputation Methods

Rationale

Single imputation involves replacing each missing item with a single value based on the distribution of the non-missing data or using auxiliary data (Vermaak, 2009). It is the easier of the two imputation methods. There are several common methods, which are discussed below.

Procedural steps

Single Imputation Methods:

6.1 Select one of the single imputation methods available. Consider the following:

6.1.1 Overall mean value hot-deck imputation.

- Replace the missing values for a variable with the mean value for that variable across the entire dataset.
- While this is a very simple method to use, it can distort the distribution of the variable with imputed values by creating a spike in the distribution at the mean value, potentially biasing the results.

6.1.2 Overall mean value cold-deck imputation.

- Replace the missing values for a variable with the mean value for that variable from an external source or dataset.

6.1.3 Sequential hot-deck imputation.

- Sort the dataset by specific, observed variables related to the statistic of interest. For example, imagine the statistic of interest is the average, yearly personal income in Spain. Assume that it is known from previous studies that the yearly personal income in Spain is related to years of education and age. The dataset would first be sorted by years of formal education and then respondent age.
- See if the first element on the sorted dataset has a value for the variable that is to be imputed; in the above example it would be reported yearly personal income.
- If the first element does not have a value, impute the mean value of the variable based on the sample elements with data on the statistic of interest.
- If the first element does have a value, keep this reported value and move to the second element. The last reported value is now the “hot-deck” value.
- If the second element is missing a value for the specified variable, impute the “hot-deck” value. The value for the second element then becomes the “hot-deck” value for the third element, etc.
- Sequential hot-deck imputation is less costly than regression imputation methods because no model fitting is necessary, and it has fewer complexities than regression imputation methods. Thus, sequential hot-deck imputation is more easily understood by analysts and can reduce variance and nonresponse bias

6.1.4 Regression imputation.

- Carefully create a regression model for a specific variable that predicts the value of the variable based on other

observed variables in the dataset. For example, one could create a regression model that predicts the number of doctor visits in the past year based on demographics, such as age, sex, race, education, and occupation.
- Check that the predictor variables do not have many missing values.
- Regression imputation can produce better imputations of missing values than hot-deck methods for variables with complex missing data patterns and for small samples.

6.2 For all variables for which at least one value was imputed, create imputation flag fields that indicate which items for each record on the data file were imputed.

Multiple Imputation Methods

Rationale

The goal of multiple imputation is to account for the decreased variance imputed values have compared to observed values. Multiply imputed values and multiple datasets are created for each missing value. Variation in the estimates across the trial runs allows for the estimation of both sampling and imputation variance. Therefore, multiple imputation creates a distribution of imputed values that have their own standard errors and confidence intervals (Vermaak, 2009). An added level of expertise is needed to perform multiple imputation, which may result in a more expensive procedure than using single imputation.

Due to the statistical complexity of multiple imputation methods, only the most commonly used method—sequential regression imputation—is briefly described below (see Little & Rubin (2002) and Rubin (1987) for additional detail). Please refer to Lepkowski & Bowles (1996) for information on other methods.

Procedural steps

6.1 Select a multiple imputation method; consider sequential regression imputation.

6.1.1 Create multiple datasets where each missing element is based on a different trial run of a regression model for each imputed item.

- This is an iterative process where one item is imputed using an imputation model and then the next item is imputed with a regression model that uses the imputed values of the first item.

- Consider using the same set of variables for all imputations to reduce the risk of over-fitting the model.

6.1.2 Several statistical software packages are capable of multiple imputation. Imputation and Variance Estimation Software (IVEWare), a package developed at the University of Michigan and available to users for free, is an example of one such package (IVEWare, 2009). R programs that perform multiple imputation are also available (Eff & Dow, 2009).

6.1.3 Use sequential regression imputation when records contain different numbers of missing items.

6.1.4 Although sequential regression imputation accounts for the increased uncertainty of imputed values, it can be time-consuming for large surveys.

Lessons learned

6.1 Researchers who employ case deletion are frequently forced to collapse regions together in order to have enough cases to analyze. By imputing data, regional distinctions can be maintained (Dow & Eff, 2009).

6.2 Sampling statisticians advise users to avoid imputing attitudinal variables since attitudes can easily change over time and missing data patterns can be difficult, if not impossible, to predict. Imputation models for factual variables are generally easier to specify because they are more static and outside validation can be provided.

6.3 If item nonresponse is missing at random (MAR) given the covariates used in the imputation process, imputation reduces bias, sometimes a lot. In MAR, the process causing missing values can be explained either by the variables in the model or by variables from auxiliary data. (See Appendix E for more information about assumptions for missing data).

6.4 Imputed data are synthetic data. Computed variances using single-imputed data methods will be smaller than the true underlying variances that would have occurred of a same sized sample without any missing data.

6.5 Data analysts must be able to identify real values and imputed values. Therefore, the imputation must be thoroughly documented.

6.6 Imputation procedures can vary across survey topics and populations. Therefore, different procedures may need to be implemented and documented within different countries, etc. For an example, see Frick and Grabka (2007).

6.7 Even with the continual improvements in statistical software, multiple imputation methods may be hard to do for many 3MC surveys because it takes a greater skill level and often more time and money than single imputation. In addition, each variable requires specific treatments and evaluation on how to impute the missing values.

6.8 Check that the imputation model fits the data correctly and is well specified. A poor imputation model can actually increase the bias of the estimate, making it worse than not using imputation.

7. When calculating the sampling variance of a complex survey design, use a statistical software package with the appropriate procedures and commands to account for the complex features of the sample design.

Rationale

The survey sample design determines the level of precision. Unfortunately, many statistical texts only discuss the sampling variance formula for simple random sampling without replacement (a sampling method that is almost never used in practice). Similarly, statistical software packages (e.g., STATA, SAS, and SPSS) assume simple random sampling without replacement unless otherwise specified by the user. However, compared to a simple random sample design, (proportionate) stratification generally decreases sampling variance while clustering increases it (see Sample Design for in-depth explanations of simple random samples, stratification, and clustering). If the correct formulas or appropriate statistical software procedures and commands are not applied, the calculation of the precision (i.e., sampling variance) of the statistic(s) of interest can be inaccurate. Therefore, analysts are cautioned to ensure they are applying the correct methods to calculate sampling variance, based on the sampling design. Always compare results with the default simple random sample selection assumptions to check for inconsistencies that might occur due to defective estimators.

Procedural steps

7.1 In order to use Taylor series variance estimation, which many statistical software packages use as a default, the survey data file must include, at a minimum, a final survey weight, a stratum identifier, and a sampling unit identifier for each responding sample element (Groves et al., 2009a). The chosen statistical software package must have the capacity to account for survey weights, stratification, and sampling units in the estimation process (Lepkowski & Bowles, 1996; Brogan, 2005).

7.1.1 If the complex survey design used clustering, the survey data should also include cluster identifiers for each responding sample element.

7.1.2 In order to estimate the sampling variance within a stratum, at least two selections must be made within the stratum. For a sampling design that selects only one primary sampling unit (PSU) per stratum, the sampling variance cannot be estimated without bias. In "one PSU per stratum" designs, the PSUs are arranged after data collection into a set of sampling error computational units (SECUs) that can be grouped into pairs for purposes of estimating approximate variances. If a participating country uses a sample design that selects only one PSU per stratum, the survey data must include the SECU of each element to make variance estimation possible.

7.2 When a survey data file is supplied with a series of replicate weights plus the final survey weight, balanced repeated replication or jackknife repeated replication could be used to estimate variances (see Appendix F).

7.3 When estimating means and variances with statistical software packages, use the appropriate procedures and commands to account for the complex survey data. For example, SAS version 9.1.3 features the SURVEYFREQ and SURVEYMEANS procedures with strata and cluster commands to account for complex survey designs.

Lessons learned

7.1 Not all countries may have access to statistical software packages or skilled personnel. Therefore, it may be necessary to arrange for reduced fees or for centralized analysis. Alternatively, consider using free, open source software, such as R.

Data Processing and Statistical Adjustment

8. Implement quality checks at each stage of the data processing and statistical adjustment processes.

Rationale

Ensuring quality is vital throughout the survey lifecycle. Even after data collection is complete, the survey organization must continue to implement quality measures to help reduce or eliminate any errors that could arise during the processing and adjustment procedures discussed above. If the emphasis on quality is relaxed during these latter activities, all of the time and money spent on maintaining quality during the previous tasks of the survey lifecycle will be compromised.

Procedural steps

8.1 Continually monitor coding activities, such as the number of responses that were coded automatically; were changed after data dictionary updates (Biemer & Lyberg, 2003).

8.2 Use data entry tools to perform keying quality checks. Have human analysts check for representativeness and outliers (United Nations, 2005).

8.3 Monitor editing using some key process statistics (Biemer & Lyberg, 2003; Granquist & Kovar, 1997). Examples are as follows (where objects can refer to fields, characters, or records):
   8.3.1 *Edit failure rate = # of objects with edit failures / # of objects edited* (estimate of amount of verification).
   8.3.2 *Recontact rate = # of recontacts / # of objects edited* (estimate of number of recontacts).
   8.3.3 *Correction rate = # of objects corrected / # of objects edited* (estimate of the effect of the editing process).

8.4 Remove any identifying information from the production data. For example, remove any names and addresses attached to each responding element or unit. (For more information, see Ethical Considerations and Data Dissemination.)

8.5 When possible, use paradata and other auxiliary data (e.g., census or population files) for post-survey adjustments and to enhance the precision of the survey estimates. For example, population files could be used to create nonresponse weighting adjustment categories. However, in 3MC surveys be aware of very different levels of accuracy across countries for such information.

8.6 Compare the sum of the base weights of the initially sampled elements to the count N of units on the sampling frame. If the sample was selected with probabilities proportional to size, then the sum of base weights is an estimate of N. If an equal probability sample was selected within strata or overall, then the sum of base weights should be exactly equal to N.

8.7 Assign a second sampling statistician to check the post-survey adjustment methodology and the statistical software syntax of the survey's primary sampling statistician. This should be done whether the statistical adjustments are done individually by each participating country or done for all countries by a statistical team selected by the coordinating center.

Lessons learned

8.1 Make certain that all identifying information is removed from the dataset before making it publicly available. In some surveys, this may require detailed geographic identifiers be removed. One survey publicly released a dataset that included variables which made it easy to personally identify each respondent. The principles of the Helsinki Declaration should be upheld (see Ethical Considerations and International Statistical Institute (2009)).

8.2 When using official statistics for poststratification adjustments, consider the reputation of the agency. It has been suggested that some countries have manipulated official statistics. Examples of potential manipulations include the adjustment of agricultural outputs or redefining terms such as unemployment (European Commission, 2010; Rawski, 2001).

9. Document the steps taken in data processing and statistical adjustment.

Rationale

Over the course of many years, various researchers may wish to analyze the same survey dataset. In order to provide these different users with a clear sense of how and why the data were collected, it is critical that all properties of the dataset be documented.

Documentation will help secondary data users better understand post-survey statistical adjustments that can become quite intricate, such as the imputation procedures and the creation of survey weights for complex survey designs. A better understanding of these adjustments will help ensure that secondary data users correctly interpret the data. In addition, post-survey documentation will indicate whether the survey organization that conducted the survey met benchmarks agreed to in the contract by the coordinating center and the survey organization.

Procedural steps

9.1 Document the procedures and quality indicators of the data processing. Examples include:
  9.1.1 Data capture process.
  9.1.2 Versions of the data dictionary and codebook.
  9.1.3 Maintain code files used to process data.

9.1.4 Training protocol and manuals for data coding, entry, and editing.
9.1.5 What items were coded or recoded.
9.1.6 What items were edited and their original values.
9.1.7 How the raw data was edited.
9.1.8 Who coded, entered, and edited the data.
9.1.9 Evaluation protocol for data coding, entry, and editing.
9.1.10 Measure of coding reliability (e.g., Cohen's kappa). See Appendix B for more details.
9.1.11 Verification protocol for coding and entry.
9.1.12 Data entry accuracy rate.
9.1.13 Protocol for editing open-ended responses (e.g., removing identifying information, correcting typographical errors, standardizing language).

9.2 If values were imputed for specific variables in the study, clearly describe the imputation method that was used in the post processing documentation. In addition, for each variable where at least one value was imputed, create an imputation indicator variable that identifies whether a value was imputed for the specific variable or record in the dataset.

9.3 Create a unique identification number for each sampling unit. Describe how the sample identification numbers/codes were assigned to each element.
9.3.1 For internal use, create and document a sample identification number for each sampling unit. It is useful to have components of the identifier describe the case (e.g., 0600500200101: first two digits identify the country, the next three digits identify the area segment, the next three digits identify the sample replicate, the next three digits identify the household, the final two digits indicate the order of selection of the respondents within the unit, where 01=main respondent selected and 02=second respondent selected).
9.3.2 Create a separate unique identification number for public use data to prevent disclosing a respondent's identity. This number should contain no identifying information about responding units; it is simply a way to uniquely identify a case. The identifier could maintain any structure necessary for understanding the relationships of sample. For example, the identification numbers for members of the same household could have the same first 4 digits.
9.3.3 Sampling frame variables that could identify respondents should be included for internal use **only** (e.g., country two digits (06), area segment three digits (005), sample replicate three digits (002), household three digits (001), respondent

selected two digits (01), etc. Sampling information can be included in public use data provided it cannot be used to disclose a respondent's identity.

- For example, the sample identifier could be sensitive information if the user knew that the country was Japan, and the area segment was Hokkaido. Using this information, responses to rarely-occurring survey items, such as on crime victimization, could be used to search newspaper articles and discover the identity of the respondent.

9.3.4 For panel studies, endeavor to maintain the same identifiers for sample across data collection periods for both the internal and public-use data files. If this cannot be achieved, create a crosswalk table that links each identifier. This is crucial for data to be comparable across collection periods.

9.4 If survey weights were generated for the study, clearly explain how each individual weight adjustment was developed and how the final adjustment weight was calculated.

9.4.1 Each explanation should include both a written description and the formula used to calculate the weighting adjustment. Below are examples of the first sentence of an explanation for different weight adjustments from different countries. These are not meant to be exhaustive explanations, and the documentation of each adjustment should include further written descriptions and formulas.

- The base weight accounted for oversampling in the Wallonia region (Belgium) strata.
- The nonresponse adjustment was the inverse of response rate in each of three regions – Vlanders, Wallonia, and Brussels.
- The poststratification adjustment factor adjusted weighted survey counts to totals from Denmark's 2003 population register by sex, education, and age.
- As of March 1, 2004, a random half of the outstanding elements in the field were retained for additional follow-up efforts, and this subsample of elements was given an extra weight adjustment factor of $W = 1/.5 = 2.0$.

9.4.2 If additional adjustments were used to calculate a final weight, provide a clear description of how these components were created. Examples of additional weight components are country-specific adjustments and adjustments that account for differential probability of selection for certain questionnaire sections.

9.4.3 Address whether there was any trimming of the weights and, if so, the process used to trim the weights.

9.4.4 Address whether a procedure was used for scaling of the weights (e.g., population (N), population (N in 1000s), sample size (centered)).

9.4.5 If a replicated weighting method was used (i.e., Jackknife Repeated Replication or Balanced Repeated Replication – see Appendix F), provide the replicate weights for variance estimation.

9.4.6 Clearly describe how each of the survey weights and adjustments should be used in data analysis.

9.5 For complex survey data, identify the cluster and stratum assignment variables made available for sampling error calculations. For instance:

9.5.1 The variable that identifies the stratum to which each sample element and sample unit belongs.

9.5.2 The variable that identifies the sampling cluster to which each sample element and sample unit belongs.

- If the sample design has multiple stages of selection, document the variables that identify each unique sample element's primary sampling unit (PSU), secondary sampling unit (SSU), etc.
- If Balanced Repeated Replication variance estimation was used, identify the stratum-specific half sample variable, i.e., a field that identifies whether a unit is in the sampling error computation unit (SECU) 1 or 2.

9.6 If the risk of disclosing respondent identities is low, consider providing the different weight components on public use datasets. However, preventing disclosure of respondent identity takes priority over providing weight components.

9.7 Discuss whether the survey met the requirements (e.g., response rates, number of interviews) outlined in the contract.

9.7.1 If the requirements were not met, provide possible reasons why the survey failed to meet these requirements.

Lessons learned

9.1 Innovations for Poverty Action provides a good guide to data and coding management (Pollock, Chuang, Wystra, 2015).

9.2 The application of a unique identification code is often underestimated by survey agencies using their internal reference systems. For instance, a European survey implemented a two-year special panel survey where the agency conducting the study did not understand the need to link the two panel waves via one variable.

Hence, the agency provided a set of hard-to-interpret 'synthetic' codes that made it difficult for users to know if they were correctly analyzing the data. Much time and money were spent disentangling these codes and clarifying dubious cases.

9.3 Secondary users of survey data often have a hard time understanding when and if they should use weights in their analyses. This issue is exacerbated in many 3MC surveys, where participating countries may apply different nonresponse and postratification adjustment strategies. Without clear documentation of how each country created their survey weights and when to use each of the weights in data analysis, the chance of secondary users either not applying or incorrectly applying weights and producing estimates that do not accurately reflect the respective target population greatly increases. Therefore, clear documentation of the statistical adjustment processes is extremely important.

9.4 A good example of how to document the key elements of the statistical adjustment process can be found in National Survey of Family Growth, Cycle 6 (2006).

Appendix A

Example of a code frame for an open-ended question:

From the Survey of Consumer Attitudes (2012):

> A2. We are interested in how people are getting along financially these days. Would you say that you (and your family living there) are better off or worse off financially than you were a year ago?
> 1. BETTER NOW
> 3. SAME
> 5. WORSE NOW
> 8. DK
> 9. NA
>
> A2a. Why do you say so? (Are there any other reasons?).

REASONS FOR MAKING YOU BETTER OFF
10. Better pay: raise in wages or salary on present job, promotions, higher commissions, change to higher paying job (include Armed Forces induction or discharge) (Any family member who gets a raise is coded 10); increased tips, bonuses

11. Higher income from self-employment or property: higher business profits or farm income, higher dividends, royalties or rents, more income from professional practice or trade

12. More work, hence more income: Head (or wife) started working (again), more members of family working; higher income, NA why, MORE MONEY (if self-employed, code 11)

13. Increased contributions from outside FU: (from private individuals, government pension, relief or welfare, gifts); inheritance

14. Lower prices: decrease in cost of living; low or reasonable prices

15. Lower taxes; low or unchanged taxes

16. Decreased expenses: fewer people to be supported by FU; spending less, NA whether 14 or 16; thrift

18. Higher interest rates

19. Better asset position: more savings; business or farm worth more; has more business/farm assets; stocks went up; investments

20. Debt, interest or debt payments low or lower: have paid, is paying bills; interest rates lower

21. Change in family composition means higher income or better off (except 16 or 12); got married, etc. (no inheritance factor)

23. Good times, no recession (not codeable above) -- refers to the general situation as being good

27. Other reasons for making FU better off: great security (job more permanent, psychological security), greater opportunities, higher standard of living, have more things, future outlook improved, got insurance; bought house, additions and repairs to house

38. Reference to government economic policy

39. Income tax refund

REASONS FOR MAKING FU WORSE OFF

50. Lower pay: decrease in wages or salary on present job, change to lower paying job (including Armed Forces induction or discharge) (Any family member who has a decrease in wages or salary is coded 50); no increase in pay; decreased tips, bonuses

51. Lower income from self-employment or property: lower business profits or farm income, lower dividends, royalties or rents, less income from professional practice or trade

52. Less work, hence less income: unemployed (refers to any unemployed family member) laid off, sick, retired, on strike, unsteady work, less overtime, fewer members of FU working, back to student status, lower income NA why (if self-employed, code 51);WORSE off because R/family member is/has been sick

53. Decreased/Unchanged contributions from outside FU, "worse because Social Security hasn't gone up" (if "same" because Social Security hasn't gone up, DO NOT USE THIS CODE); "worse because on a fixed income"

54. High(er) prices: increase in cost of living; prices rise faster than income; inflation; worse because raises have been too small --code "no raise" or decrease in pay in 50

55. Higher interest rates

56. High, higher taxes (except 57)

57. Income taxes

58. Increased expenses; more people to be supported by FU; spending more, NA whether 54, 55, 56, or 58

59. Worse asset position: savings used up wholly or partially; less business, farm or personal assets; stocks declined in value; interest rates lower

60. Debt: interest, debt, or debt payments high or higher

61. Change in family composition means lower income or worse off (except 58); divorced, death, etc.

63. Bad times, recession (not codeable above--refers to the general situation as being bad)

64. Strike(s)--not codeable in 52

67. Other reasons for making FU worse off: less security (job less secure); lower standard of living

78. Reference to government economic policy

98. DK

99. NA

00. Inap, no change and no pro-con reason given; 9 in PAGO; no second mention

Appendix B

Cohen's kappa

Cohen's kappa can be used to assess the reliability of an individual (item) code.

- In the formula for kappa below, Pr(a) is the relative observed agreement among coders for a given item, and Pr(e) is the hypothetical probability of chance agreement in the observed data calculated from the probabilities of each coder randomly reporting each possible code category for that item.

$$\kappa = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)}$$

- If the coders are in complete agreement then kappa equals 1. If there is no agreement among the coders (other than what would be expected by chance) then kappa is less than or equal to 0.
- Kappa values between 0.7 – 0.8 are considered reliable.

Appendix C

Coder design effect (Groves, et al., 2009)

The coder design effect ($deff_c$) applies much of the same logic as the interviewer design effect (see Interviewer Recruitment, Selection, and Training chapter).

- In the formula for coder design effect below, ρ_c is the intraclass coefficient for coders, m is the average number of cases coded per coder, and r is the reliability of a particular (item) code.

$$deff_c = 1 + \rho_c\,(m\text{-}1)(1\text{-}r)$$

- The intraclass coefficient for coders is a measure of the ratio of coder variance to the total variance and is defined as:

$$\rho_c = \frac{(\textit{between-coder variance})}{(\textit{between-coder variance}) + (\textit{within-coder variance})}$$

Appendix D

Loss in precision of estimate due to weighting in household surveys

While overall survey weights help decrease three different sources of bias (coverage, nonresponse, and sampling), the variability of the weights also can increase the sampling variance in household surveys. The formula below is a simple model to measure the loss in precision (L_w) due to weighting. It assumes that the weights and the variable of interest are not related.

$$L_w = \left[\frac{\sum_{i=1}^{n} w_i^2}{\left(\sum_{i=1}^{n} w_i \right)^2} \right] (n) - 1$$

- For example, if $L_w = .156$, then the sampling variance of the estimate increased by 15.6% due to differential weighting.
- L_w can also be calculated for subgroups.
- Note: This formula does not apply to surveys of institutions or business establishments where differential weighting can be efficient.
- This is only one method for measuring the variability of the weights.

Appendix E

<u>Assumptions of missing data</u>

There are there mechanisms for missing data (Biemer & Lyberg, 2003). The difference between the three mechanisms depends on the relationship of the variable of interest to the missing observations and the variables available to explain the missingness.

- Missing Completely at Random (MCAR)
  - This missing data mechanism assumes the underlying process causing missing data are uncorrelated with any of the variables in the dataset. In other words, the probability of an observation for variable y being missing does not depend on measurements (x or y in the diagram below) in the dataset itself. An example of MCAR data are missing data due to an instrument malfunction. If MCAR holds, listwise deletion (i.e., an entire record is excluded from analysis if any one value is missing) can be employed because the available cases constitute a random subsample. Therefore, under MCAR, valid inferences to the target population can be made when analyzing only those units with complete data. If there are variables in the dataset (x, y) that help predict the missing values, the assumption does not hold. MCAR rarely holds, and, thus, listwise deletion will seldom be appropriate.
  - The concept of MCAR is illustrated below where y is the variable of interest with missing values, x is a predictor of y, m is the process causing missingness, and q is a variable not in the dataset.

$$x \longrightarrow y$$

$$q \longrightarrow m$$

- Missing at Random (MAR)
  - MAR is a weaker assumption about missingness than MCAR. In MAR, the process causing missing values can be explained by observed, non-missing data (x in the diagram below) other than the variable of interest (y). The probability of data missing on variable y is not related to the value of y, controlling for other variables. For data that are MCAR or MAR, the missing data mechanism is deemed ignorable. Note that the missing data *mechanism* is what is ignorable, not the missing data themselves. For data that are MAR, imputation will reduce bias.
  - The concept of MAR is illustrated below where y is the variable of interest with missing values, x is a predictor of y and also can

predict the mechanism for missing values, m. q is auxiliary to the dataset and also predicts m.

$$\begin{tikzcd} x \arrow[r] \arrow[dr] & y \\ q \arrow[r] & m \end{tikzcd}$$

- Missing Not at Random (MNAR)
  - For data that are MNAR, even after controlling for other observed variables in the dataset (x in the diagram below), the reason for a variable y having missing observations still depends on the unseen observations of y itself. One example of data that could be MNAR is reported income. Individuals with either high or low incomes can be reluctant to report how much they earn. If this is true, the probability of obtaining a measure of a person’s income will depend upon the amount the person earns. Nonignorable nonresponse creates data that are MNAR, and, hence, a method of imputation that accounts for this is necessary.
  - In the diagram below, y is the variable of interest with missing values, x is a predictor of y in the dataset, and q is unobserved auxiliary data. The three variables y, x, and q all predict m, the mechanism of missing values.

$$\begin{tikzcd} x \arrow[r] \arrow[dr] & y \arrow[d] \\ q \arrow[r] & m \end{tikzcd}$$

Appendix F

Estimating complex statistics when sample size is not fixed

Whenever the sample size is not fixed, use the Taylor Series estimation or one of the replicated methods, such as Balanced Repeated Replication (BRR) or Jackknife Repeated Replication (JRR), to estimate ratio means or other complex statistics.

- Taylor Series estimation.
  - This method computes the sampling variance of an approximation to a complex function like a ratio or regression coefficient. (See Kish, 1965 for the exact formulas.)
  - Advantages:
    - Used by most statistical software packages.
  - Disadvantages:
    - Requires analytic manipulations and computation of derivatives (but these have been done by developers of the software packages for common type of estimates).
    - Not useful if estimate cannot be expressed as a function of sample totals.
    - Taylor Series estimates in most software packages do not account for the variability of nonresponse adjustments.

- Balanced Repeated Replication (or Half-Sample Replication).
  - This method assumes a paired selection design (i.e., 2 PSUs per stratum) and selects H^* half sample replicates (H^* is the smallest multiple of 4 greater than or equal to the number of strata) by deleting one primary sampling unit (PSU) from each stratum according to the pattern in a Hadamard matrix. Each remaining element in the half sample receives a replicate weight of two. Fay's method of BRR is an alternative that retains both PSUs in a pair but modifies their survey weights (Kalton, 1983b).
  - Advantages:
    - More useful for complex estimates, such as medians, than Taylor Series.
    - Easily applied to user-specified statistics like differences or ratios of domain means.
    - Accounts for variability due to multiple steps in adjustment more easily than does Taylor Series.
  - Disadvantages:
    - Best used only with a paired selection stratification design.
    - Appending replicate weights to each record increases file size.

- Combining of strata and PSUs is sometimes done to reduce number of replicates. This must be done carefully to avoid biased variance estimates.

- Jackknife Repeated Replication.
  - This method creates a replicate by dropping a PSU from one stratum and weights up the other PSUs in the stratum to maintain the sampling distribution across the strata.
  - Advantages:
    - More useful for complex estimates than the Taylor Series.
    - Easily applied to user-specified statistics like differences or ratios of domain means.
    - Can handle designs other than paired selection.
    - Accounts for variability due to multiple steps in adjustment more easily than does Taylor Series.
  - Disadvantages:
    - Not appropriate for the variance of quantiles like the median.
    - Appending replicate weights to each record increases file size.
    - Combining of strata and PSUs is sometimes done to reduce number of replicates. This must be done carefully to avoid biased sampling variance estimates.

References

Afrobarometer Survey. (2014). Round 6 Survey Manual. Retrieved June 5, 2016, from http://www.afrobarometer.org/sites/default/files/survey_manuals/ab_r6_survey_manual_en.pdf

American Association for Public Opinion Research (AAPOR). (2015). *Standard definitions: Final dispositions of case codes and outcome rates for surveys* (8th ed.). Retrieved March 2, 2016, from https://www.esomar.org/uploads/public/knowledge-and-standards/codes-and-guidelines/AAPOR_Standard-Definitions2015_8theditionwithchanges_April2015_logo.pdf

Bethlehem, J. G. (2002). Weighting nonresponse adjustments based on auxiliary information. In R. M. Groves, A. D. Dillman, J. Eltinge, & R. Little (Eds.), *Survey nonresponse* (ch. 18). New York, NY: John Wiley & Sons.

Biemer, P. P., & Christ, S. L. (2008). Weighting survey data. In E. D. de Leeuw, J. J. Hox, & D. A. Dillman (Eds.), *International handbook of survey methodology.* New York/London: Lawrence Erlbaum Associates.

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

Brogan, D. (2005). Chapter XXI:Sampling error estimation for survey data. In United Nations Statistical Division, United Nations Department of Economic and Social Affairs (Eds.), *Household surveys in developing and transition countries*. New York, NY: United Nations. Retrieved April 8, 2015, from http://unstats.un.org/unsd/HHsurveys/pdf/Household_surveys.pdf

Busch C, De Maret P.S., Flynn T., Kellum R., Le S., Meyers B, Saunders M., White R., Palmquist M. (2012). Content Analysis. Writing@CSU. Colorado State University. Retroved Nov 9., 2015 from http://writing.colostate.edu/guides/guide.cfm?guideid=61

Dow, M. M., & Eff, E. A. (2009). Multiple imputation of missing data in cross-cultural samples. *Cross-Cultural Research, 43*(3), 206-229.

Eff, E. A., & Dow, M. M. (2009). How to deal with missing data and Galton’s problem in cross-cultural survey research: A primer for R. *Structure and Dynamics, 3*(3). Retrieved April 8, 2015, from http://www.escholarship.org/uc/item/7cm1f10b

Elias, P. (1997). Occupational classification (ISCO-88): Concepts, methods, reliability, validity and cross-national comparability. *Labour Market and Social Policy Occasional Papers* (No. 20). Paris: OECD.

European Commission. (2010). Report on Greek government deficit and debt statistics. Brussels: European Commission. Retrieved April 8, 2015, from http://ec.europa.eu/eurostat/documents/4187653/6404656/COM_2010_report_greek/c8523cfa-d3c1-4954-8ea1-64bb11e59b3a

European Social Survey. (2015). ESS5 – 2010 Documentation Report. Retrieved May 5, 2016, from http://www.europeansocialsurvey.org/docs/round5/survey/ESS5_data_documentation_report_e04_0.pdf

Federal Committee on Statistical Methodology. (1983). *Statistical policy working paper 9: Contracting for surveys*. Washington, DC: Office of Management and Budget.

Fellegi, I. P., & Holt, D. (1976). A systematic approach to automatic edit and imputation. *Journal of the American Statistical Association, 71*(353), 17-35.

Frick, J. R. & Grabka, M. (2007). Item Non-Response and Imputation of Annual Labor Income in Panel Surveys from a Cross-National Perspective.. IZA Discussion Paper No. 3043; SOEPpaper, No. 49. Available at SSRN: http://ssrn.com/abstract=1020605

Gabler, S., & Häder, S. (1997). Deviations from the population and optimal weights. In W. E. Saris, & M. Kaase (Eds.) *Eurobarometer: measurement instruments for opinions in Europe* (ZUMA-Nachrichten Spezial 2) (pp. 32-44). Mannheim, Germany: ZUMA. Retrieved April 8, 2015, from http://www.gesis.org/fileadmin/upload/forschung/publikationen/zeitschriften/zuma_nachrichten_spezial/znspezial2.pdf

Granquist, L., & Kovar, J. (1997). Editing of survey data: How much is enough? In L. E. Lyberg, P. P. Biemer, M. Collins, E. D. de Leeuw, C. Dippo, N. Schwarz. (Eds.), *Survey measurement and process quality* (pp. 415-435). New York, NY: John Wiley & Sons.

Groves, R. M., Fowler, F. J. Jr., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2009a). *Survey methodology* (2nd ed.). Hoboken, NJ: John Wiley & Sons.

International Standard Classification of Occupations (ISCO). (2007). ISCO-8 Structure and preliminary correspondence with ISCO-88. Retrieved November 9, 2015 from http://www.ilo.org/public/english/bureau/stat/isco/isco08/index.htm

International Statistical Institute. (2009). *International Statistical Institute declaration on professional ethics*. The Hague, The Netherlands: ISI. Retrieved March 01, 2010, from http://isi.cbs.nl/ethics0index.htm

Jordan, K., Porcheret, M., & Croft, P. (2004). Quality of morbidity coding in general practice computerized medical records: A systematic review. *Family Practice, 21*(4), 396-412.

Kalton, G. (1983b). *Compensating for missing survey data*. Ann Arbor, MI: Institute for Social Research, University of Michigan.

Kalton, G., & Kasprzyk, D. (1986). Treatment of missing survey data. *Survey Methodology, 12*, 1-16.

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

Kish, L. (1965). *Survey sampling.* New York, NY: John Wiley & Sons.

Lepkowski, J. M., & Bowles, J. (1996). Sampling error software for personal computers. *The Survey Statistician*, *35*, 10-17.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). New York, NY: John Wiley & Sons.

Living Standard Measurement Study Survey. (1996). Working paper 126: A manual for planning and implementing the Living Standard Measurement Study Survey. Retrieved July April 8, 2015, from http://www-wds.worldbank.org/external/default/WDSContentServer/WDSP/IB/2000/02/24/000009265_3961219093409/Rendered/PDF/multi_page.pdf

Lyberg, L & 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. Lyberg, P.P. Mohler, B. Pennell, T.W. Smith(Eds.) Survey Methods in Multinational, Multiregional, and Multicultural Contexts. (ch. 13). Hoboken, NJ. John Wiley & Sons.

Marker, D. A, Judkins, D. R., & Winglee, M. (2002). Large-scale imputation for complex surveys. In R. M. Groves, D. A. Dillman, J. Eltinge, & R. Little (Eds.), *Survey nonresponse*. New York, NY: John Wiley & Sons.

National Survey of Family Growth. (2006). Cycle 6: Sample Design, Weighting, Imputation, and Variance Estimation. Series 2, Number 142.

Office of Management and Budget. (2006). *Standards and guidelines for statistical surveys*. Washington, DC: Office of Information and Regulatory Affairs, OMB. Retrieved April 10, 2015, from http://www.whitehouse.gov/omb/inforeg/statpolicy/standards_stat_surveys.pdf

O'Reagan, Robert T., Maureen P. Lynch, and Susan M. Odell. (1988) "INDEPENDENT VERSUS DEPENDENT QUALITY CONTROL OF CLERICAL CODING: ONE ENLIGHTENING EXPERIENCE." Retrieved July 29, 2015 from http://www.amstat.org/sections/srms/proceedings/papers/1988_092.pdf

Pollock, H. D., Chuang E., Wykstra S. (2015). *Reproducible Research: Best Practices for Data and Code Management*. Innovations for Poverty Action. Retrieved April 22, 2016 from http://www.poverty-action.org/publication/ipas-best-practices-data-and-code-management

Rawski, T. G. (2001). What is happening to China's GDP statistics? *China Economic Review. 12*(4)*,* 347-354.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York, NY: John Wiley & Sons.

Saldaña, J. (2012). The coding manual for qualitative researchers (No. 14). Sage.

Särndal, C. E., & Lundström, S. (2005). *Estimation in surveys with nonresponse*. *New* York, NY: John Wiley & Sons.

Singh, M. P., Hidiroglou, M. A., Gambino, J. G., & Kovačević, M. S. (2001). Estimation methods and related systems at Statistics Canada. *International Statistical Review / Revue Internationale de Statistique, 69*(3), 461-486.

Statistics Canada. (2009). *Statistics Canada quality guidelines.* Ottawa: Statistics Canada. Retrieved April 8, 2015, from http://unstats.un.org/unsd/dnss/docs-nqaf/Canada-12-539-x2009001-eng.pdf

Taylor, C., & Gibbs, G. R. (2010) "How and what to code", *Online QDA Web Site.* Retrieved January 11, 2016 from http://onlineqda.hud.ac.uk/Intro_QDA/how_what_to_code.php

Tourigny, J. Y.,& Moloney J. (1991). The 1991 Canadian Census of Population Experience with Automated Coding. Chapter 6 Automated Coding. United Nations Economic Commission for Europe. Retrieved August 6, 2015 from http://www.unece.org/fileadmin/DAM/stats/publications/editing/SDE2ch6.pdf

United Nations. (2005). *Household surveys in developing and transition countries.* New York, NY: United Nations, Department of Economic and Social Affairs. Retrieved April 8, 2015, from http://unstats.un.org/unsd/HHsurveys/pdf/Household_surveys.pdf

University of Michigan. Survey Research Center. Economic Behavior Program. Survey of Consumer Attitudes and Behavior, December 2012. ICPSR35464-v1. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2014-10-27. http://doi.org/10.3886/ICPSR35464.v1 van de Vijver, F.J.R., & Chasiotis, A. (2010). Making methods meet: Mixed designs in cross-cultural research. In J.A. Harkness, M. Braun, B. Edwards, T.P. Johnson, L. Lyberg, P. Mohler, B-E. Pennell & T.W. Smith (Eds.), *Survey methods in multinational, multiregional, and multicultural contexts* (pp. 455-473). Hoboken, NJ: John Wiley & Sons.

Vermaak, C. (2009). *The impact of multiple imputation on estimates of the working poor in South Africa.* Paper presented at the Development Policy Research Unit Conference "The Regulatory Environment and Its Impact on the Nature and Level of Economic Growth in South Africa."

Wagner, J. & Stoop, I. (Forthcoming). Comparing nonresponse and nonresponse biases cross-nationally. In T. P. Johnson, B-E. Pennell, I. A. L. Stoop, & B. Dorer (Eds.), *Advances in Comparative Survey Methods: Multicultural, Multinational and Multiregional Contexts (3MC).* New York, NY: John Wiley & Sons.

Wurdeman, K. (1993). *Quality of data keying for major operations of the 1990 census.* Paper presented at the Proceedings of American Statistical Association: Survey Research Methods Section. Retrieved April 8 2015, from http://www.amstat.org/Sections/Srms/Proceedings/papers/1993_049.pdf

Internet Links

Afrobarometer Survey. Retrieved June 1, 2016, from http://www.afrobarometer.org/sites/default/files/survey_manuals/ab_r6_survey_manual_en.pdf

Asian Barometer. Retrieved June 1, 2016, from http://www.asianbarometer.org/

Census and Survey Processing System (CSPro). (2010). Retrieved May 6, 2010, from http://www.census.gov/population/international/software/cspro/

European Social Survey (ESS). (2010). Retrieved July 10, 2015, from http://www.europeansocialsurvey.org

Imputation and Variance Estimation Software (IVEWare). (2009). Retrieved May 6, 2010, from http://www.isr.umich.edu/src/smp/ive/

Further Reading

Alexa, M. (1997). Computer assisted text analysis methodology in the Social Sciences. ZUMA.

Groves, R. M. (1989). *Survey errors and survey costs*. Hoboken, NJ: John Wiley & Sons.

Groves, R. M., Dillman, D. A., Eltinge, J. L., & Little, R. J. A. (Eds.), (2002). *Survey nonresponse*. New York, NY: John Wiley & Sons.

Horvitz, D. G., & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, *47*(260), 663-665.

Kish, L., & Hess, I. (1959). On variances of ratios and their differences in multistage samples. *Journal of the American Statistical Association*, *54*(286), 416-446.

Lavallée, P., & Beaumont, J.F. (2016). Weighting: principles and practicalities. In C. Wolf, D. Joye, T.W. Smith, & Y. Fu (Eds.), *The SAGE Handbook of Survey Methodology*. Thousand Oaks, CA: Sage.

Lessler, J., & Kalsbeek, W. (1992). *Nonsampling error in surveys*. New York, NY: John Wiley & Sons.

Lyberg, L. E., Biemer, P. P., Collins, M., de Leeuw, E. D., Dippo, C., Schwarz, N. et al. (Eds.). (1997). *Survey measurement and process quality*. New York, NY: John Wiley & Sons.

Raghunathan, T. E., Lepkowski, J. M., van Hoewyk, J., & Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology, 27*(1), 85-95.

Smith, T. W. (2003). A review of methods to estimate the status of cases with unknown eligibility. *Report of the Standard Definitions Committee for the American Association for Public Opinion Research.*

Spiess, M. (2016). Dealing with missing values. In C. Wolf, D. Joye, T.W. Smith, & Y. Fu (Eds.), *The SAGE Handbook of Survey Methodology*. Thousand Oaks, CA: Sage.

Wirth, H. (2016). Anonymization and analytical potential: finding the optimal balance. In C. Wolf, D. Joye, T.W. Smith, & Y. Fu (Eds.), *The SAGE Handbook of Survey Methodology*. Thousand Oaks, CA: Sage.

Worcester, R., Lagos, M., & Basañez, M. (2000). *Problems and progress in cross-national studies: Lessons learned the hard way.* Paper presented at the 2000 AAPOR/WAPOR Conference, Portland, OR.

Data Dissemination

Peter Granda and Emily Blasczyk, 2016

Introduction

Dissemination is the process by which producers of microdata from surveys and from public and official statistics make their data available to other users. These users may include government officials, academic researchers, policymakers, and the general public. Data may be disseminated publicly without any restrictions (public use files) or only to certain users under specific conditions (restricted use files). The availability of microdata is often dependent on national laws and regulations. Data and documentation may be disseminated in various formats but the goal is to provide complete information in a non-proprietary format that is amenable to long-term preservation.

Several aspects of making data and documentation files available to analysts require special consideration. More is involved in the dissemination process than merely providing data access to interested researchers. Data producers and archivists must assure analysts that the data they provide accurately reflects the efforts of the data collection process, is trustworthy, fully documented, has no confidentiality concerns, and is securely preserved for future use. Disseminating data from multinational, multicultural, or multiregional surveys, which we refer to as “3MC” surveys, can include specific processes such as standardization, harmonization, and multi-lingual documentation which may not apply to surveys done in a single country.

An additional aspect of dissemination is how to share research findings with interested parties. Determining who is using the data and why they are using it is important to consider as part of a comprehensive dissemination strategy. Many international organizations, social science data archives, and survey research projects also embrace these objectives. Although focused on micro-economic data, the International Monetary Fund (IMF), for example, established a set of guidelines on macroeconomic data for member countries to follow in order to provide the public with “comprehensive, timely, accessible, and reliable economic, financial, and socio-demographic data” (IMF, 2015; Gutmann et al., 2009).

Guidelines

Goal: To ensure that survey and statistical research teams in all cultures and countries involved in a 3MC survey follow accepted standards for the long-term preservation and dissemination of data to the social science research community and the wider public.

1. Make a dissemination and data preservation plan that includes archiving, publishing, and distribution, early in the project lifecycle.

Rationale

Dissemination is an integral part of the survey research process. It involves the documentation of major steps in the data lifecycle from initial planning to the production of final data files. This includes, when available and appropriate, detailed information about the survey process (paradata), all data editing steps, and protocols which determine what types of data and documentation files are made available to which users.

Procedural steps

1.1 For multi-lingual surveys, decide on the standard documentation language to be used.

1.2 Identify any documents that should be published in their original language such as individual country questionnaires, codes, verbatim responses and nation-specific data files.

1.3 Have a system in place to preserve all major planning and operational documents as soon as they are created.

1.4 Consider including information about the survey process when disseminating data, documentation, and reports. Producers may want to balance the amount of paradata they release with the need to maintain proprietary information about the data collection process.

Lessons learned

1.1 All studies must develop a system for preserving and storing materials. There are a variety of methods that can be utilized which rely on centralized depository. Some examples of dissemination strategies are below:

1.2 Round 4 of the Afrobarometer Survey strongly recommends that participating countries scan their completed paper-and-pencil questionnaires. Hard copies are acceptable where circumstances (e.g., cost) prevent scanning. National partners are responsible for either the scanning or the storing of their own questionnaires. Each national partner is responsible for entering and cleaning their own data and delivering a clean SPSS data set (Afrobarometer Survey, 2014).

1.3 The Demographic and Health Surveys Program (DHS) (2015) provides to users both the raw data and a "standard recode" datasets. The recode datasets contain the same data as the raw datasets, but in a standardized format where variable names and definitions are, wherever possible, consistent across all surveys. DHS also provides standard data tabulations (http://www.dhsprogram.com).

1.4 All documents related to each round of the European Social Survey (ESS) are uploaded to a server. This includes, but is not limited to, original unedited (raw) data, fieldwork documents, metadata, and population statistics for coverage and response rates (ESS, 2015).

1.5 Documentation of International Social Survey Programme (ISSP) survey methods and data files are sent to a central data archive no later than nine months after fieldwork is completed. Data are to be sent unweighted, but descriptions of weighting procedures should accompany the datasets (ISSP, 2015).

1.6 Master copies of all important Living Standard Measurement Study Survey (LSMS) files are kept in a separate archive which is backed-up (see http://go.worldbank.org/BKW704K6Q0).

1.7 Documentation for the World Mental Health (WMH) Survey is done using the Survey Metadata Documentation System designed by the WMH Data Collection Coordination Centre (Kessler, Ustun, & World Health Organization, 2008).

1.8 Countries participating in the World Values Survey are required to submit documentation of their survey methods and data to a central data archive no later than three months after fieldwork has been completed. Documentation must include a completed methodology questionnaire, a report of any questions omitted or added to the original official questionnaire, a report of additional and/or country specific codes to any questions, official demographic statistics, weights used, and a copy of the original country questionnaire (http://www.worldvaluessurvey.org)

1.9 Many institutions which provide research grants for data collection now strongly recommend that grantees prepare a data sharing plan as part of the proposal process. The National Institutes of Health in the United States (NIH) provide the following justification for their emphasis on dissemination: “Data sharing promotes many goals of the NIH research endeavor. It is particularly important for unique data that cannot be readily replicated. Data sharing allows scientists to expedite the translation of research results into knowledge, products, and procedures to improve human health. There are many reasons to share data from NIH-supported studies. Sharing data reinforces open scientific inquiry, encourages diversity of analysis and opinion, promotes new research, makes possible the testing of new or alternative hypotheses and methods of analysis, supports studies on data collection methods and measurement, facilitates the education of new researchers, enables the exploration of topics not envisioned by the initial investigators, and permits the creation of new datasets when data from multiple sources are combined." This policy has resulted in more data becoming available in the public domain (NIH Data Sharing Policy and Implementation Guidance, 2003).

1.10 The International Federation of Data Organizations conducted an informal web survey of institutional data policies in the social sciences in 2013. IFDO found there was a growing awareness and interest in data sharing. However, the implementation varies across countries and research funders. The results indicate that the social sciences have more developed policies than the medical and health science (International Federation of Data Organizations for Social Science, 2014).

1.11 More than ten years ago the International Monetary Fund (IMF) began to develop a set of dissemination standards "to guide countries in the provision to the public of comprehensive, timely, accessible, and reliable economic, financial, and socio-demographic data" (IMF, 2015). These standards were considered best practices but their implementation was completely voluntary depending on the policies and wishes of each nation. The Fund published a report (Alexander, Cady, & Gonzalez-Garcia, 2008) about the success of this initiative over the first ten years of the initiative. It concluded that more accurate and reliable statistical information is now being produced by more nations than ever before but also recognized that dissemination mechanisms are not fully developed in many locations. Nations also have internal challenges and constraints in addressing dissemination goals from resource constraints, shifting priorities, and in their ability to generate periodic and timely statistical data.

1.12 The Organisation for Economic Co-Operation and Development has produced guidelines for access to research data from public funding. It aims to help governments, research support and funding organizations, research institutions and researchers themselves in deal with challenges in improving the international access and sharing of research data (Organisation for Economic Co-Operation and Development, 2007).

2. Preserve sustainable copies of all key data and documentation files produced during the data collection process, as well as those made available for secondary analyses.

Rationale

Preservation is an important part of the survey lifecycle, a prerequisite for long-term access to valuable physical objects and digital materials. The materials that need to be preserved and kept available to members of the research community include such objects as public use data and documentation files (including key files used in their construction), copies of the data collection instruments, user guides, information about the data collection process, and reports on field operations. Since dissemination policies may differ among countries, it is important that data producers take the necessary steps to make their collections as accessible as possible to members of the research community. If appropriate repositories are not available, producers may need to organize dissemination of their materials themselves.

Procedural steps

2.1 Define the long-term preservation standards and protocols to be used. Consider digitizing physical objects, commonly-used questionnaires, or other administrative materials documenting the whole data lifecycle including the design phase of the project.

2.2 There are several digital preservation metrics that can be used to assess digital repositories. Two metrics are the Trusted Repository Audit Checklist (TRAC) and The Trusted Digital Repository Checklist (TDR) or ISO 16363 (Center for Research Libraries, 2015).

2.3 Protect digital materials through storage of multiple copies in multiple locations. An ideal preservation storage situation includes a minimum of several off-site copies of digital materials undergoing regularly scheduled back-ups. If it is not possible to store materials at multiple sites, preserve at least one copy in a different location.

2.4 Make certain that digital materials remain retrievable through constant refreshment of the media on which they are stored. This is particularly important if removable media, such as tapes, are used for storage, since formats and the machines required to read these media change quickly over time.

2.5 Implement a system of version control to maintain older versions of important data and documentation files. Users should be able to follow the changes made from one version to the next. Version control is necessary for users to replicate previous analysis or to test analysis done by others.

2.6 At a minimum, store a copy of all data and metadata files in software-independent formats such as ASCII files or XML which, with proper accompanying documentation, can be read into all major statistical packages.

2.7 Investigate the protocols and standards of digital repositories, such as availability of extracting data and in the areas of multi-site storage, security, and costs.

2.8 Make test runs of copied data to ensure error-free copy processes.

2.9 Work if possible with a trusted digital repository, such as a national or public social science data archive, to preserve all study materials. In doing so, data producers do their best to ensure that their data collections will remain available to the research community.
 2.9.1 Such repositories make an explicit commitment to preserving digital information by:
 - Complying with the Open Archival Information System (OAIS) in the US and other similar standards in other countries which have their own digital preservation standards and practices (National Digital Archive of Datasets, 2010; Royal Statistical Society & the UK Data Archive, 2002; Van Diessen & Steenbergen, 2002).
 - Ensuring that digital content can be provided to users and exchanged with archives without damaging its integrity.
 - Participating in the development and promotion of digital preservation community standards, practice, and research-based solutions.
 - Developing a reliable, sustainable, and auditable digital preservation repository that has the flexibility to grow and expand.
 - Managing the hardware, software, and storage media components of the digital preservation function in accordance with environmental standards, quality control specifications, and security requirements

2.10 If no national or public social science data archives exist, consider depositing data with an archive in another country or investigate the possibility of doing so with a national statistical agency or certified provider. Consider archiving collections in one archive which would keep master copies of files in several locations but minimize the possibility of conflicting versions of data and documentation files.

Lessons learned

2.1 The German National Science Foundation (2013) requires data to be archived for a minimum of 10 years as part of its anti-fraud activities.

2.2 Some earlier studies, such as older Eurobarometer surveys, did not preserve individual country data, thus issues about harmonization emerging some decades later could not be easily settled.

2.3 Data producers should make every effort to extract data that is on media which may no longer be easy to read. Too many data files have been irretrievably lost because the files were never copied to newer types of media.

3. Conduct effective disclosure analysis to protect respondent confidentiality.

Rationale

Any plan to disseminate survey data must include very specific procedures for understanding and minimizing the risk of breaching the promise of confidentiality that is made to respondents at the time of the survey or collection of data. The key goal of disclosure risk analysis and processing is to ensure that the data maintain the greatest potential usefulness while simultaneously offering the strongest possible protection to the confidentiality of the individual respondents. Disclosure analysis has become increasingly important as more and more datasets become available online and as the possibility of linking survey data to other contextual and administrative databases has grown exponentially (Inter-university Consortium for Political and Social Research [ICPSR], 2015; National Human Research Protections Advisory Committee, 2002).

Procedural steps

3.1 Be aware of and adhere to the different legislation for disclosure control in each country.

3.2 Disclosures can be categorized as: identity disclosure or attribute disclosure (Hundepool et al., 2012).

 3.2.1 Identity disclosure results from using a single identifying characteristic or combination of characteristics to discover an individual respondent (e.g. name and address).

 3.2.2 Attribute disclosure results from using a combination of indirect characteristics to associate a given with an individual (e.g. an outlier for a large number of employees and the industry sector in a business survey identifies the company, which reveals the company’s annual employee turnover).

3.3 Implement a disclosure protocol. A proper disclosure protocol includes an analysis of the most likely outside sources which might allow the identification of respondents or households.

3.4 Search systematically in the data file for sensitive information such as transcripts of open ended answers including International Standard Classification of Occupations (ISCO) occupational variables, identification of PSUs, birth dates, income, or housing and dwelling information.

3.5 Search also for unusual characteristics and for cells in tables with very low frequencies.

3.6 Undertake both practical and statistical steps to identify cases and variables. This allows the identification of areas or variables that need to be further masked in order to prevent identification of subjects, either through analysis or by matching study data with data from other external databases. After having decided on which variables present unacceptable risks, mask the relevant information.

3.7 Evaluate data files once those cases and variables are identified. In virtually every case, the data can be masked in various ways that make it possible for public use data to be distributed, usually through a Web-based system.

3.8 Use appropriate masking procedures to preserve respondent confidentiality while also trying to optimize the usefulness of the resultant data file for analysis. These procedures might include top or bottom coding of key demographic variables such as income, removing data for very sensitive variables, and swapping data values between similar cases (O'Rourke et al., 2006).

3.9 Document all confidentiality assurance processes and make a final assessment about the anonymity of the data file.

Lessons learned

3.1 With the enhanced emphasis on privacy in almost all countries, confidentiality reviews of microdata are increasingly important, if not indispensable, to assuring the future availability of public use data.

3.2 A 2011 experiment used individual-level reoffending and sentencing data in the UK to demonstrate the possibility of disclosure prior to public release. Disclosure resulted from matching data to a local news website (Tudor, Cornish, & Spicer, 2014).

3.3 The practice of reporting examples of privacy violations, particularly in the health care field in the United States, has increased awareness of this issue (Health Insurance Portability and Accountability Act Privacy and Security, 2015).

4. **Consider the production of both public- and restricted-use data files.**

Rationale

In order to ensure that researchers have access to the greatest amount of data without compromising respondent confidentiality, data producers, when appropriate, must make every effort to create both public- and restricted-data documentation files, and make these files available to the research community through secure and predictable channels.

Procedural steps

4.1 Make data files fully available to the research community as soon as possible within the confines of how the project is organized and financed. If general distribution is not feasible, establish clear rules under which researchers can obtain the data.

4.2 Remain cognizant of any copyright restrictions that data may have. In some cases, even after dissemination the ownership of the data remains with the principal investigators.

4.3 Provide access directly by the data producer if resources permit, but also always send copies to a trusted digital repository for permanent preservation, in case the data producer should cease to provide access at some time in the future.

4.4 Consider the creation of less thoroughly masked versions that can be distributed under restricted-use contracts, or made available within a research data center or "enclave" (i.e., a secure environment in which the user has access to restricted data and analytic outputs under controlled conditions).

4.5 Establish clear policies for how researchers may access restricted data files by creating a set of application materials and restricted-use data agreements that specify how researchers can obtain and use such data (ICPSR, 2012).

4.6 Distribute restricted files through signed data use agreements. These may incorporate data protection plans, formal licenses, and travel to a special facility at which researchers can access the data in a very controlled environment.

4.7 Create special files for researchers that cannot be matched with public use files (for example, provide finer grained local information and simultaneously change respondents' IDs and other matching variables).

Lessons learned

4.1 Consider making clear agreements on data heritage (i.e., copyright transfer after the original principal investigator retires). A German elite study was nearly lost to the academic public due to heritage issues.

4.2 Most data are already paid for by taxpayer money or foundations. Thus foundations and public funders often ask for free data access (i.e., they deny the principal investigator's sole ownership on collected data).

4.3 Despite general agreement about the advantages of making data accessible to other researchers, as well as strong data-sharing cultures in many nations, too few social science data collections are effectively preserved. Data archives should do as much as possible to facilitate the deposit process by contacting principal investigators and data producers as they prepare data and documentation files.

5. Produce data files that are easy for researchers to use.

Rationale

An effective data processing strategy focuses on the production of data files that will provide optimal utility for researchers. Such files have been thoroughly checked and cleaned, possess uniform and consistent coding strategies, use common formats, and address the potential research needs of secondary analysts.

Procedural steps

Processors should perform a series of steps to ensure the integrity and maximum utility of public-use files. Such steps include:

5.1 Address the various ways data may be utilized by creating tools within a web-based system that permits online analysis, subsetting, and access to documentation. Be aware that online analysis must use fully anonymized data. Data users may be policymakers seeking summary information, analysts browsing for new data sources, or individuals seeking summary analytic information, or wanting to quickly download specific variables.

5.2 In order to provide optimal utility for researchers, produce a variety of products for varied constituencies.
  5.2.1 Produce setup files and ready-to-use 'portable' files in SAS, SPSS, and Stata to address the needs of those who seek to do intensive statistical analyses with particular software packages.
  5.2.2 Consider disseminating data on removable media (e.g., CD-ROM or DVD) if appropriate.
  5.2.3 Clearly identify the master version and provide access to any previously released versions.

5.3 Format the data files in a way that permits access through a wide variety of statistical packages, all of which will produce the same results no matter how complicated the analysis requested, particularly with any variable where decimal precision is an important consideration.

5.4 Consider creating simplified versions of datasets for use by a wider public such as journalists and policymakers (i.e., by creating recode variables such as age of respondents in groups, income in groups, removing detailed information such as household lists, setting missing data properly, etc.). Make such datasets accessible via web-analysis.

5.5 Make a thorough investigation of any undocumented code numbers or inconsistent responses. Whenever possible provide labels for such codes such as 'not ascertained' if there is no alternative.

5.6 Standardize all missing data values, unless it is not possible to do so because of different cultural understandings (flag such issues carefully). Users doing analyses will appreciate that all "does not apply," "don't know," "refused," and "no data available" responses are coded the same way in the data file.

5.7 Create complete and concise variable and value labels which will provide researchers with clear descriptions of their analytic results.

5.8 Provide a printable questionnaire that contains all variable names and values in an appropriate format.

5.9 Consider producing ancillary files for those data collection efforts which cover multiple waves of respondents or several geographic areas. Such files may include recoded variables to summarize information contained in many questions or special constructed variables that producers feel will aid researchers in their analyses.

5.10 Create special subsets of data which take advantage of the longitudinal richness of long-term collections and provide unique opportunities to study important social, political, and economic issues from different perspectives, particularly with regard to the changing characteristics of the sampled respondents. Some examples include:

5.10.1 The Integrated Demographic and Health Series (IDHS) project integrated a subset of data from the Demographic and Health Surveys for women of childbearing age and their children from 18 countries (IDHS, 2015).

5.10.2 The International Social Survey Programme created modules on specific topics that integrated data for repeating years and across countries. Example modules include Religion; Role of Government; and Leisure Time and Sports (ISSP, 2015).

5.11 Whenever possible and expedient, make individual country datasets available in 3MC surveys.

Lessons learned

5.1 Users increasingly expect data files to come in a variety of formats that will work easily with their statistical package of choice. In some settings this may be just an SPSS portable file, but in others data producers and/or archives might need to create the same file in a variety of formats, particularly if a standard database conversion package, such as STAT-TRANSFER, is not available.

5.2 Be very clear about coding responses that refer to “item response refused,” “item response does not apply due to filtering,” “can’t choose all” or “don’t know,” and especially “no code in data file where a code should be.” All these have different meanings and must get different values. The “no code in data file” indicates either an interviewer error or error in data editing.

5.3 “Don’t know”/ “Can’t choose” responses may have different meaning in different countries based on different response styles. Treating all of these responses as missing data may lead to unwarranted conclusions about the attitudes of whole populations (Sicinski, 1970).

5.4 Established 3MC studies share their data in a variety of ways:

5.4.1 The Afrobarometer Survey publicly releases all data and documentation via their website, one year after the completion of fieldwork (Afrobarometer Survey, 2014).

5.4.2 The European Social Survey (ESS) releases anonymized data onto the public website within one year of the onset of data collection (ESS, 2015).

5.4.3 The International Social Survey Programme makes individual national and/or combined datasets available to the scientific community by the Data Archive one year after the calendar year to which it relates (ISSP, 2015).

5.4.4 Living Standard Measurement Study Survey (LSMS) data are usually available within twelve months of the end of fieldwork and is published on the World Bank website for the LSMS study, as well as each country's statistics office website. (http://go.worldbank.org/BKW704K6Q0).

5.4.5 Survey of Health, Ageing, and Retirement in Europe (SHARE) data are distributed through their Research Data Center (http://www.share-project.org).

5.4.6 The World Values Survey provides data only to participating countries for a period of two years after fieldwork has been completed; after this period, the data are made available to the worldwide social science community in the form of data archives (http://www.worldvaluessurvey.org).

6. Develop finding aids to guide users in their quest to locate data collections they want to use.

Rationale

The capability to query for specific information is critical to all data dissemination systems, from individual data producers, with only a few data collections, to social science archives with thousands of such collections.

Procedural steps

6.1 Create a robust search engine to query the fielded metadata so that the user can find variables of interest efficiently.

6.2 Allow the search engine to run against a study's bibliography to enable two-way linking between variables and publications based on analyses of those variables.

6.3 Display the abstracts of the publications with links to the full text whenever possible, in order to realize the full potential of the online research environment.

6.4 Dedicate staff time to continuously search journals and online databases to discover new citations where the data have been used. Many search engines have the ability to set up "alerts" that notify a user when new items are found based on a query.

6.5 Encourage data archives to create metadata records for surveys they do not preserve and distribute these records to facilitate their discovery and use.

Lessons learned

6.1 Data usage increases when the data are easy to find and when users know of publications scholars have produced from the data. There are many datasets that would be of interest to secondary analysts if the analysts only knew about them. For example, many surveys were conducted in Latin America and Africa in the 1960s and 1970s which might offer opportunities for interesting comparative analyses with the more recent and much more popular Latino and Afrobarometer surveys. These are not always as visible to researchers, however, as they might not possess immediately obvious substantive or methodological interest.

7. Create comprehensive training, outreach, and user support programs to inform the research community about the dataset.

Rationale

Training and support of users will increase usage of the data and encourage comprehensive analyses. It is very important that major survey research producers or archives reach out to the user community effectively, in order to explain the structure of new datasets and to encourage the greatest possible use. The most straightforward way to reach out is to develop an effective online presence, ensuring that the data are easily located and acquired, and that metadata and bibliographical citations are also available. Good user support will prevent obvious misuse or possible misunderstanding of the structure and content of the dataset.

Procedural steps

7.1 Organize workshops at relevant professional organizations or attend conferences where 3MC research is a focus soon after the data are released, in order to bring early users together to discuss important preliminary results, as well as to ensure that the data are used effectively and that any problems with the data are recognized and corrected.

7.2 Maintain a presence at professional meetings even when the data have been released for a long time. Staff from the project can describe the data, distribute documentation and sample data and encourage researchers to make use of the data.

7.3 Hold training workshops in different countries to ensure that novice users have a chance to learn about the data from experts and, if possible, from the data production team itself. Users should learn about specific issues involved in data collected in their own countries as well as how comparable the data collection experience was in other countries.
   - 7.3.1 Without specialized instruction and training, analyses of cross-cultural, longitudinal data and repeated cross-sectional data are particularly challenging.
   - 7.3.2 These training courses can be brief half-day or one-day sessions at the time of professional meetings, or they can continue for longer periods (e.g., three- or five-day sessions with a more detailed focus).
   - 7.3.3 Provide the training materials online so people unable to attend can have access to the information.

7.4 Provide easy access to user support through phone, email, online chat, user forums, and tutorials.

7.5 Track all user questions in a database that creates an accumulating knowledge base and that can also serve to generate Frequently Asked Questions.

7.6 Create tutorials, some of which may be offered in video format, to provide help in using the data, the online analysis system, and the major statistical software packages.

7.7 Establish moderated user forums to provide the foundation for an online community of researchers and students who can discuss their experiences using data and learn from each other.

7.8 While all of these procedures can increase the effective use of 3MC datasets, each country must decide on which steps would be most beneficial for their own research communities.

Lessons learned

7.1 In order for participants to fully benefit from the experience, training programs must be well-planned, with a high level of substantive, methodological, and technical expertise. While data producers are usually those who best understand their data, they may not have the resources or desire to provide ongoing user support for the research community. Some may delegate this task to a data archive, but a joint approach, with data archives providing basic user support and data producers addressing more complicated substantive questions, often works best. In countries where national data archives do not exist, data producers may want to partner with university social science departments or research centers to increase awareness and use of important datasets.

7.2 Complex data sets often require specialized training. Data collection methods or sampling frames often change between different waves or in different countries and weighting variables may require extensive descriptions. In this context, there is no real substitute for intensive training and ongoing user support.

7.3 The Demographic Health Surveys have an online user forum for users to post and discuss issues (http://www.dhsprogram.com/). This type of tool is increasingly more common.

8. Produce comprehensive documentation for all public and restricted use data files.

Rationale

High-quality documentation is essential for effective data use in all surveys but particularly in 3MC datasets because of the need to provide comparable information from all countries or study populations. As resources permit, data producers must strive to provide documentation, commonly referred to as metadata, on all aspects of the survey or statistical lifecycle, from initial planning through final data production and its release to the research community. For more information on data processing techniques used preceding dissemination, see Data Processing and Statistical Adjustment.

Procedural steps

8.1 Keep detailed records from the very beginning of the project and make every attempt to record important project events at the time they occurred. This will assist analysts in understanding the goals and purpose of each survey.

8.2 Update documentation continually during the entire lifecycle of the project and preserve old versions of key files.

8.3 For 3MC surveys, provide complete information about how the survey was conducted in each country or study population, and describe specific procedures and practices involving data collection and data processing activities.

8.4 Consider adopting the Data Documentation Initiative (DDI) standard for producing metadata (DDI, 2015). The use of this standard, which is based on the use of Extensible Markup Language (XML), allows for specification of each metadata element (e.g., title of the survey, name of the principal investigators, type of sampling) for storage and future searching.

8.4.1 Define a database structure that will be used to store XML elements.

8.4.2 Identify appropriate tools that will access and create XML coded information in a natural language environment such as a browser displaying a web-based form generator.

8.5 XML metadata markup offers opportunities for data producers to create their documentation, as well as several advantages to users of the documentation:

8.5.1 All information that the analyst needs is available in a core document, from which other products, such as text files that contain the necessary information to run statistical analyses in software programs, can be produced.

8.5.2 The XML file can be viewed with Web browsers and lends itself to Web display and navigation.

8.5.3 Because the content of each field of the documentation is tagged, the documentation can serve as the foundation for extract and analysis programs, search engines, and other software agents written to assist the research process.

8.5.4 Preparing documentation in DDI format at the outset of a project means that the documentation will also be suitable for archival deposit and preservation, because it will contain all of the information necessary to describe all of aspects of the corresponding data files. DDI XML should ideally be generated by the CAI system used to collect data, but can also be collected from paper and pencil surveys through access to the information in the original questionnaire.

8.5.5 There are many examples of projects that utilize DDI-compliant metadata, both at individual study level and multi-study data repositories (DDI, 2015). These studies illustrate the value of using these standards, such as:

- The presentation of instrument documentation, so that users can track the logic of the questionnaire.
- The creation of questions banks, comprising all items asked in multi-year studies, years items were asked, differences in question wording, and so on. XML marked up information gets its full potential when coupled with a database management system and powerful front end tools.
- The establishment of links to the documentation of related surveys (e.g., those conducted in other countries) with variable text viewable in the native languages, assists

analysts who want to study relationships among all of the survey items.

Lessons learned

8.1 Many 3MC studies provide extensive documentation online. Some examples include:
   8.1.1 The Demographic Health Survey provides their questionnaires and manuals via their website (http://www.dhsprogram.com/)
   8.1.2 The European Social Survey produces an annual survey documentation report, as well as a report summarizing field work and any deviations for each round. (ESS, 2015)

8.2 Even though the amount of documentation that 3MC studies provide has increased in recent years, there is still a need to provide users with more information about the entire survey life cycle, particularly through detailed quality profiles (see Survey Quality).

9. Consider disseminating research findings.

Rationale

Dissemination is more than storing (archiving) data. Presenting research findings in addition to making the data file available to other users, is an important step in quality dissemination practices. This section of the chapter discusses dissemination in terms of presenting results of the study, and considering who will use the information and why. This guideline is based on the guidelines written by the Community Advisory Board of the University of California, San Francisco, Center for AIDS Prevention Studies (CAB CAPS) and is adapted for the 3MC context (Fernández-Peña et al., 2008).

Procedural Steps

9.1 Create a dissemination plan
   9.1.1 Include presenting findings in the study's initial budget. This may include salary, translation, printing, mailing, and/or meeting costs (see Tenders, Bids, and Contracts and Translation: Budgeting).
   9.1.2 Create a team which will organize and create dissemination materials.
   9.1.3 Get input from study participants, community representatives, and other potentially interested parties on the preferred forum for viewing findings, such as press releases, websites, newsletters, or conferences. Consider offering multiple venues, if possible.

9.1.4 Remember that there may be a need to disseminate findings several times as new information is collected and updated.

9.2 Make research results accessible to the desired audience(s). Potential audiences and effective methods include:

9.2.1 Study participants:
- Ask participants if and how they would want to receive results. This can be incorporated as a question in the survey instrument.
- Create a newsletter for participants.
- Write any information disseminated in accessible language, and keep in mind the literacy and language needs of the study population.

9.2.2 Community members/Target populations:
- Consider multiple methods including articles in the media such as newspapers, radio, or TV news in order to reach many people.
- As with study participants, consider the language needs of the community.
- Explore how research results from cross-national surveys can be disseminated to as many participating countries as possible. Different dissemination strategies may need to be employed in different countries/cultures.

9.2.3 Agencies and Service Providers:
- Prioritize contacting agencies that aided with participant recruitment and/or serve the target population.
- Emphasize practical use of the study results.

9.2.4 Policymakers:
- Evaluate if research results have potential to impact policy.
- Send newsletters/articles or reports to local and national government representatives.

9.3 Consider the ethical and legal policies within each country and culture. Individual countries may have different dictates on sharing data within and between countries. See Ethical Considerations for further discussion.

Lessons Learned

9.1 Traditionally, researchers disseminate work in peer-reviewed journals. However, practitioners, as well as the general public, rarely have the time, or even ability, to read these types of articles. The CAB CAPS guidelines were created by a committee of activists, teachers, and other stakeholders. Committee members who had participated in research studies were concerned over the lack of accessible findings, and developed the above points in order to address dissemination needs. Making the attempt to disseminate results in this way provides more benefit to those who funded the research project and encourages discussion about the strengths and weaknesses of the original data.

9.2 The Afrobarometer Survey issues reports or bulletins within three months of the end of fieldwork. An advance briefing is offered to top policy makers in the executive and legislative branches of participating countries; immediately thereafter, results are released publicly to the national and international media, civil society, and donors. Releases must be approved by a core partner (Afrobarometer Survey, 2014). Similarly, data from the World Mental Health Survey is available to policy makers in participating countries (Kessler, Ustun, & World Health Organization, 2008).

10. Make quality control an integral part of all dissemination steps.

Rationale

The value of data depends on the quality of the data itself. Dissemination requires strict compliance to archiving, editing, publishing, and distribution protocols. Dissemination also requires the long-term availability of data and documentation files though constantly new versions of hardware, software, and possible changes in management and staff. Clear procedures must be in place to make certain all files are readable as statistical and word processing software systems change over time.

Procedural steps

10.1 Establish a quality compliance protocol: an overall plan for regularly monitoring the integrity and validity of all data and documentation files that are available for secondary use.

10.2 Consult with institutions, research associations and analysts to develop appropriate quality standards. Standards should be developed with researchers to ensure they meet the needs of the relevant discipline (The Organisation for Economic Co-Operation and Development, 2007).

10.3 Check all dissemination production steps throughout.

10.4 Test archived files periodically to verify user accessibility.

10.5 Establish procedures early in the survey lifecycle to insure that all important files are preserved.

10.6 Create digitized versions of all project materials, whenever feasible.

10.7 Develop specific procedures for assessing disclosure risk to respondents and execute these procedures whenever public-use files are produced.

10.8 Produce and implement procedures to distribute restricted-use files if applicable.

10.9 Provide data files in all the major statistical software packages and test all content thoroughly before they are made available for dissemination. In addition, provide data in a non-proprietary format so that users may utilize the statistical package of their choice.

10.10 Designate resources to provide user support and training for secondary researchers.

10.11 Discuss with users their experiences working with the data. This may include surveying users, conference presentations, and collecting user data.

Lessons learned

10.1 The Centers for Disease Control (CDC) in the United States, working with other federal agencies, did a study of web-based systems for the dissemination of health data and produced a *Guide for Public Health Agencies Developing, Adopting, or Purchasing Interactive Web-based Data Dissemination Systems.* The *Guide* was developed based on the experiences of many health agencies in disseminating their data and attempts to establish a set of general standards and practices. A checklist is provided to guide agencies in developing a comprehensive web dissemination system (Centers for Disease Control, 2010).

References

Afrobarometer Survey. (2014). *Round 6 Survey Manual*. Retrieved June 5, 2016, from http://www.afrobarometer.org/sites/default/files/survey_manuals/ab_r6_survey_manual_en.pdf

Alexander, W. E., Cady, J., & Gonzalez-Garcia, J. (2008). *The IMF's data dissemination initiative after 10 years*. Washington, DC: International Monetary Fund. Retrieved March 9, 2015, from http://www.imf.org/external/pubs/ft/books/2008/datadiss/dissemination.pdf

Center for Research Libraries. (2015). *Digital preservation metrics*. Retrieved March 9, 2015 from http://www.crl.edu/archiving-preservation/digital-archives/metrics

Data Documentation Initiative (DDI). (2015). Retrieved March 9, 2015, from http://www.ddialliance.org/

The Demographic and Health Surveys Program (DHS). (2015). *Survey process*. Retrieved June 14, 2016, from http://dhsprogram.com/What-We-Do/Survey-Process.cfm

European Social Survey. (2015). ESS5 – 2010 Documentation Report. Retrieved May 5, 2016, from http://www.europeansocialsurvey.org/docs/round5/survey/ESS5_data_documentation_report_e04_0.pdf

Fernández-Peña, J. R., Moore, L., Goldstein, E., DeCarlo, P., Grinstead, O., Hunt, C., et al. (2008). Making sure research is used: Community generated recommendations for disseminating research. Progress in Community Health Partnerships: Research, *Education, and Action, 2*(2), 171-176.

German National Science Foundation. (2013). *Safeguarding good scientific practice*. Retrieved Oct 14, 2015, from http://www.dfg.de/en/research_funding/principles_dfg_funding/good_scientific_practice/

Gutmann, M. P., Abrahamson, M., Adams, M. O., Altman, M., Arms C., Bollen. K., et al. (2009). From preserving the past to preserving the future: The Data-PASS project and the challenges of preserving digital social science data. *Library Trends, 57*(3). Retrieved March 9, 2015, from http://muse.jhu.edu/journals/library_trends/summary/v057/57.3.gutmann.html

Health Insurance Portability and Accountability Act Privacy and Security. (2015). *Examples of privacy violations*. Retrieved March 9, 2015, from http://www.hhs.gov/ocr/privacy/hipaa/enforcement/examples/allcases.html

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E. S., Spicer, K., & de Wolf, P. (2012). *Statistical disclosure control*. West Sussex UK: John Wiley & Sons.

Integrated Demographic and Health Series. (2015). Integrated demographic and health series version 2.0. Retrieved October 14, 2015, from https://www.idhsdata.org/idhs/

Integrated Social Survey Programme. (2015). ISSP Modules by Topic. Retrieved October 14, 2015, from http://www.gesis.org/issp/modules/issp-modules-by-topic/

Inter-university Consortium for Political and Social Research (ICPSR). (2012). Data sharing for demographic research. Retrieved March 9, 2015, from http://www.icpsr.umich.edu/icpsrweb/DSDR/

Inter-university Consortium for Political and Social Research (ICPSR). (2015). Data management & curation. Retrieved October 20, 2015, from http://www.icpsr.umich.edu/icpsrweb/content/datamanagement/confidentiality/

International Federation of Data Organizations for Social Science. (2014). Policies for sharing research data in social sciences and humanities. Retrieved March 17, 2015 from http://ifdo.org/wordpress/?page_id=309

International Monetary Fund. (2015). Dissemination standards bulletin board. Retrieved March 9, 2015 from http://dsbb.imf.org/Applications/web/dsbbhome/

International Social Survey Programme (ISSP). (2015). Retrieved March 9, 2015, from http://www.issp.org/

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

National Digital Archive of Datasets. (2010). Retrieved March 9, 2015, from http://www.ndad.nationalarchives.gov.uk/

National Human Research Protections Advisory Committee. (2002). Recommendations on confidentiality and research data protections. Retrieved October 20, 2015 from https://cyfar.org/sites/default/files/Confidentiality_0.pdf

National Institute of Health. (2003). NIH data sharing policy and implementation guidance. Retrieved March 9, 2015 from http://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm

The Organisation for Economic Co-Operation and Development. (2007). OECD principles and guidelines for access to research data from public funding. Retrieved March 17, 2015 from http://www.oecd.org/science/sci-tech/38500813.pdf

O'Rourke, J. M., Roehrig, S., Heeringa, S. G., Reed, B. G., Birdsall, W.C., Overcashier, M. (2006). Solving problems of disclosure risk while retaining key analytic uses of publicly released microdata. *Journal of Empirical Research on Human Research Ethics, 1*(3), 63-84.

Royal Statistical Society & the UK Data Archive. (2002). *Preserving & sharing statistical material.* Colchester, Essex: University of Essex. Retrieved March 9, 2015, from http://www.rss.org.uk/Images/PDF/publications/rss-reports-preserving-sharing-statistical-data-2002.pdf

Sicinski, A. (1970). "Don't Know" answers in cross-national surveys. *Public Opinion Quarterly, 34*(1), 126-129.

Tudor, C., Cornish, G., & Spicer, K. (2014). Intruder testing on the 2011 UK census: Providing practical evidence for disclosure protection. *Journal of Privacy and Confidentiality, 5*(2), 3.

Van Diessen, R., & Steenbergen, J. (2002). *Long term preservation study of the DNEP Project – An overview of the results*. Amsterdam: IBM Netherlands. Retrieved June 9, 2016, from http://ibrarian.net/navon/paper/IBM___KB___Long_Term_Preservation_Study.pdf?paperid=3165076

Further Reading

Allum, P., & Mehmut A. (2001). Economic data dissemination: What influences country performance on frequency and timeliness? (IMF Working Paper No. 01/173). Retrieved March 9, 2015 from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=880222

The Dataverse Network Project. (2015). Retrieved March 9, 2015, from http://dataverse.org/

Doyle, P., Lane, J. I., Theeuwes, J. J. M., & Zayatz, L. V. (Eds.) (2001). *Confidentiality, disclosure, and data access: Theory and practical applications for statistical agencies*. Amsterdam: Elsevier Science.

Duncan, G. T., & Lambert, D. (1986). Disclosure-limited data dissemination. *Journal of the American Statistical Association, 81*(393), 10-18.

Hoelter, L., Pienta, A., & Lyle, J. (2016). Data preservation, secondary analysis, replication: Learning from existing data. In C. Wolf, D. Joye, T. W. Smith, & Y. Fu (Eds.), *The SAGE handbook of survey methodology*. Thousand Oaks, CA: Sage.

Inter-university Consortium for Political and Social Research (ICPSR). (2009). The guide to social science data preparation and archiving: Best practice throughout the data life cycle (4th ed.).Retrieved March 9, 2015 from http://www.icpsr.umich.edu/files/ICPSR/access/dataprep.pdf

Joint UNECE/EUROSTAT Work Session on Statistical Data Confidentiality. (2013). Retrieved March 9, 2015, from http://www.unece.org/index.php?id=31958#/

United Nations. (1998). *Handbook on civil registration and vital statistics systems: Policies and protocols for the release and archiving of individual records* (Studies in Methods Series F, No. 70). New York, NY: Department of Economic and Social Affairs, United Nations Statistics Division. Retrieved March 9, 2015, from http://unstats.un.org/unsd/publication/SeriesF/SeriesF_70E.pdf

United Nations Statistics Division. (2015). Retrieved March 9, 2015, from http://unstats.un.org/unsd/default.htm

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*. Thousand Oaks, CA: Sage.

Other Resources

For an example presentation of study findings, watch

Lesclingand, M., & Hertrich V. (2007). *When the population is changing. A presentation of research findings in Mali*. Paris, INED (CD).

The CD-ROM is available free of charge. Contact M. Lesclingand (marie.leschingand@unice.fr) or U. Herfrich (hertrich@ined.fr).

Statistical Analysis

Mengyao Hu, 2016

Introduction

In recent years, the number and scope of multinational, multicultural, or multiregional surveys, which we refer to as “3MC” surveys, has increased dramatically. With the increased availability of large datasets covering multiple countries, such as the European Social Survey (ESS) and the Survey of Health, Ageing and Retirement in Europe (SHARE), more researchers have become engaged in analyzing these data (Davidov, Schmidt, & Billiet, 2011). Not surprisingly, there has been increased interest in the development of the statistical tests appropriate to cross-cultural survey data analysis. This chapter aims to provide a comprehensive introduction of different statistical methods, from basic statistics to advanced modeling approaches. Note that this chapter does not aim to teach statistics, but rather to provide an overview of what statistical tests are available and when to apply them in 3MC research. We also provide links and references to each statistical method for those who would like additional detail.

1. Exploring the variables

1.1 Types of variables

The classification of variable types is important because it will help to determine which statistical procedure should be used. For example, when the dependent variable is continuous, a linear regression can be applied (see Guideline 2.2); when it is categorical (binary), a logistic model can be applied (see Guideline 3.2); when it is categorical (nominal or ordinal), multinomial or ordinal logistic regressions may be used (see Guideline 3.3). If, in latent variable models (see Guideline 6), the latent variable is continuous, Confirmatory Factor Analysis (CFA) or Item Response Theory model (IRT) can be used (see Guidelines 6.1 and 6.5). Table 1 and Table 2 below list the choices of regression and latent variable measurement models, regarding the variable types of the dependent and independent variables.

Several commonly used variable types are listed as below:

- **Nominal variables:** Variable values assigned to different groups. For example, respondent gender may be “male” or “female”.
- **Ordinal variables:** Categorical variables with ordered categories. For example, “agree,” “neither agree nor disagree,” or “disagree”.

- **Continuous variables:** Variables which take on numerical values that measure something. "If a variable can take on any value between two specified values, it is called a continuous variable; otherwise, it is called a discrete variable"(Rovai, Baker, & Ponton, 2013). Continuous variables are understood to have equal intervals between each adjacent pair of values in the distribution. Income is an example of a continuous variable.
- **Discrete (ratio) variable:** "A discrete variable can only take on a finite value, typically reflected as a whole number" (Randolph & Myers, 2013). The variables have an absolute '0' value. One example is the number of children a person has.

Table 1. Variable type and choices of regression models

Dependent Variables	**Regression models**
Continuous	Linear Regression
Categorical (Binary)	Logistic Regression
Categorical (Ordinal)	Ordinal Logistic Regression
Categorical (Nominal)	Multinomial Logistic Regression

Table 2. Variable type and choices of latent variable measurement models (see Guideline 6 for more detail)

		Latent Variables	
		Continuous	**Categorical**
Indicators	**Continuous**	CFA; Exploratory Factor Analysis (EFA)	Latent profile model
	Categorical	CFA with categorical indicators; IRT	Latent class analysis

1.2 The distribution of variables

1.2.1 Graphical illustrations of distributions

It is commonly recommended to look at graphical summaries of both continuous and categorical distributions before fitting any models. Details of the graphical options listed below can be found here: http://onlinestatbook.com/2/graphing_distributions/graphing_distributions.html

- For categorical variables:
  - Bar graphs
- Pie charts

  For continuous or discrete variables:
  - Stem and Leaf Plots
  - Histograms
- Box plots

  For any type of variable:
  - Frequency distributions

In 3MC data analysis, to get a direct visual comparison, researchers can plot distributions by country or racial group.

1.2.2 Numerical summaries of distributions

A distribution can be summarized with various descriptive statistics. The mean and median capture the center of a distribution (central tendency) while the variance describes the distribution spread or variability (see http://onlinestatbook.com/2/summarizing_distributions/summarizing_distributions.html).

- **Mean**: the average of a number of values. It is calculated by adding up the values and dividing by the number of the values (how many the values there are).
- **Median**: The "…median is the number separating the higher half of a data sample, a population, or a probability distribution, from the lower half" (Reviews, 2013). For a highly skewed distribution, the median may be a more appropriate measure of central tendency than the mean. For example, the median is more widely used to characterize income, since potential outliers (e.g., those with very high incomes) have much more impact on the mean.
- **Variance**: **Variance** is a measure of the extent to which a set of numbers are "spread out".
- **Precision**: **Precision** is the reciprocal of the variance and is most commonly seen in Bayesian analysis (see Guideline 9).

1.3 Suggested reading

- Tests of the equality of two means: https://onlinecourses.science.psu.edu/stat414/node/224
- van de Vijver and Leung (1997)
- Braun and Johnson (2010)

1.4 Potential uses in 3MC research

A good starting point of an analysis is to look at the distributions of variables of interests and the graphical illustrations of the variables in each cultural group.

One way of comparing survey estimates across various cultures is to directly compare mean estimates. A two sample t-test can be used to evaluate the equality of two means (see Guideline1.3). However, researchers need to be aware that the observed mean differences are not necessarily equal to the latent construct mean differences (see Guideline 6) and direct comparison using observed mean differences may lead to invalid results (see Braun & Johnson, 2010). In addition, factors irrelevant to the question content, such as response style differences in different cultures, may influence the comparability across cultures. More advanced models (such as latent variable models) can be used to evaluate and control for these factors.

2. Simple and Multiple Linear Regression Models

2.1 Bivariate relationships

A bivariate relationship is the relationship between two variables. For example, one may be interested in knowing how height is associated with weight (i.e., whether those who are taller tend to weigh more). Basic information about bivariate relationships can be found here: http://onlinestatbook.com/2/describing_bivariate_data/bivariate.html

- **Scatterplots**
  Before running any models, a scatterplot is essential to explore the associations (negative or positive) between variables.
- **Correlations between variables**
  Pearson's correlation is the most commonly used method of evaluating the relationship between two variables. Refer to this website for more information:
  http://onlinestatbook.com/2/describing_bivariate_data/pearson.html

2.2 Linear regression models

Linear regression models can allow researchers to predict one variable using other variable(s). The dependent variable in linear regression models is a continuous variable. Basic information about simple linear and multiple regression models can be found here:

http://onlinestatbook.com/2/regression/regression.html

- ANOVA table
  In the output of regression model results, an Analysis of Variance (ANOVA) table is usually provided. It "consists of calculations that provide information about levels of variability within a regression model and form a basis for tests of significance" (Filler & DiGabriele, 2012).
  - https://onlinecourses.science.psu.edu/stat414/node/215
  - https://onlinecourses.science.psu.edu/stat414/node/221
  - An example of a regression model results output using Stata: http://www.princeton.edu/~otorres/Regression101.pdf
- Dummy predictor variables
  As described by Skrivanek (2009), a dummy variable or indicator variable is an artificial variable created to represent an attribute with two or more distinct categories/levels. If a categorical variable is added to the regression models directly, without being specially specified, the software will treat it as continuous. However, the differences between the categories (e.g., category 2 minus category 1) do not have an actual meaning. Dummy variables are usually created in this situation to make sure that such categorical variables are correctly specified in the model.

  For example, in 3MC data analysis, to compare country A to Country B on the level of the dependent variable, one can create a country dummy variable, using one of the countries as a reference group, and add it as an independent variable to the model. When multiple countries exist, one can use one of the countries as the reference category, and treat the variable as categorical in the model Piccinelli & Simon, 1997.
  - For information on dummy variables and how they are created and used, see Skrivanek, (2009).
  - For information on regression models with categorical predicators using SAS, see http://www.ats.ucla.edu/stat/sas/webbooks/reg/chapter3/sasreg3.htm
- Interactions of predictor variables
  Sometimes a regression model is used to test whether the relationship between the dependent variable (DV) and one specific independent variable (IV) depends on another IV. To test this, an interaction term between the two IVs can be added to the model.
  - http://www.jerrydallal.com/lhsp/reginter.htm
    - http://www.kenbenoit.net/courses/quant1/Quant1_Week10_interactions.pdf

- Transformations of variables
  When non-linearity is found for predictors, transformations may be considered to “normalize” a variable which has a skewed distribution. For more detail, see LaLonde (2005).
- Lack of fit testing
  Various techniques are available to test for the lack of fit in regression models, including visual (e.g., plots) and numerical methods (e.g., R^2 and F tests):
- Model diagnostics
  Techniques are available to test the appropriateness of the model and whether the model assumptions hold.
  - http://www.stat.columbia.edu/~yangfeng/W4315/lectures/lecture-10/lecture_10.pdf
  - Using R:
    http://web.stanford.edu/class/stats191/notebooks/Diagnostics%20for%20multiple%20regression.pdf
- Selecting reduced regression models (variable selection)
  Techniques for determining the model which contains the most appropriate independent variables, giving the maximum R^2 value.
  - http://www.ndsu.nodak.edu/ndsu/horsley/Stepwise%20regression.pdf (Including SAS code)
  - Steel & Uys (2007)

2.3 Suggested reading

- Applied Statistical Analysis and Data Display: An Intermediate Course with Examples in S-PLUS, R, and SAS (Heiberger & Holland, 2004)
- Statistical Methods, 8th ed. (Snedecor & Cochran, 1994)
- The Little SAS Book, 4th ed. (Delwiche & Slaughter, 2012)

2.4 Potential uses in 3MC research

- As in linear regression models, a country variable / indicator can be added to the regression model as a covariate (e.g., Piccinelli & Simon, 1997).

3. Categorical Data Analysis

3.1 Analysis of two-way tables

Categorical data are often displayed in a two-way table. Sometimes, one or both variables are continuous. If so, the continuous variable(s) can be categorized into groups. A two-way table can then be constructed using the new variables. Note that this approach may lead to a loss of information on the continuous variables.
http://www.stat.purdue.edu/~mhonerla/stat301/Chapter_9.pdf

3.1.1 Pearson chi-square
The Pearson chi-square test evaluates whether the row and column variables in a two-way table are associated.
- http://www.stat.yale.edu/Courses/1997-98/101/chisq.htm

3.1.2 Odds ratios (OR) and relative risks (RR)
OR and RR describe the proportions in contingency tables. See Sistrom and Garvan (2004) for a comprehensive introduction.
- http://ocw.jhsph.edu/courses/fundepiii/PDFs/Lecture16.pdf
- Schmidt and Kohlmann (2008)

3.1.3 Log-linear models
Log-linear models are commonly used to model the cell counts of contingency tables, such as two-way tables.
- http://www.biostat.umn.edu/~dipankar/bmtry711.11/lecture_22.pdf

3.2 Logistic regression
Logistic regression models can be used when the dependent variable is a binary categorical variable. The technique allows researchers to model or predict the probability an individual will fall into one specific category, given other independent variables. Logistic regression is a type of generalized linear model, where the logit function of selecting one category is expressed through a linear function of the predictors. Thus, as in other linear regression models, the predictors can include both continuous and categorical variables.
- McDonald (2009) or http://www.biostathandbook.com/simplelogistic.html
- http://www.stat.cmu.edu/~cshalizi/uADA/12/lectures/ch12.pdf

3.3 Multinomial and ordinal logistic regressions
When the DV is a nominal variable, a multinomial logistic regression model can be used. If the DV is an ordinal variable, an ordinal logistic regression can be used.
- http://www.kenbenoit.net/courses/ME104/ME104_Day8_CatOrd.pdf

3.4 Suggested reading
- Agresti (1990)
- Bishop, Fienberg, & Holland (2007)
- Feinberg (1978)

3.5 Potential uses in 3MC research
- To evaluate responses to a categorical variable across two different cultures, one can construct a two-way table using the categorical variable and the country indicator as the rows and

columns. A Pearson chi-square test can be used to evaluate whether the variable differs by cultures.
- As in logistic regression models, a country variable / indicator can be added to the logistic regression model as a covariate.

4. Multilevel Models

Multilevel models are usually used when there is a hierarchical structure, such as when sampling units are nested in geographical areas (e.g., cluster sampling) and when they are selected in longitudinal studies. Multilevel models are also known as hierarchical linear models, mixed models, random effects models, and variance components models. The Center for Multilevel modeling at the University of Bristol offers a free online course on multilevel modeling. See http://www.bristol.ac.uk/cmm/learning/online-course/ for more information. Additional information on multilevel modeling can be found at http://www.bristol.ac.uk/cmm/learning/multilevel-models/, and in van de Vijver, van Hemert and Poortinga (2008).

When many cultural groups are present, a multilevel model framework can be used, with country treated as a random variable. Multilevel models with latent variable can also be run, such as multilevel structural equation models (MLSEM), as discussed by Cheung (2006) and Fischer (2009). See Guideline 6.4 for more information on SEM models.

4.1 Suggested reading
- Bryan and Jenkins (2015)
- Gill and Womack (2013)
- Merlo, Chaix, Yang, Lynch, & Råstam (2005)
- West and Galecki (2011)

4.2 Potential uses in 3MC research
Many 3MC studies have a multilevel data structure with respondents nested within countries. Recent research on multilevel cross-cultural research has emerged in last several decades. For more information, see Van de Vijver, van Hemert, and Poortinga (2015).

5. Longitudinal analysis

Longitudinal data analysis refers to techniques used to evaluate data collected through repeated measures.
- https://onlinecourses.science.psu.edu/stat510/node/41

5.1 Modeling longitudinal / panel data
In panel surveys, respondents are interviewed at multiple points in time, producing “panel data” or “longitudinal data.” The first step in

analyzing longitudinal data is to look at the descriptive plots; then select one of several possible methods of analysis. The traditional technique is the repeated measures analysis of variance (rmANOVA), although this has several limitations. More commonly used approaches include multilevel models and marginal models.

5.1.1 Descriptive plots

The "spaghetti" plot "involves plotting a subject's values for the repeated outcome measure (vertical axis) versus time (horizontal axis) and connecting the dots chronologically" (Swihart, et al., 2010). Plots can be created at both the individual data level and the mean level. For binary outcomes, proportions can be used to generate the plot for different population groups. In 3MC studies, the plots can be generated for different cultural or country groups.

- http://www.phuse.eu/download.aspx?type=cms&docID=4672
- http://www.pharmasug.org/proceedings/2013/CC/PharmaSUG-2013-CC27.pdf

5.1.2 Repeated measures analysis of variance (rmANOVA)

- http://www.ats.ucla.edu/stat/sas/library/repeated_ut.htm
- The rmANOVA approach is not recommended due to the limitations as mentioned below:
- Subjects missing any data will not be included in the analysis.
- A limited number of covariance structures are allowed.
- Time-varying covariates are not allowed.

5.1.3 Multilevel models for longitudinal data

Multilevel models account for between respondent variance by including random effects in the model, such as random slope and random intercept.

- http://www.ats.ucla.edu/stat/seminars/mlm_longitudinal/
- Steele (2008) and presentation slides at http://www.bristol.ac.uk/media-library/sites/cmm/migrated/documents/longitudinal.pdf
- http://www.gllamm.org/handout_2012.pdf

5.1.4 Marginal modeling approaches

If the between subject variation is not of interest, the marginal modeling approach, where only the correlated error terms are included in the model, can be used – no random effects are added to the model.

- Welch (2009)
- http://www.ats.ucla.edu/stat/sas/examples/alda/

5.2 Suggested reading

- Chatfield (2013)
- Ballinger (2004)
- Steele (2008)
- Singer (1998)

- West (2009)
- Halekoh, Højsgaard, and Yan (2006)
- Kreuter and Muthen (2008)
- Twisk (2006)

5.3 Potential uses in 3MC research

- A country variable / indicator can be added to the marginal models as a covariate, or it can be added in a multilevel model as a fixed effect.

6. Latent Variable Models

Latent variable models include both observed variables (the data) and latent variables. A latent variable is unobserved, which represents hypothetical constructs or factors (Kline, 2011). A latent variable can be measured by several observed variables. An example of latent variable provided by Kline (2011) describes construct of intelligence. As mentioned by Kline (2011), "there is no single, definitive measure of intelligence. Instead, researchers use different types of observed variables, such as tasks of verbal reasoning or memory capacity, to assess various facets of intelligence." Examples of such latent variables are usually measured in a measurement model, which evaluates the relationship between latent variables and their indicators.

6.1 Exploratory Factor Analysis and Confirmatory Factor Analysis

Exploratory Factor Analysis (EFA) and Confirmatory Factor Analysis (CFA) are two types of measurement models, where latent variables are indicated by multiple observed variables. The difference between EFA and CFA is related whether you have a hypothesis about the measurement model before doing the analysis. As mentioned by Yong and Pearce (2013), "CFA attempts to confirm hypotheses and uses path analysis diagrams to represent variables and factors, whereas EFA tries to uncover complex patterns by exploring the dataset and testing predictions". As seen in Table 3, since both the indicators and the latent variables are all continuous, both EFA and CFA are based on linear functions. Table 3 shows the differences between EFA and CFA. Since EFA is purely data-driven which may be arbitrary in nature, it is thus suggested by some literature to always use CFA, which is theory-driven, rather than EFA (Sansone, Morf, & Panter, 2004). As mentioned by Sansone et al. (2004), in selecting items, it is more appropriate to use EFA rather than CFA, when the theory is not well established.

See Yong and Pearce (2013) for a comprehensive overview on EFA and Brown (2015) for CFA. The code for conducting EFA and CFA are included in the Appendix A.

Table 3. Comparisons between EFA and CFA

EFA	CFA
Purpose: To identify latent factors that account for variance and covariance among a set of observed variables (both based on common factor model)	
Descriptive / exploratory procedure	Requires strong empirical or conceptual foundation
Input: correlation matrix (all variables standardized)	Input: variance-covariance matrix (standardized and unstandardized solution)
Factor selection based on eigenvalue procedures and model fit statistics	Prespecification of number of factors pattern of factor loadings
Factor rotation to obtain simple structure	Simple structure is achieved by fixing (most) indicator cross-loadings to zero
Unique variances / measurement error uncorrelated	Unique variances / measurement error can be modelled
Overall, CFA offers more parsimonious solutions and greater modelling flexibility than EFA	

Adapted from: http://www.psychometrics.cam.ac.uk/uploads/documents/JumpStartFeb10/efacfajumpstart.pdf

Multi-group CFA (MCFA) is commonly used in 3MC research for measurement equivalence testing. The basic idea is to start with the same model but allow the coefficients differ by groups (assuming configural equivalence), and then start introducing constrains in the model coefficients – such as to make them equal across the groups. Then, the model fit of the previously run models can be compared. Among all the models, the parsimonious model with a good fit solution will be selected to evaluate the data. If the model reveals no violations of scalar equivalence, the country means can be compared directly. In a panel study, with data available at different time points, one can also evaluate measurement equivalence across cultures over time. See Guideline 6.2 below for more information on measurement equivalence testing.

For more information of MCFA, see:

- http://www.unc.edu/~rcm/psy236/measinv.pdf
- Steinmetz, Schmidt, Tina-Booh, Wieczorek, and Schwartz (2008)
- Brown (2015)

6.2 Measurement equivalence in 3MC research

As mentioned by Kankaraš and Moors (2010), “measurement equivalence implies that a same measurement instrument used in different cultures measures the same construct." There are different levels of equivalence. Three most widely discussed levels are: configural, metric and scalar equivalence. These three levels are hierarchical, where the higher ones have higher requirements of equivalence, and require the achievement of the lower ones (Kankaraš & Moors, 2010).

Configural equivalence refers to similar construction of the latent variable. In other words, same indicators are associated with the latent concepts in each culture. It does not require each culture view the concept in the same way. For example, it allows the strength (i.e., loadings) to be different across cultures. ***Metric equivalence*** requires same slope across cultures which capture the associations between indicator and the latent variable. In other words, it implies "the equality of the measurement units or intervals of the scale on which the latent concept is measured across cultural groups" (Kankaraš & Moors, 2010; Steenkamp & Baumgartner, 1998). ***Scalar equivalence*** implies that on the basis of equality of the measurement units, the scales of the latent variable also have the same origin across cultures (Kankaraš & Moors, 2010). Under this equivalence level, the model achieves full measurement equivalence, and researchers can compare the country scores (i.e., country scores) directly.

In situations where full equivalence is difficult to achieve, researchers also evaluate the conditions under which different cultures achieve partial equivalence. An example of partial equivalence is when most of the indicators are equivalent across cultures, but only one has a different slope and thresholds across cultures. One can then conclude that the different cultures achieve partial equivalence, where they differ on one specific indicator. As mentioned by Kankaraš and Moors (2010), "partial equivalence enables a researcher to control for a limited number of violations of the equivalence requirements and to proceed with substantive analysis of cross-cultural data" (Kankaraš & Moors, 2010; Steenkamp & Baumgartner, 1998).

The aforementioned approaches to assessing measurement equivalence have been widely used in 3MC survey analysis. However, it has recently been criticized for being overly strict. As mentioned by Davidov et al. (2015), it is difficult to achieve scalar equivalence or even metric equivalence in surveys with many countries or cultural groups. A Bayesian approximate equivalence testing approach has been recently proposed by Davidov et al. (2015). This approach allows "small variations" in parameters across different cultural groups (Davidov et al., 2015). Thus, when approximate scalar measurement equivalence is reached, one can compare across cultures meaningfully, even though the traditional method may indicate scalar inequivalence (Davidov et al., 2015). For introductions and references of Bayesian methods, see Guideline 9.

6.3 Latent Class Analysis (LCA)

Unlike the previously mentioned approach, such as CFA and SEM, where the latent variables are continuous, LCA treats the latent variables as categorical – nominal or ordinal (see Table 3). The categories of the latent variable in LCA are referred to as classes, which represent “a mixture of subpopulations where membership is not known but is inferred from the data” (Kline, 2011). That is to say, LCA can classify respondents into different groups based on their attitudes or behaviors, such as classifying respondents by their drinking behavior. Respondents in the same group are similar to each other, regarding the behavior / attitudes, and they differ from those in other groups – i.e., heavy drinkers vs. nonalcoholic drinkers. One can also add covariates in the model if those measures can influence the class membership. In a second-step, the class membership from the model can then be used for follow-up analysis. For example, to better understand the differences between respondents, a logistic (or multinomial logistic, if more than two groups) regression model can be run in which selected covariates are used to predict the class membership. Or, to evaluate the influence of the class membership on other variables, LCA can also be used in regression models as a covariate to predict other outcomes.

For more information on LCA, please see:

- http://www.restore.ac.uk/latentvariablemodels/workshopfiles/Latent%20Class%20Analaysis_Orla%20McBride/Course%20Presentation/Powerpoint%20slides%20-%20LCA%20workshop.pdf
- McCutcheon (1987)

As mentioned by Kankaraš, Moors, and Vermunt (2010), when testing for measurement invariance with latent class analysis, “the model selection procedure usually starts by determining the required number of latent classes or discrete latent factors for each group. … If the number of classes is the same across groups, then the heterogeneous model is fitted to the data; followed by a series of nested, restricted models which are evaluated in terms of model fit”. That is to say, unlike multi-group CFA, the multi-group LCA will need to identify if the number of classes are the same across groups, before testing for models at different invariance levels. See

Kankaraš, Moors, and Vermunt (2010), Eid, Langeheine, and Diener (2003), and Kankaraš and Moors (2009) for more information.

6.4 Structural Equation Modeling (SEM)

Structural Equation Modeling is a multivariate analysis technique used in many disciplines, which aims to test the causal relationship hypothesis between variables (Holbert & Stephenson, 2008). It usually includes two components: 1) the measurement model which summarizes several observed variables using their latent construct (e.g., CFA as discussed in 7.2) and the structural model, which describes the relationship between multiple constructs (e.g., relationships among both latent and observed variables).

6.4.1 Variables

Similar to previously discussed latent variable models, SEM can have both observed and latent variables, where observed variables are the data collected from respondents and latent variables represent unobserved construct and factors (Kline, 2011). The observed variables which are used as measures of a construct are indicators of the latent variable. In other words, the latent variable is indicated by these observed variables.

Besides observed and latent variables, SEM models also include error terms, similar to the error terms in a regression analysis. As mentioned by Kline (2011), "a residual term represents variance unexplained by the factor that the corresponding indicator is supposed to measure. Part of this unexplained variance is due to random measurement error, or score unreliability".

6.4.2 Analysis of Covariance Structure

In SEM analysis, the parameter estimation is done by comparing the model-based covariance matrix with the data-based covariance matrix. The goal of this approach is to evaluate whether the model with best fit is supported by the data—that is, whether the two covariance matrixes are consistent with each other and whether the model can explain as much of the variance of the data.

6.4.3 Means of latent variables

Structural equation models can also estimate the means of latent variables. It also allows researchers to analyze between and within-group mean differences (Kline, 2011). In 3MC analysis, one can estimate the group mean differences on latent variables, such as between two cultures.

6.4.4 Suggested Reading

- Kline (2011)
- Bollen (1989)

- Hoyle (2012)

6.5 Item Response Theory (IRT)

IRT is commonly used for psychometric and educational testing, and is becoming more popular in 3MC analysis. It begins with the idea that when answering a specific question, the response provided by an individual depends on the ability / qualities of the individual and the qualities of the question item. As mentioned by Ostini and Nering (2005), "The mathematical foundation of IRT is a function that relates the probability of a person responding to an item in a specific manner to the standing of that person on the trait that the item is measuring. In other words, the function describes, in probabilistic terms, how a person with a higher standing on a trait (i.e., more of the trait) is likely to provide a response in a different response category to a person with a low standing on the trait." Therefore, IRT allows researchers to model the probability of a specific response to a question item, given the item and the individual's trait level.

The simplest IRT model is the Rasch model, also called one-parameter IRT model, which assumes equal item discrimination ("the extent to which the item is able to distinguish between individuals on the latent construct" (Chan, 2000)). This model starts from the premise that the probability of giving a "positive" answer to a yes/no question is "a logistic function of the distance between the item's location, also referred to as item difficulty, and the person's location on the construct being measured", also known as the person's latent trait level (Mneimneh, Heeringa, Tourangeau & Elliott, 2014). There are other types of IRT models available. They can be categorized by the number of parameters and the question response option format, such as binary or multiple response options and whether ordinal or nominal. Table 4 below summarizes different types of IRT models.

In a two-parameter (2PL) IRT model, an item discrimination parameter is also included in the model. The parameters are "analogous" to the factor loadings in CFA and EFA, since they all represent "the relationship between the latent trait and item responses" (Brown, 2015). A three-parameter (3PL) IRT model also includes a "guessing" parameter. It describes the situation that when a question can be answered by guessing, the probability of giving a correct answer is higher than zero even for those with low latent trait level.

For items with multiple response options (ordinal or nominal variables), polytomous IRT models can be used. See Table 4 for details. In this chapter, we will not discuss these models in detail. See suggested readings on IRT models for more information.

Table 4. Variable type and IRT model choices.

Type of observed variable	**Model**
Binary	1 parameter-logistic model (1 - PL model) / Rasch model
	2 - PL model
	3 - PL model
Multiple response options (Ordinal)	Graded Response model / Thurstone/Samejima polytomous models
	Partial Credit model (PCM) & Graded PCM
Multiple response options (Nominal)	Rating Scale model
	Nominal response model / Bock's model

6.5.1 Suggested Reading

- Ostini and Nering (2005)
- Davidov, Schmidt, and Billiet (2011)
- Hambleton, Swaminathan, and Rogers (1991)
- Van der Linden & Hambleton (2013)
- Nering & Ostini (2011)
- Mneimneh, Heeringa, Tourangeau and Elliott (2014)

6.6 Other types of latent variable models

Besides what is discussed above, other types of latent variable models are available. Some examples are listed below.

6.6.1 Latent Transition Model

Latent Transition Model is "a special kind of latent class factor model that represents the shift from one of two different states, such as from nonmastery to mastery of a skill, is a latent transition model" (Kline, 2011). See http://r2ed.unl.edu/presentations/2012/SRM/033012_RyooWu/033012_RyooWu.pdf for more information.

6.6.2 Latent Profile Model

In latent profile models, the latent variable is categorical and the indicators are continuous. It is commonly used for cluster analysis. See Vermunt (2004) for more information.

6.6.3 Mixed Rasch Model

Mixed Rasch model is “a combination of the polytomous Rasch model with latent class analysis” (Quandt, 2011). See Quandt (2011) for more information.

6.6.4 Multilevel Structural Equation Modeling (MLSEM)

When we have data where the population of individuals are divided into different groups, such as in a 3MC context, a Multilevel Structural Equation Modeling (MLSEM) can be used. This model decomposes individual data into within group and between group components, and can simultaneously estimate of within and between group models (Muthén and Muthén, 2007). For more information on MLSEM, see Rabe-Hesketh, Skrondal, & Zheng (2007).

6.7 Potential uses in 3MC research

As mentioned by Steinmetz (2011), the observed mean does not equal to the latent mean, where the observed mean is a function of item intercepts, factor loadings and the latent mean. Similarly, “observed mean differences between two or more groups (e.g., cultures) do not necessarily indicate latent mean differences as unequal intercepts and/or factor loadings will also lead to observed differences” (Steinmetz, 2011). To conduct more valid comparisons across different groups (e.g., cultures), measurement invariance testing is a widely used method, which aims to evaluate whether the latent means of various groups are comparable. In other words, it evaluates whether the different groups differ in factor loadings and intercepts of the measures. See Steinmetz (2011) for more information.

Measurement invariance testing is usually conducted within the multigroup analysis (MGA) framework. The most commonly used is multigroup confirmatory factor analysis (MGCFA) (e.g., Steinmetz, 2011). Other types of MGA include multigroup structural equation modeling (MGSEM) analysis (e.g., Meuleman & Billiet, 2011), multigroup latent class analysis (e.g., Kankaraš, Vermunt, & Moors, 2011), multigroup IRT model (e.g., Janssen, 2011) and multigroup mixed rasch model (e.g., Quandt, 2011). See Davidov et al. (2011) for more information.

A recent paper by Welzel and Inglehart discusss the misconceptions in measurement equivalence analysis. Using data from World Value Survey, they show that “constructs can entirely lack convergence at the individual level and nevertheless exhibit powerful and important linkages at the aggregate level” (Welzel & Inglehart, 2016).

7. Differential Item Functioning (DIF)

Differential Item Functioning (DIF) is a statistical concept developed to identify to what extent the question item might be measuring different properties for individuals of separate groups, such as ethnicity, culture, region, language, sex or other demographic groups. It can be used as indicators for "item bias" if the items function in a systematic different way across cultures. To detect DIF, different methods can be used, as listed below:

7.1 Mantel Haenszel.

Mantel-Haenszel (MH) statistic is regarded as a "reference" technique of detecting DIF due to its ease of use and the fact that it can be applied to small samples (Padilla, Hidalgo, Benítez, & Gómez-Benito, 2012). The disadvantage of MH statistic is that it does not allow statistical significance - testing.

7.2 Logistic regression.

Logistic regression can be used as an alternative method to detect DIF. For more information, see Clauser, Nungester, Mazor, & Ripkey (1996).

7.3 Techniques based on IRT models.

DIF can be detected using an IRT framework. Item characteristic curves (ICCs) of the same item but from different groups can be compared to evaluate whether there is DIF. For more information, see Thissen, Steinberg, & Wainer (1993) and Zumbo (2007).

8. Machine learning

Machine learning is "a general term for a diverse number of classification and prediction algorithms" (Lee, Lessler, & Stuart, 2010) which has applications in many different fields. Unlike statistical modeling approaches, machine learning evaluates the relationship between outcome variable and predictors using a "learning algorithm without an a priori model" (Lee et al., 2010). Below we introduce several machine learning methods.

8.1 Classification tree

The Classification tree is a data-driven method which allows researchers to evaluate the complex interaction between variables when there are many predictor variables present. In binary trees, the nodes of the tree are divided into two branches. To reasonably construct and prune a given tree, deviance measure is used to choose the splits. In R, "rpart" package is used for classification tree analysis (see Appendix A for more information). The classification tree result can be evaluated through apparent error rate and true error rate. The former one is the error rate when the tree is applied to a training data set, and the latter one is when it is applied to a new data set or a test data. In evaluating the true error rate, researchers usually divide the data into two parts: training data and test data, and validate the tree based on the test data set.

8.2 Random forest
Random forest is an algorithm for classification which uses an "ensemble" of classification trees (Díaz-Uriarte & De Andres, 2006). Through averaging over a large ensemble of "low-bias, high-variance but low correlation trees", the algorithm yields an ensemble that can achieve both "low bias and low variance" (Díaz-Uriarte & De Andres, 2006).

8.3 Suggested reading
- Ledolter (2013)
- Lemon, Roy, Clark, Friedmann, and Rakowski (2003)
- Lewis and Street (2000)
- Loh (2014)
- http://www-stat.wharton.upenn.edu/~stine/mich/DM_07.pdf
- http://www.ams.org/samplings/feature-column/fc-2014-12

8.4. Potential use in 3MC research
Classification tree analysis in cross-cultural research allows researchers to evaluate 1) the important factors for each culture, and 2) how the factor interactions differ across cultures. One study used classification tree to evaluate college student alcohol consumptions across American and Greek students, and found that "student attitudes toward drinking were important in the classification of American and Greek drinkers" (Kitsantas, Kitsantas, & Anagnostopoulou, 2008).

9. Incorporate complex survey data features

It is usually difficult to draw a simple random sample from the population, due to cost and practical considerations such as no comprehensive sampling frame available. As discussed in Sample Design, complex samples, such as surveys involving stratified / cluster sample design, are commonly used in surveys. In a simple random sample, one can assume that observations are independent from each other. However, in a complex sample design, such as multi-stage samples of schools, classes and students, students from one classroom are likely to be more correlated than those from another classroom. Therefore, as described in Sample Design, in the analysis phase, we need to compensate for complex survey designs with features including, but not limited to, unequal likelihoods of selection, differences in response rates across key subgroups, and deviations from distributions on critical variables found in the target population from external sources, such as a national Census, most commonly through the development of survey weights for statistical adjustment. If complex sample designs are implemented in data collection but the analysis assumes simple random sampling, the variances of the survey estimates can be underestimated and the confidence interval and test statistics are likely to be biased (Heeringa, West, & Berglund, 2010).

In a recent meta-analysis of 150 sampled research papers analyzing several surveys with complex sampling designs, it is found that analytic errors caused by ignorance or incorrect use of the complex sample design features were frequent. Such analytic errors define an important component of the larger total survey error framework, produce misleading descriptions of populations and ultimately yield misleading inferences (Aurelien, West, & Sakshaug, 2016). It is thus of critical importance to incorporate the complex survey design features in statistical analysis. For many of the aforementioned statistical models, various statistical software programs have enabled the analysis of complex survey data features, such as "svy" statement in Stata, and SURVEY procedures in SAS. See Appendix A for more information.

9.1 Suggested Reading:
- Heeringa, West, and Berglund (2010)
- Carle (2009)
- Rabe-Hesketh and Skrondal (2006)
- Stapleton (2006)
- Valliant, Dever, and Kreuter (2013)

10. Introduction to Bayesian Inference

This section presents an overview of the Bayesian Theory, which follows closely the overview of Lee (2012), Barendse, Albers, Oort, & Timmerman (2014), and Kaplan & Depaoli (2013). In surveys, respondents' answers, denoted as y, reflects our measure of the true population's Y – a random variable takes on a realized value y. In other words, Y is unobersever, and the probability distribution Y is of researchers' interests. We use θ to denote a parameter which reflects the characteristics of the distribution of Y. For example, θ can be the mean of the distribution. The goal is to estimate the unknown parameter θ based on the data, which is $p(\theta|y)$. Based on Bayes' theorem,

$$p(\theta|y) = \frac{p(\theta, y)}{p(y)} = \frac{p(y|\theta)p(\theta)}{p(y)}$$

where $p(y)$ is the probability distribution of the data, which is known for researchers, $p(y|\theta)$ refers to the probability of the data given the unknown parameter θ, and $p(\theta)$ is the *prior distribution* of the parameters. $p(\theta|y)$ is thus referred to as the *posterior distribution* of the parameter θ given the data, which is also the results of the model.

In summary, Bayesian methods use both the prior information (which indicates the distribution of parameters) and the distribution of data to estimate the model results – the posterior distributions of the parameters. The key difference between Bayesian and frequentist approach is relates with the unknown parameter θ. In frequentist approach, θ is viewed as unknown but fixed. On the other hand, in Bayesian approach, θ is random, which has a posterior distribution taking into account the uncertainty of θ.

10.1 Priors

There are generally two types of priors, noninformative and informative priors. The choice between the two types depends on our confidence about how much information we have about the priors and how accurate they are. Noninformative priors are also referred to as "vague" or "diffuse" priors. It is used when there is little information about the priors, and thus its influence on the posterior distribution of θ is minimal (Lee, 2012). An example of a noninformative prior can be a density with a huge variance, so that the Bayeisan estimation is mainly affected by the data. Informative priors are used when we have sufficient information about the priors, such as from knowledge of experts and similar data set.

10.2 Bayesian model comparison

There are multiple Bayesian model comparison statistics. Two most commonly used are Bayes factor and DIC. The Bayes factor quantifies the odds that the data favor one hypothesis over another. As discussed in Guideline 5, Bayes factors are not well defined when using noninformative priors (Berg, Meyer, & Yu, 2004), and the evaluations can be computationally difficult (Lee, 2012). DIC compromises both goodness of fit and model complexity. In practical applications, the model with the smaller DIC value is preferred.

10.3 Credible interval

When we have estimated the posterior distributions of the parameters, we would like summaries of the distribution, such as mean and variance, for hypothesis testing. One important way to evaluate the distribution is based on the credible interval, which is often referred to as a similar measure as the "confidence interval" in frequentist approach. Credible interval is based the quantiles of the posterior distributions. Based on the quantiles, we can directly evaluate the probability that the parameter lies in a particular interval. When this probability is 0.95, it is referred to as 95% credible interval. If the credible intervals from two models do not overlap, we say that the two posterior distributions of this parameter differ.

10.4 Markov Chain Monte Carlo (MCMC) Methods

MCMC is the most common computational algorithms for Bayesian methods. It generates Markov Chains, which simulate the posterior distribution. The basic idea is that by simulating a sufficiently large number of observations from the posterior distribution, $p(\theta|y)$, we can approximate the mean and other summary statistics of the distribution. The use of MCMC for posterior simulation in latent variable models is to treat the latent variables as missing data, which enables the augmentation of the observed variables. The most common MCMC algorithm is the Gibbs sampler, which performs on alternating conditional sampling at each of its iteration. More specifically, it draws each component conditional on the values of all the other components (Lee, 2012). In a Markov Chain, early proportion of the chain which may not converge to target distribution is called burn-in.

10.5 Convergence diagnostics

Multiple convergence diagnostics exist. In practice, it is common to inspect several different diagnostics, since there is no single adequate assessment. One of the most common statistics in a multiple-chain condition is the Gelman and Rubin diagnostic (Gelman & Rubin, 1992), which compares the within-chain and between-chain variance. A value above 1.1 is an indication of lack of convergence. The common diagnostics for single chain condition include the Geweke (1992) convergence diagnostic, and the Raftery and Lewis (1992) convergence diagnostic, which can help to decide how many iterations needed, and how many can be treated as burn-in in a long-enough chain.

10.6 Suggested reading

- Davidov et al. (2015)
- Lee and Song (2012)
- Fox (2010)
- Stone and Zhu (2015)
- Muthén and Asparouhov (2012)

10.7 Potential uses in 3MC research

As previously mentioned in Guideline 6, approximate Bayesian measurement equivalence approach can be used for cross-cultural comparison research (e.g., Davidov et al., 2015; Bolt, Lu, & Kim, 2014). See suggested readings in Guideline 9.6 above for more information.

Appendix A. Code resources for different types of analyses using different software packages

	SAS code	R code	Stata	Mplus	LISREL
Descriptive Analysis	▪ http://www.ats.ucla.edu/stat/sas/modules/descript.htm	▪ http://www.statmethods.net/stats/descriptives.html ▪ http://www.statmethods.net/stats/frequencies.html	▪ http://www.ats.ucla.edu/stat/stata/modules/descript.htm		
Linear Regression	▪ http://www.ats.ucla.edu/stat/sas/webbooks/reg/chapter1/sasreg1.htm ▪ http://www.ats.ucla.edu/stat/sas/webbooks/reg/chapter2/sasreg2.htm	▪ http://www.statmethods.net/stats/regression.html	▪ http://www.ats.ucla.edu/stat/stata/webbooks/reg/chapter1/statareg1.htm		
Log-Linear Models	▪ http://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug_genmod_sect059.htm	▪ http://ww2.coastal.edu/kingw/statistics/R-tutorials/logilin.html	▪ http://www.ats.ucla.edu/stat/stata/examples/icda/icdast6.htm		
Logistic regression	▪ http://www.ats.ucla.edu/st	▪ http://www.ats.ucla.edu/st	▪ http://www.ats.ucla.edu/st	▪ http://www.ats.ucla.edu/st	

	SAS code	R code	Stata	Mplus	LISREL
	at/sas/seminars/sas_logistic/logistic1.htm	at/r/dae/logit.htm	at/stata/dae/logit.htm	at/mplus/dae/logit.htm	
Ordinal Logistic Regression	▪ http://www.ats.ucla.edu/stat/sas/dae/ologit.htm	▪ http://www.ats.ucla.edu/stat/r/dae/ologit.htm	▪ http://www.ats.ucla.edu/stat/stata/dae/ologit.htm	▪ http://www.ats.ucla.edu/stat/mplus/dae/ologit.htm	
Multinomial Logistic Regression	▪ http://www.ats.ucla.edu/stat/sas/dae/mlogit.htm	▪ http://www.ats.ucla.edu/stat/r/dae/mlogit.htm	▪ http://www.ats.ucla.edu/stat/stata/dae/mlogit.htm	▪ http://www.ats.ucla.edu/stat/mplus/dae/mlogit.htm	
Longitudinal data analysis	▪ http://www.ats.ucla.edu/stat/sas/examples/alda/				
Multilevel modeling	▪ http://www.ats.ucla.edu/stat/examples/imm/				
Classification Tree	▪ Chapter 4 in http://support.sas.com/documentation/onlinedoc/miner/getstarted.pdf	▪ http://www.statmethods.net/advstats/cart.html	▪ http://statweb.stanford.edu/~lpekelis/talks/13_datafest_cart_talk.pdf		
	(Al Ghoson, 2010)				
EFA	▪ http://www2.sas.com/proceedings/sugi31/200-31.pdf ▪ http://www.ats.ucla.edu/st	▪ http://www.statmethods.net/advstats/factor.html	▪ http://dss.princeton.edu/training/Factor.pdf ▪ http://www.ats.ucla.edu/stat/stata/outp	▪ http://www.ats.ucla.edu/stat/mplus/seminars/introMplus_part1/efa_52.htm	▪ Page 8 on http://www.ssicentral.com/lisrel/techdocs/Session1.pdf

	SAS code	R code	Stata	Mplus	LISREL
	at/sas/library/factor_ut.htm		ut/fa_output.htm		
CFA	▪ http://www.salaswright.com/wp-content/uploads/2012/05/CFA.SEM_using_Stata12.0.pdf	▪ http://www.statpower.net/Content/312/Handout/Confirmatory%20Factor%20Analysis%20with%20R.pdf	▪ http://www.salaswright.com/wp-content/uploads/2012/05/CFA.SEM_using_Stata12.0.pdf	▪ http://www.iu.edu/~statmath/stat/all/cfa/cfa2008.pdf ▪ https://www.statmodel.com/usersguide/chapter5.shtml	▪ http://www.unc.edu/~rcm/psy236/lisrel.intro.pdf
MCFA		▪ http://pareonline.net/getvn.asp?v=19&n=7			▪ http://www.unc.edu/~rcm/psy236/measinv.pdf
SEM	▪ http://support.sas.com/documentation/cdl/en/statugstructequmodel/61765/PDF/default/statugstructequmodel.pdf ▪ http://support.sas.com/rnd/app/stat/papers/JSM2010_Yung.pdf	▪ http://www.personality-project.org/r/r.sem.html ▪ (Fox, 2006)	▪ http://www.stata.com/manuals13/sem.pdf	▪ http://www.ats.ucla.edu/stat/mplus/seminars/introMplus_part2/sem.htm ▪ https://www.statmodel.com/SEM.shtml	▪ http://www.ucdenver.edu/academics/colleges/nursing/Documents/PDF/HowToUseLISREL.pdf ▪ http://www.ssicentral.com/lisrel/complexdocs/chapter5_web.pdf

	SAS code	R code	Stata	Mplus	LISREL
LCA	▪ (Lanza, Collins, Lemmon, & Schafer, 2007)	▪ (Linzer & Lewis, 2001)	▪ http://www.stata.com/meeting/2nasug/lclass.pdf	▪ http://www.ats.ucla.edu/stat/mplus/seminars/introMplus_part2/lca.htm ▪ http://www.ats.ucla.edu/stat/mplus/seminars/IntroMplus/lca.htm	
Complex survey data	▪ http://www2.sas.com/proceedings/sugi27/p263-27.pdf	▪ http://www.jstatsoft.org/v09/i08/paper	▪ http://www.stata.com/meeting/mexico10/mex10sug_canette.pdf		

See also: Data Analysis Examples. UCLA: Statistical Consulting Group, from http://www.ats.ucla.edu/stat/dae/ (accessed August 3, 2015).

References

Agresti, A. (1990). *Categorical data analysis*. Hoboken, NJ: John Wiley & Sons.

Al Ghoson, A. M. (2010). Decision tree induction & clustering techniques in SAS Enterprise. *International Journal of Management and Information Systems, 14*(3), 57–70.

Aurelien, G. A. S., West, B. T., & Sakshaug, J. W. (2016). *How big of a problem is analytic error in secondary analyses of survey data?* The 71th Annual Conference of the American Association for Public Opinion Research, Austin, Texas.

Ballinger, G. A. (2004). Using generalized estimating equations for longitudinal data analysis. *Organizational Research Methods, 7*(2), 127-150.

Barendse, M. T., Albers, C. J., Oort, F. J., & Timmerman, M. E. (2014). Measurement bias detection through Bayesian factor analysis. *Frontiers in Psychology, 5*(September), 1–9.

Berg, A., Meyer, R., & Yu, J. (2004). Deviance information criterion for comparing stochastic volatility models. *Journal of Business & Economic Statistics, 22*(1), 107–120.

Bishop, Y. M., Fienberg, S. E., & Holland, P. W. (2007). *Discrete multivariate analysis: Theory and practice*. New York, NY: Springer Science & Business Media.

Bollen, K. A. (1989). *Structural equations with latent variables*. Wiley series in probability and mathematical statistics. Applied probability and statistics section. Oxford, England: John Wiley & Sons.

Bolt, D. M., Lu, Y., & Kim, J. S. (2014). Measurement and control of response styles using anchoring vignettes: A model-based approach. *Psychological Methods, 19*(4), 528.

Braun, M., & Johnson, T. P. (2010). An illustrative review of techniques for detecting inequivalences. In J. A. Harkness, M. Braun, B. Edwards, T. P. Johnson, L. E. Lyberg, P. Ph. Mohler …T. W. Smith (Eds.), *Survey methods in multinational, multiregional, and multicultural contexts* (pp. 373-393). Hoboken, NJ: John Wiley & Sons.

Brown, T. A. (2015). *Confirmatory factor analysis for applied research*. New York, NY: Guilford Press.

Bryan, M. L., & Jenkins, S. P. (2015). Multilevel modelling of country effects: A cautionary tale. *European Sociological Review*, jcv059.

Carle, A. C. (2009). Fitting multilevel models in complex survey data with design weights: Recommendations. *BMC Medical Research Methodology, 9*(1), 1.

Chan, D. (2000). Detection of differential item functioning on the Kirton Adaption-Innovation Inventory using multiple-group mean and covariance structure analyses. *Multivariate Behavioral Research, 35*(2), 169-199.

Chatfield, C. (2013). *The analysis of time series: An introduction, sixth edition (Vol. 19)*. Boca Raton, FL: CRC Press.

Cheung, M. W. L. (2006). Evaluating multilevel models in cross-cultural research: An illustration with social axioms. *Journal of Cross-Cultural Psychology, 37*(5), 522–541.

Clauser, B. E., Nungester, R. J., Mazor, K., & Ripkey, D. (1996). A comparison of alternative matching strategies for DIF detection in tests that are multidimensional. *Journal of Educational Measurement*, 33(2), 202-214.

Cross-cultural survey guidelines. (2011). Guidelines for best practice in cross-cultural surveys. Ann Arbor, MI: Survey Research Center, Institute for Social Research, University of Michigan.

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*(S1), 244-266.

Davidov, E., Schmidt, P., & Billiet, J. (Eds.). (2011). *Cross-cultural analysis: Methods and applications.* New York, NY: Routledge Taylor & Francis Group.

Delwiche, L. D., & Slaughter, S. J. (2012). *The little SAS book: A primer*. Cary, NC: SAS Institute.

Díaz-Uriarte, R., & De Andres, S. A. (2006). Gene selection and classification of microarray data using random forest. *BMC Bioinformatics, 7*(1), 1.

Eid, M., Langeheine, R., & Diener, E. (2003). Comparing typological structures across cultures by multigroup latent class analysis: A primer. *Journal of Cross-Cultural Psychology, 34*(2), 195–210.

Feinberg, S. (1978). *The analysis of cross-classified categorical data*. Cambridge, Massachusetts: MIT Press.

Filler, M. G., & DiGabriele, J. A. (2012). *A quantitative approach to commercial damages: Applying statistics to the measurement of lost profits*. Hoboken, NJ: John Wiley & Sons.

Fischer, R. (2009). Where is culture in cross cultural research? An outline of a multilevel research process for measuring culture as a shared meaning system. *International Journal of Cross Cultural Management, 9*(1), 25–49.

Fox, J. (2006). Structural equation modeling with the SEM package in R. *Structural Equation Modeling, 13*(3), 465–486.

Fox, J. P. (2010). *Bayesian item response modeling: Theory and applications*. New York, NY: Springer Science & Business Media.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science, 7*(4), 457–472.

Geweke, J. (1992), *Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments*. Proceedings of the Fourth Valencia International Conference on Bayesian Statistics.

Gill, J., & Womack, A. (2013). *The multilevel model framework. The SAGE handbook of multilevel modeling.* Thousand Oaks, CA: Sage Publications.

Halekoh, U., Højsgaard, S., & Yan, J. (2006). The R package geepack for generalized estimating equations. *Journal of Statistical Software, 15*(2), 1-11.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage Publications.

Heeringa, S. G., West, B. T., & Berglund, P. A. (2010). *Applied survey data analys*is. Boca Raton, FL: CRC Press.

Heiberger, R. M., & Holland, B. (2004). *Statistical analysis and data display: An intermediate course with examples in S-plus, R, and SAS.* .New York, NY: Springer Texts in Statistics.

Holbert, R. L., & Stephenson, M. T. (2008). Commentary on the uses and misuses of structural equation modeling in communication research. In A. F. Hayes, M. D. Slater, & L. B. Snyder (Eds.), *The SAGE sourcebook of advanced data analysis methods for communication research* (pp. 185–219). Thousand Oaks, CA: Sage Publications.

Hoyle, R. H. (Ed.). (2012). *Handbook of structural equation modeling*. New York, NY: Guilford Press.

Janssen, R. (2011). Using differential item functioning approach to investigate measurement invariance. In E. Davidov, P. Schmidt, & J. Billiet (Eds.), *Cross-cultural analysis: Methods and applications* (pp. 415-432). New York, NY: Routledge Taylor & Francis Group.

Kankaraš, M., & Moors, G. (2009). Measurement equivalence in solidarity attitudes in Europe: Insights from a multiple-group latent-class factor approach. *International Sociology, 24*(4), 557–579.

Kankaraš, M., & Moors, G. (2010). Researching measurement equivalence in cross-cultural studies, *Psihologija 2010, 43*(2), 121-136.

Kankaraš, M., Moors, G., & Vermunt, J. K. (2010). Testing for measurement invariance with latent class analysis. In E. Davidov, P. Schmidt, & J. Billiet (Eds.), *Cross-cultural analysis: Methods and applications,* (pp. 359–384). New York, NY: Routledge Taylor & Francis Group.

Kankaraš, M., Vermunt, J. K., & Moors, G. (2011). Measurement equivalence of ordinal items: A comparison of factor analytic, item response theory, and latent class approaches. *Sociological Methods & Research 40*, 279-310.

Kaplan, D., & Depaoli, S. (2013). Bayesian statistical methods. Retrieved from http://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199934874.001.0001/oxfordhb-9780199934874-e-020

Kitsantas, P., Kitsantas, A., & Anagnostopoulou, T. (2008). A cross-cultural investigation of college student alcohol consumption: A classification tree analysis. *The Journal of Psychology, 142*(1), 5–20.

Kline, R. (2011). *Structural equation modeling*. New York, NY: Guilford Press.

Kreuter, F., & Muthén, B. (2008). Analyzing criminal trajectory profiles: Bridging multilevel and group-based approaches using growth mixture modeling. *Journal of Quantitative Criminology, 24*(1), 1-31.

LaLonde, S. M. (2005). *Transforming variables for normality and linearity—when, how, why and why not's.* In SAS Conference Proceedings NESUG.

Lanza, S. T., Collins, L. M., Lemmon, D. R., & Schafer, J. L. (2007). PROC LCA: A SAS procedure for latent class analysis. *Structural Equation Modeling, 14*(4), 671–694.

Ledolter, J. (2013). Further discussion on regression and classification trees, computer software, and other useful classification methods. In J. Ledolter (Ed.), *Data mining and business analytics with R* (pp. 185–195). Hoboken, NJ: John Wiley & Sons.

Lee, B. K., Lessler, J., & Stuart, E. A. (2010). Improving propensity score weighting using machine learning. *Statistics in Medicine, 29*(3), 337-346.

Lee, S. Y. (2012). *Structural equation modeling: A Bayesian approach*. Hoboken, NJ: John Wiley & Sons.

Lee, S. Y., & Song, X. Y. (2012). *Basic and advanced Bayesian structural equation modeling: With applications in the medical and behavioral sciences*. Hoboken, NJ: John Wiley & Sons.

Lemon, S. C., Roy, J., Clark, M. A, Friedmann, P. D., & Rakowski, W. (2003). Classification and regression tree analysis in public health: Methodological review and comparison with logistic regression*. Annals of Behavioral Medicine : A Publication of the Society of Behavioral Medicine, 26*(3), 172–181.

Lewis, R. J., & Street, W. C. (2000). *An introduction to classification and regression tree (CART) analysis.* Presented at the 2000 Annual Meeting of the Society for Academic Emergency Medicine in San Francisco.

Linzer, D. A, & Lewis, J. B. (2001). poLCA : An R package for polytomous variable. *Journal of Statistical Software, 42*(10).

Loh, W.-Y. (2014). Fifty years of classification and regression trees. *International Statistical Review, 82*(3), 329-348.

McCutcheon, A. L. (1987). *Latent class analysis (No. 64)*. Thousand Oaks, CA: Sage Publications.

McDonald, J. H. (2009). *Handbook of biological statistics* (Vol. 2, pp. 173-181). Baltimore, MD: Sparky House Publishing.

Merlo, J., Chaix, B., Yang, M., Lynch, J., & Råstam, L. (2005). A brief conceptual tutorial of multilevel analysis in social epidemiology: Linking the statistical concept of clustering to the idea of contextual phenomenon. *Journal of Epidemiology and Community Health, 59*(6), 443-449.

Meuleman, B., & Billiet, J. (2011). Religious involvement: Its relation to values and social attitudes. In E. Davidov, P. Schmidt, & J. Billiet (Eds.), *Cross-cultural analysis: Methods and applications* (pp. 173-206). New York, NY: Routledge Taylor & Francis Group.

Mneimneh, Z. N., Heeringa, S. G., Tourangeau, R., & Elliott, M. R. (2014). Bridging psychometrics and survey methodology: Can mixed Rasch models identify socially desirable reporting behavior? *Journal of Survey Statistics and Methodology, 2*(3), 257-282.

Muthén, B. O., & Asparouhov, T. (2012). Bayesian structural equation modeling: A more flexible representation of substantive theory. *Psychology Methods 17*:313–35

Muthén, L. K., & Muthén, B. (2007). *Multilevel modeling with latent variables using Mplus*. Unpublished manuscript. Retrieved from: http://www.ats.ucla.edu/stat/mplus/seminars/mlmMplus_JH/Hopkins_Day2_fixed_May2005_two_per_page.pdf

Nering, M. L., & Ostini, R. (Eds.). (2011). *Handbook of polytomous item response theory models*. New York, NY: Routledge Taylor & Francis Group.

Ostini, R., & Nering, M. L. (2005). *Polytomous item response theory models (No. 144)*. Thousand Oaks, CA: Sage Publications.

Padilla, J. L., Hidalgo, M. D., Benítez, I., & Gómez-Benito, J. (2012). Comparison of three software programs for evaluating DIF by means of the Mantel-Haenszel procedure: EASY-DIF, DIFAS and EZDIF. *Psicologica: International Journal of Methodology and Experimental Psychology*, 33(1), 135-156.

Piccinelli, M., & Simon, G. (1997). Gender and cross-cultural differences in somatic symptoms associated with emotional distress. An international study in primary care. *Psychological Medicine, 27*(2), 433–444.

Quandt, M. (2011). Using the mixed Rasch model in the comparative analysis of attitudes. In E. Davidov, P. Schmidt, & J. Billiet (Eds.), *Cross-cultural analysis: Methods and applications* (pp. 433-460). New York, NY: Routledge Taylor & Francis Group.

Rabe-Hesketh, S., & Skrondal, A. (2006). Multilevel modelling of complex survey data. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 169*(4), 805-827.

Rabe-Hesketh, S., Skrondal, A., & Zheng, X. (2007). Multilevel structural equation modeling. In S. Y. Lee (Ed.). *Handbook on structural structural equation models* (pp. 209-227). Amsterdam: Elsevier.

Raftery, A. E., & Lewis, S. M. (1992). [Practical Markov Chain Monte Carlo]: comment: One long run with diagnostics: Implementation strategies for Markov Chain Monte Carlo. *Statistical Science, 7*(4), 493-497.

Randolph, K. A., & Myers, L. L. (2013). *Basic statistics in multivariate analysis*. OUP USA.

Reviews, C. T. (2013). e-Study guide for reading and understanding research. In L. F. Locke (Ed.), *Statistics, research methods. Cram101*.

Rovai, A. P., Baker, J. D., & Ponton, M. K. (2013). *Social science research design and statistics: A practitioner's guide to research methods and IBM SPSS*. Chesapeake, VA: Watertree Press LLC.

Sansone, C., Morf, C. C., & Panter, A. T. (2004). *The SAGE handbook of methods in social psychology.* Thousand Oaks, CA: Sage Publications.

Schmidt, C. O., & Kohlmann, T. (2008). When to use the odds ratio or the relative risk? *International Journal of Public Health, 53*(3), 165–167.

Singer, J. D. (1998). Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models. *Journal of Educational and Behavioral Statistics, 23*(4), 323-355.

Sistrom, C. L., & Garvan, C. W. (2004). Proportions, odds, and risk. *Radiology, 230*(1), 12–9.

Skrivanek, S. (2009). *The use of dummy variables in regression analysis*. Retrieved from https://www.moresteam.com/WhitePapers/download/dummy-variables.pdf

Snedecor, G. W., & Cochran, W. G. (1994). *Statistical methods (eighth edition).* Calcutta, India: Oxford & IBH Publishing Co.

Stapleton, L. M. (2006). An assessment of practical solutions for structural equation modeling with complex sample data. *Structural Equation Modeling, 13*(1), 28-58.

Steel, S. J., & Uys, D. W. (2007). Variable selection in multiple linear regression : The influence of individual cases. *Orion, 23*(2), 123–136.

Steele, F. (2008). Multilevel models for longitudinal data. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 171*(1), 5-19.

Steenkamp, J.-B., & Baumgartner, H. (1998). Assessing measurement invariance in consumer JAN-BENEDICT. *Journal of Consumer Research, 25*(1), 78–90.

Steinmetz, H. (2011). Estimation and comparison of latent means across cultures. In E. Davidov, P. Schmidt, & J. Billiet (Eds.), *Cross-cultural analysis: Methods and applications* (pp. 85-116). New York, NY: Routledge Taylor & Francis Group.

Steinmetz, H., Schmidt, P., Tina-Booh, A., Wieczorek, S., & Schwartz, S. H. (2008). Testing measurement invariance using multigroup CFA: Differences between educational groups in human values measurement. *Quality & Quantity, 43*(4), 599–616.

Stone, C. A., & Zhu, X. (2015). *Bayesian analysis of item response theory models using SAS®*. Cary, NC: SAS Institute Inc.

Swihart, B. J., Caffo, B., James, B. D., Strand, M., Schwartz, B. S., & Punjabi, N. M. (2010). Lasagna plots: A saucy alternative to spaghetti plots. *Epidemiology, 21*(5), 621–5.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models.

Twisk, J. W. (2006). *Applied multilevel analysis: A practical guide for medical researchers*. Cambridge, UK: Cambridge University Press.

Valliant, R., Dever, J. A., & Kreuter, F. (2013). *Practical tools for designing and weighting survey samples*. New York, NY: Springer.

van de Vijver, F. J. R., & Leung, K. (1997). *Methods and data analysis for cross-cultural research*. Thousand Oaks, CA: Sage Publications.

van de Vijver, F. J., van Hemert, D. A., & Poortinga, Y. H. (2008). *Individuals and cultures in multilevel analysis*. Mahwah, NJ: Lawrence Erlbaum.

van de Vijver, F. J., van Hemert, D. A., & Poortinga, Y. H. (2015). *Multilevel analysis of individuals and cultures*. New York, NY: Psychology Press.

van der Linden, W. J., & Hambleton, R. K. (Eds.). (2013). *Handbook of modern item response theory*. New York, NY: Springer Science & Business Media.

Vermunt, J. K. (2004). Latent profile analysis. In M. Lewis-Beck, A. Bryman, & T. F. Liao (Eds.), *The SAGE encyclopedia of social sciences research methods* (pp. 554–555). Thousand Oaks: Sage Publications..

Welch, K. (2009). *Random effects vs. marginal models: Different approaches to analyzing repeated measures/longitudinal data*. Presented to the Midwest SAS Users Group. Retrieved from http://www.misug.org/uploads/8/1/9/1/8191072/kwelch_repeated_measures.pdf

Welzel, C., & Inglehart, R. F. (2016). Misconceptions of measurement equivalence time for a paradigm shift. *Comparative Political Studies*, 0010414016628275.

West, B. T. (2009). Analyzing longitudinal data with the linear mixed models procedure in SPSS. *Evaluation and the Health Professions, 32*(3), 207-228.

West, B. T., & Galecki, A. T. (2011). An overview of current software procedures for fitting linear mixed models. *The American Statistician, 65*(4), 274-282.

Yong, A. G., & Pearce, S. (2013). A beginner's guide to factor analysis: Focusing on exploratory factor analysis. *Tutorials in Quantitative Methods for Psychology, 9*(2), 79–94.

Zumbo, B. D. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly*, 4(2), 223-233.

Survey Quality

Sue Ellen Hansen, Grant Benson, Ashley Bowers, Beth-Ellen Pennell, Yu-chieh (Jay) Lin, Benjamin Duffey, Mengyao Hu, and Kristen Cibelli Hibben, 2016

Introduction

This chapter presents a framework for assessing quality in multinational, multicultural, or multiregional surveys, which we refer to as “3MC” surveys, followed by guidelines for managing and assessing quality throughout the survey lifecycle.

To monitor and control the performance of survey operations, survey organizations have developed methods and procedures to prevent and correct problems which can affect survey data quality (Lyberg & Biemer, 2008). Such methods and procedures are often referred to as quality assurance (Guideline 3), which refers to the planned procedures and activities an organization uses to ensure that the study meets quality requirements. A concept closely related to quality assurance is survey process quality management, otherwise referred to as quality control (see Figure 4), which refers to a planned system of process monitoring, verification and analysis of indicators of quality, and updates to quality assurance procedures, to ensure that quality assurance works. In some respects, quality control (Guideline 4) may be viewed as part of quality assurance (Lyberg & Biemer, 2008). However, these are separated in this set of guidelines to make monitoring and controlling performance and product quality an explicit part of quality management (Guideline 2).

In mono-cultural surveys, assessing the quality of survey data requires adequate documentation of the entire survey lifecycle and an understanding of protocols used to assure quality. In such surveys, there may be challenges in overcoming methodological, organizational, and operational barriers when performing quality assurance and quality control. For example, a country may not have the infrastructure or an organization may not have the means to implement a study entirely according to survey best practices.

In 3MC survey projects, the challenges increase. As noted by Lyberg and Stukel (2010), “quality assurance (Guideline 3) and quality control (Guideline 4) programs are, in general, less prominent and visible in cross-national comparative studies than in national surveys”(Lyberg & Stukel, 2010). 3MC surveys hinge on the comparability or equivalence of data across cultures. Moreover, 3MC survey quality assessment procedures and criteria become more complex with additional survey processes, such as adaptation and translation of questions and harmonization of data across multiple surveys (see Adaptation, Translation, and Data Harmonization), leaving little room for quality assurance and quality control. This is especially true in countries with limited “financial, methodological, and technological resources and expertise” (Mneimneh, Lyberg, Sharma, Vyas, & Sathe (2017). Given the magnified quality assurance and quality control problems in 3MC surveys, as suggested by Lyberg and Stukel (2010), it is of critical importance to include these quality programs, insofar as possible, in 3MC surveys.

As the survey lifecycle illustrates, quality and ethical considerations are relevant to all processes throughout the survey production lifecycle. Survey quality can be assessed in terms of fitness for intended use (also known as fitness for purpose (Lyberg et al., 2001), total survey error, and the monitoring of survey production process quality, which may be affected by survey infrastructure, costs, respondent and interviewer burden, and study design specifications.

Quality Framework

Survey quality is a vague concept, which has multiple definitions and has origins in two different developmental paths (Biemer & Lyberg, 2003; Lyberg, 2012). One path is the total survey error paradigm; the other path focuses more on quality management sciences, including fitness for use and survey process quality (see below) (Lyberg, 2012). The development of the overall paradigm of survey quality from both the total survey error perspective, as well as the quality management sciences perspective, as mentioned by Lyberg (2012), has taken place mainly in official statistics and organizations and has been triggered by the rapid development of technology and other developments. Technological development have largely improved the efficiency of quality monitoring and control procedures, and has influenced potential quality dimensions like "accessibility, timeliness, and coherence" (Lyberg, 2012). In addition, given the increased demands for harmonized and comparable survey statistics and complex decision-making, it is essential to develop a quality framework which can accommodate all of these demands (Lyberg, 2012). Various quality frameworks have been developed for this reason. See Lyberg (2012) for more information on the development of the survey quality paradigm and different frameworks.

The framework adopted in this chapter for assuring and assessing quality is informed by research on survey errors and costs and quality management, and highlights three aspects of quality: total survey error (Groves, 1989; Groves et al., 2009), fitness for intended use (Defeo & Juran, 2010), also known as "fitness for purpose" (Lyberg et al., 2001), and survey process quality (Biemer & Lyberg, 2003; Lyberg et al., 1997; Morganstein & Marker, 1997). The three aspects of quality are described in turn below:

Total survey error

The total survey error (TSE) paradigm is widely accepted as a conceptual framework for evaluating survey data quality (Anderson, Kasper, Frankel, & Associates, 1979; Cochran, 1977). Errors in survey estimates consist of variances of estimates (reflecting estimate instability over conceptual replications) and systematic deviations from a target value ("biases"). TSE defines quality as the estimation and reduction of the mean square error (MSE) of statistics of interest, which is the sum of random errors (variance) and squared systematic errors (bias). The MSE for each individual statistic in a survey is not typically calculated, due to the following practical problems (see Vehovar, Slavec, and Berzelak (2012) for detailed discussions). First, MSE needs to be calculated differently for different survey parameters (e.g., the survey population mean and variance). It can also differ for each survey item. The fact that a survey usually contains many items and many parameters poses a challenge for the practical application of MSE. Second, the true scores used in bias estimation are often unknown and are usually obtained from a benchmark survey such as Census data or "gold-standard" estimates such as from a face-to-face survey. The accuracy of these estimates, however, is not guaranteed. Third, given that MSE is often a combination of different error sources, it is sometimes difficult to distinguish and separate these error sources. These practical issues become more complicated in 3MC surveys, posing additional challenges to the use of MSE. Despite the challenges, however, the TSE framework helps organize and identify error sources and estimates their relative magnitude, which can assist those planning 3MC surveys to evaluate design and implementation tradeoffs.

TSE takes into consideration both measurement (construct validity, measurement error, and processing error)—i.e., how well survey questions measure the constructs of interest—as well as representation (coverage error, sampling error, nonresponse error, and adjustment error) (Groves et al., 2009)—i.e., whether one can generalize to the target population using sample survey data. In the TSE perspective, there may be cost-error tradeoffs; that is, there may be tension between reducing these errors and the cost of reducing them.

Although the TSE paradigm is increasingly used as an organizing framework in the design and evaluation of single-country surveys (Biemer, 2014), it is rarely cited in reference to 3MC research. Pennell, Lyberg, Mohler, Hibben, and Worku (2017) offer a total survey error framework adapted and expanded from Groves et al. (2009), Tourangeau, Rips, and Rasinski (2000), Smith (2011a), and Smith (2017) for 3MC survey research that integrates error sources with methodological and operational challenges that are unique to or may be more prominent in 3MC surveys (see Figure 1 below).

It is important to note, that like the framework proposed by Groves et al. (2009), Pennell et al.'s (2017) framework does not elaborate on the distinction between systematic and variable error discussed above for the sake of parsimony.

Figure 9.1 TSE (a) representation in a cross-cultural context

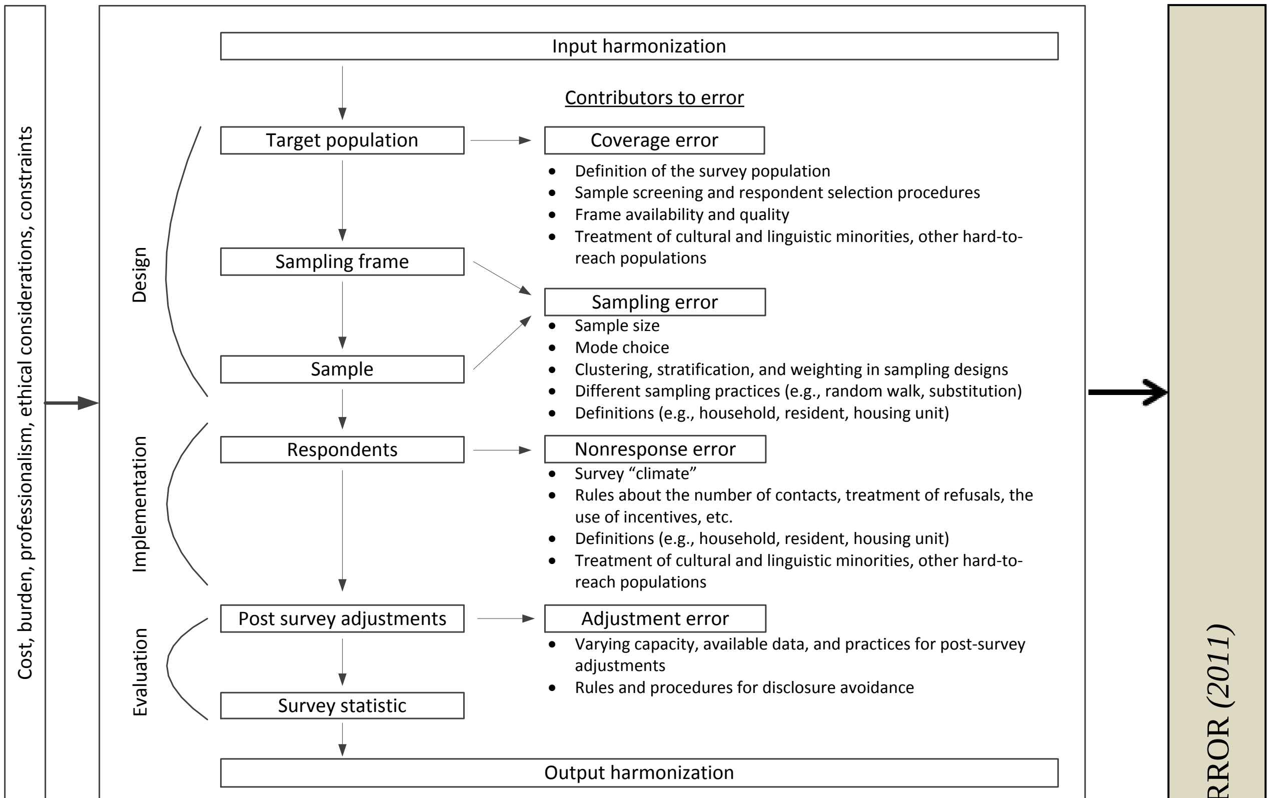
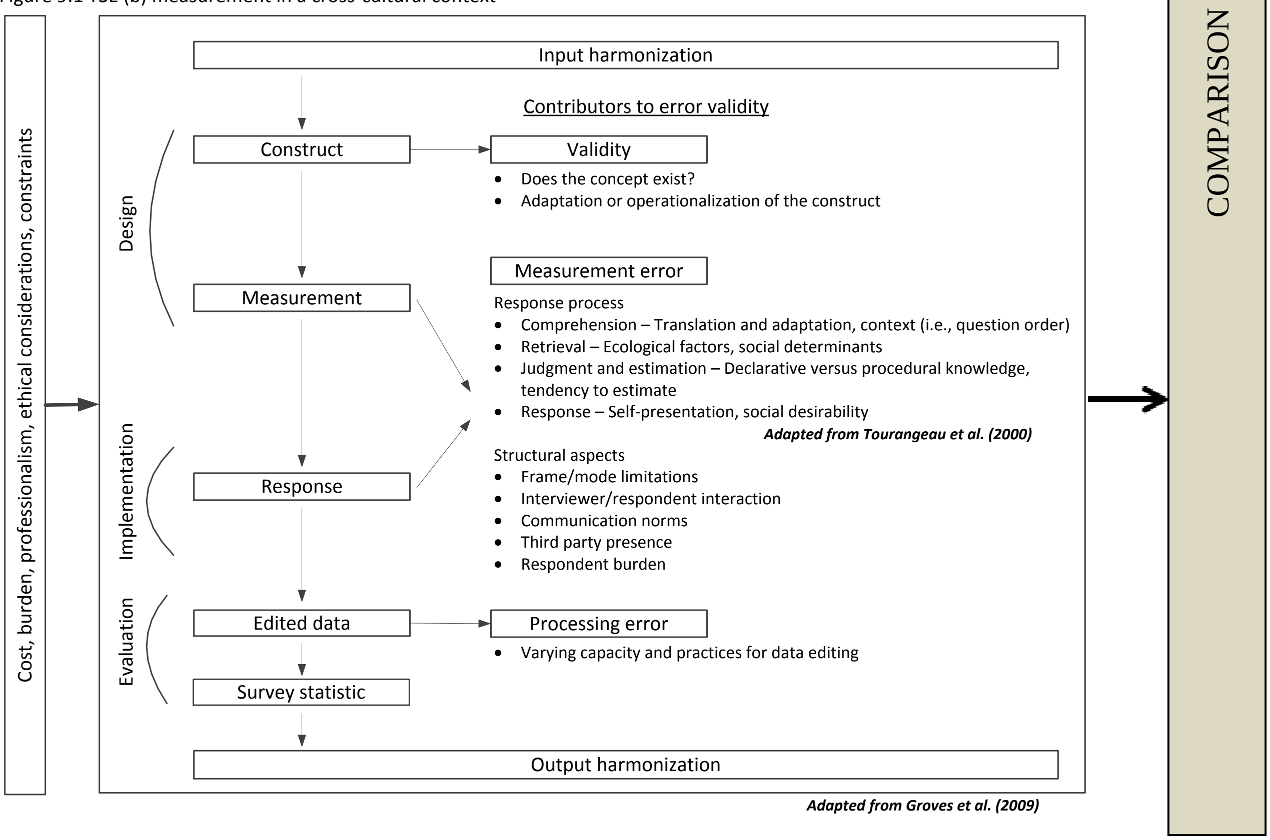


Figure 9.1 TSE (b) measurement in a cross-cultural context



Adapted from Groves et al. (2009)

The following describes the main elements of Pennell et al.'s (2017) TSE framework:

- The framework links error sources to the key stages of the survey process: design, implementation, and evaluation.

- Part A of Figure 1 outlines representation error—including coverage error, sampling error, nonresponse error, and adjustment error—which are indicators of how well survey estimates generalize to the target population.

- Part B of Figure 1 encompasses measurement related error—including validity, measurement error, and processing error—which are indicators of how well survey questions measure the constructs of interest.

- As denoted by the resulting "survey statistic" at the end of Part A and Part B, the framework produces statistic-specific error profiles for representation and measurement errors for a single survey statistic. The framework produces statistic specific error profiles because the presence and scale of error may, and frequently does, vary across individual survey statistics.

- The framework incorporates the dimensions of cost, burden, professionalism, ethics, and other design constraints that frequently impose constraints on 3MC survey design and have an important influence on the quality of 3MC surveys.

- The framework includes the role of input harmonization and output harmonization, which are unique to 3MC surveys. Input and output harmonization represent two general approaches to harmonization, which is a term for procedures aimed at achieving, or at least improving, the comparability of different surveys. See Harmonization for further discussion.

- "Comparison error"—a concept introduced by Smith (2011b) —is the conceptual error introduced across each component of a 3MC survey as well as the aggregate of error across all components, which could threaten comparability across surveys.

- For each error component (e.g., coverage error, sampling error, measurement error, etc.), key potential sources of error are identified that may contribute to TSE in individual populations and may present particular challenges to standardizing design and implementation (or establishing suitable localized equivalents) across populations, thereby potentially increasing comparison error. See Pennell et al. (2017) for a detailed discussion of key potential contributions to error and design and implementation challenges across the main stages of the survey lifecycle.

With advances in computerized interviewing software and sample management systems, data related to quality increasingly can be collected with survey data, and can be used to measure various components of error. These include paradata (see Paradata and Other Auxiliary Data) (Biemer & Lyberg, 2003; Brackstone, 1999), data from experiments embedded in a survey, and supplementary data, such as nonresponse follow-up measures. Each of these facilitates evaluation of survey data in terms of TSE.

Fitness for intended use

Biemer and Lyberg (2003) argue that the TSE framework lacks a user perspective, and that it should be supplemented by using a more modern quality paradigm—one that is multidimensional and focuses on criteria for assessing quality in terms of the degree to which survey data meet user requirements (fitness for intended use). 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, such as comparability and timeliness. In this perspective, ensuring quality on one dimension (comparability) may conflict with ensuring quality on another dimension (timeliness); and there may be tension between meeting user requirements and the associated cost of doing so on one or more dimensions. There are a number of multidimensional quality frameworks in use across the world (see, for example, Brackstone (1999), Couper (1998), International Monetary Fund (2003), Statistics Canada (2002), and Tupek (2006)).

Table 1 shows seven dimensions that are often used to assess the quality of national official statistics in terms of both survey error and fitness for use: comparability, relevance, accuracy, timeliness and punctuality, accessibility, interpretability, and coherence. In this framework, TSE may be viewed as being encompassed by the accuracy dimension.

Table 1. Dimensions of Quality

Quality Dimension	Description
Comparability	Are the data from different countries or cultures comparable to each other (i.e., equivalent)?
Coherence	Do the data form a coherent body of information that can be rearranged or combined with other data?
Relevance	Do the data meet the requirements of the client and users?
Accuracy	Are the data 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?
Timeliness and punctuality	How much time has elapsed between the end of the data collection and when the data are available for analysis? Are the data available when expected, based on client specifications?
Accessibility	Can users easily obtain and analyze the data?
Interpretability	Do the data make sense in terms of users' hypotheses? Are supplementary data available to facilitate analysis: e.g., data that describe the major characteristics and structure of the data (metadata) as well as data about the survey processes (paradata and other auxiliary data)?

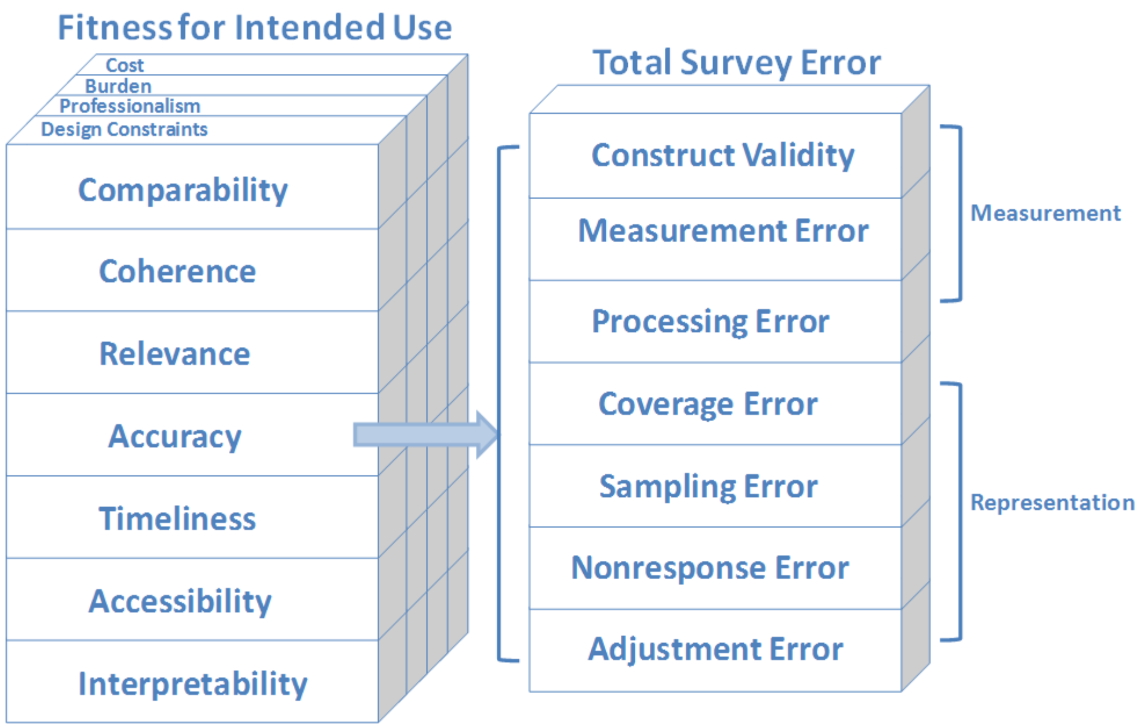
Cost, burden, professionalism, and design constraints are factors that may also affect fitness for use on these dimensions:

- **Cost** – are monetary resources optimized?
- **Burden** – given the necessary information obtained, are interviewer and respondent burden minimized?
- **Professionalism** – are staff provided with clear behavioral guidelines and professional training, are there adequate provisions to ensure compliance with relevant laws, and is there demonstration that analyses and reporting have been impartial? Note that the professionalism can vary a lot across countries, and tailored specifications and trainings may be needed.
- **Design Constraints** – are there context-specific constraints on survey design that may have had an impact on quality (for example, use of a different mode of data collection in one culture than in others, or use of different sample frames betweem countries)?

The aim is to optimize costs, minimize burden, and recognize and document design constraints where appropriate—based on the need to be sensitive to local survey contexts, and to maximize professionalism. Figure 2 shows the dimensions of quality as well as those factors that affect quality in terms of fitness for use (see Australian Bureau of Statistics (2009), Brackstone (1999), Couper (1998), International Monetary Fund (2003), Statistics Canada (2002), and

Tupek (2006) for examples of dimensions of quality used by statistical agencies). It also shows the accuracy dimension in terms of TSE (Anderson et al., 1979; Groves, 1989; Groves et al., 2009).

Figure 2. Fitness for Intended Use (Quality Dimensions) and Total Survey Error (Accuracy Dimension)



The dimensions of quality (comparability, coherence, relevance, accuracy, etc.) and factors that may have an impact on quality (cost, burden, professionalism, and design constraints) apply to all surveys. However, in a 3MC context, challenges increase:

- The quality dimensions of **coherence** and **comparability** are the *raison d'être* for cross-national and cross-cultural survey research. Fitness for intended use cannot be met without quality on these dimensions.

- **Relevance** may be harder to achieve in comparative research, in that decisions have to be made about what level of relevance to aim for with a standardized survey across many cultures and countries.

- **Accuracy** in terms of TSE may be difficult to estimate consistently across 3MC surveys.

- **Timeliness and punctuality** may be a challenge in 3MC research; for example, data collection may occur in vastly different climates or with varying organizational infrastructures (see Data Collection: General Considerations).

- **Accessibility** in the 3MC context can mean more than simply making survey data publicly available, particularly in countries with fewer resources, where it also may be necessary to include capacity building or data user training to make the data truly accessible to local users. Country-level data access laws and regulations can also come into play (see Data Dissemination).

- **Interpretability** of data may be difficult without metadata documentation about the data that would facilitate comparison across 3MC surveys (see Data Dissemination).

Appendix A highlights recommendations from other guidelines in relation to dimensions of quality.

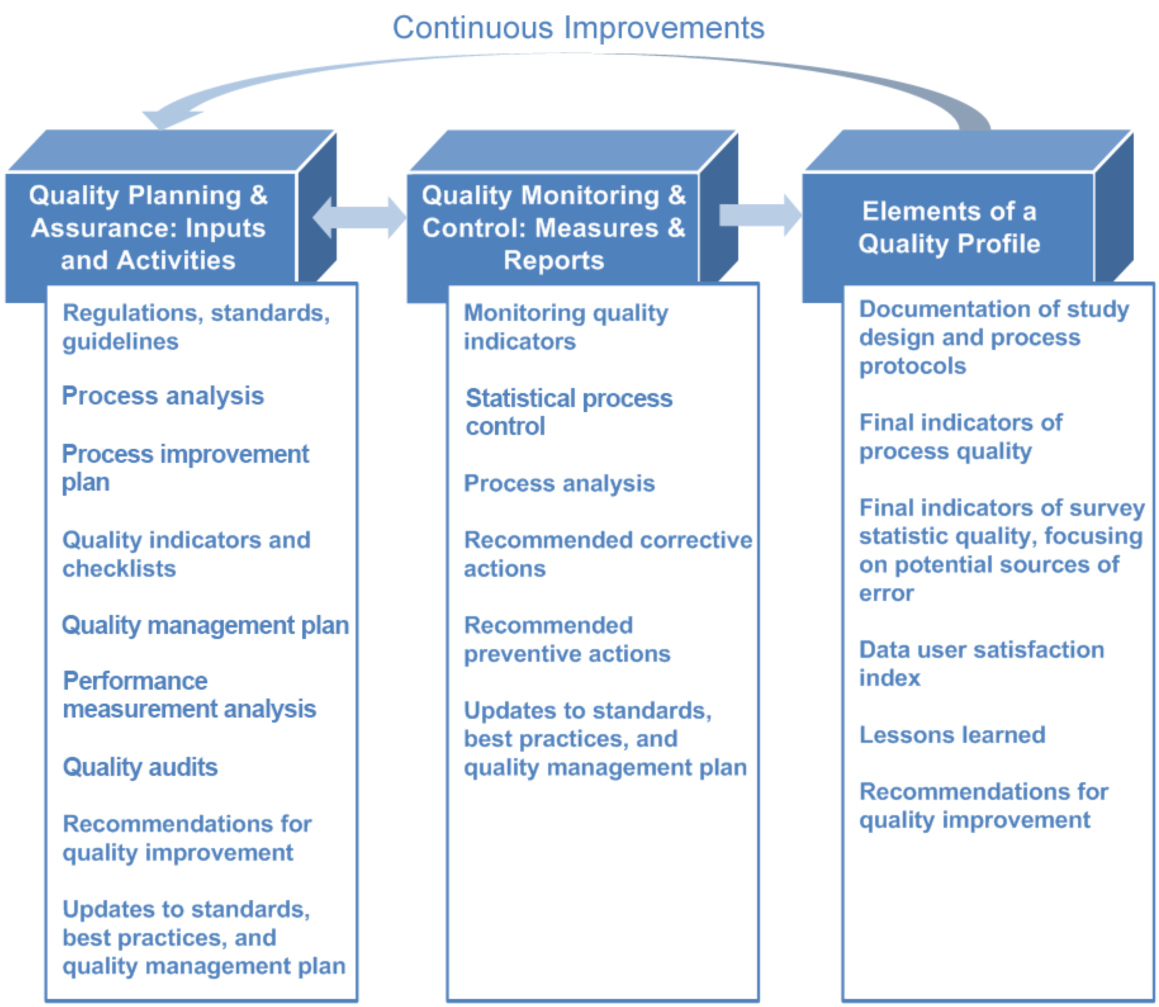
Survey process quality

Fitness for intended use provides a general framework for assessing the quality of 3MC surveys and defines the essential dimensions of quality, one of which is accuracy (i.e., TSE). A third approach to quality monitoring and assessment is survey process quality management and the notion of continuous process improvement (Groves et al., 2009). This approach focuses on quality at three levels: the product, the process, and the organization (Lyberg & Biemer, 2008). The product quality, as mentioned by Lyberg and Stukel (2010) is the expected quality of survey deliverables, which is often decided by clients or the users. The process quality refers to the quality of process that generates the product. One way to monitor and control process quality is to choose, measure, and analyze process variables relevant to the particular survey (Lyberg & Stukel, 2010). However, process quality depends on the third level —organization quality—an important element of quality assurance and quality control in 3MC surveys is the choice of survey organization in each country. As mentioned by Lyberg and Stukel (2010), without high quality organizations, there may be insufficient resources to apply processes as specified, which might lead to deviations from the specifications. Thus, product quality cannot be achieved without good process quality, and having good process quality requires a good organization that manages for quality.

A focus on survey production process quality requires the use of quality standards and collection of standardized study metadata, question metadata, and process paradata (Couper, 1998). Figure 3 shows the elements of survey process quality management that allow users to assess the quality of processes throughout the survey lifecycle: quality assurance, quality control (Lyberg & Stukel, 2010; Lyberg & Biemer, 2008), and a quality profile (or a quality report, which documents all that is known about the quality of a survey) (Biemer & Lyberg, 2003; Eurostat, 2003b). The arrows in Figure 3 indicate that quality planning and assurance will guide the quality control process, and the latter will in turn influence the quality planning and assurance. The quality profile will reflect the quality control activities and will also make recommendations for quality improvements, which will be reflected in the future quality planning and assurance. These are discussed further in the guidelines below.

Data collection organizations involved in a 3MC survey may vary in what cost-quality tradeoffs they can make, as well as the processes they generally monitor for quality purposes. However, if each organization reaches a minimum standard through adherence to the quality guidelines of the study's coordinating center (which monitors and oversees the 3MC survey activities), the coordinating center can assess the quality of each survey based on quality indicators (paradata) from each organization, and create a quality profile that allows users to assess survey data quality and comparability across cultures. Appendix B summarizes, for each set of guidelines, examples of elements of quality planning and assurance, quality monitoring and control, and a quality profile.

Figure 3. Survey Process Quality Management



Guidelines

Goal: To ensure the quality of survey production processes and consequently the survey data throughout the survey lifecycle, as well as clear and comprehensive documentation of study methodology, and to provide indicators of process and data quality.

1. Build a solid coordinating center

Rationale

To minimize comparison error in 3MC surveys, survey designers develop various guidelines and requirements for each step in the survey lifecycle. One might expect that providing such detailed guidelines and requirements to participating countries would be sufficient and that the countries would understand and follow the prescribed protocols closely. However, experience shows that this is often not the case. Countries can deviate largely from the instructions, which will greatly jeopardize the comparability. One such example is the 1994 International Adult Literacy Survey (IALS) (Murray, Kirsch, & Jenkins, 1998; Pennell et al., 2017). The coordinating center did not perform any quality control programs to make sure that each country followed the instructions strictly. This led to various deviations from the original design, and one country had to be withdrawn from the study (Kalton, Lyberg, & Rempp, 1998). The European Commission decided to improve future IALS by implementing standardized procedures including setting up a strong central infrastructure (Carey, 2000). Such infrastructure is essential for maintaining countries' adherence to the survey instructions and requirements and should be part of the quality assurance plan. In addition, most 3MC surveys are on an on-going basis, where a central infrastructure is also needed to "plan, coordinate, support, and improve international survey endeavors" (Lyberg & Stukel, 2010). See Study Design and Organizational Structure for further discussion on the establishment of coordinating center.

Procedural Steps

1.1 Clearly document the role and tasks of the coordinating center and the responsibilities of each country's coordinator in the quality management plan (see Guideline 2).

1.1.1 The tasks of the coordinating center include (Lyberg & Stukel, 2010):

- Study design
- Coordination with national bodies as necessary.

- Development of specifications for processes and participating survey organizations.
- Providing instructions and support to each participating country.
- Implementing and supervising survey operations.
- Maintaining adherence to user demands and study design.
- Design of quality control procedures.
- Suggestions for future improvements.
- Documentation of survey process and results.

1.1.2 The responsibilities of the country or local coordinator include:

- Organizing scheduling.
- Communication with the coordinating center.
- Deliverables to and from the coordinating center, such as questionnaire designs (see Questionnaire Design).

1.2 Set up the supporting system for the coordinating center. For example, the Central Coordinating Team (CCT) of the European Social Survey (ESS) is supported by a Scientific Advisory Board, as well as four Specialist Advisory Groups on question module design, methods, sampling, and translation (Lyberg & Stukel, 2010).

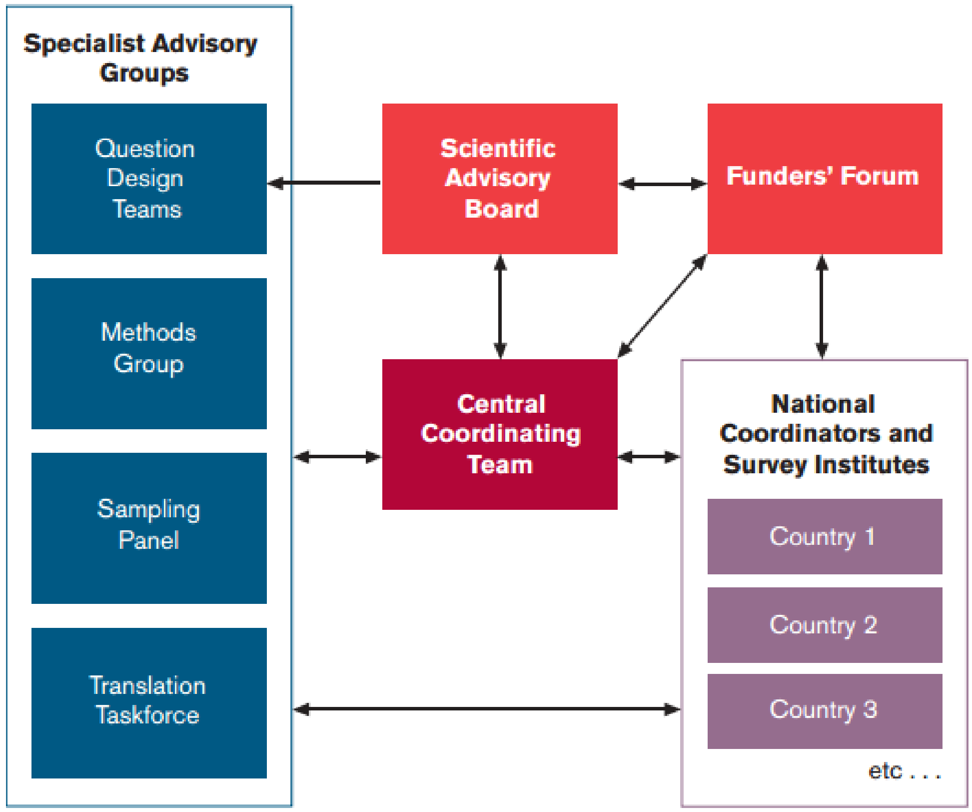
1.3 Set up rules and routines for coordination, such as who the country coordinators should be (e.g., subject matter expert or survey manager), when to communicate, contents of the deliverables, and communication methods (email, telephone, or video conference call).

1.3.1 If possible, appoint an experienced survey manager as a coordinator instead of a subject matter specialist with minimal survey experience. As mentioned by Lyberg and Stukel, (2010), it is important that the coordinator be familiar with error structures associated with different survey operations and appreciate the need for quality assurance and quality control.

Lessons Learned

1.1 To pursue optimal comparability, the European Social Survey (ESS) has set up an infrastructure with a Central Coordinating Team (CCT) at the center. This team receives feedback from national coordinators (NCs) in participating countries, and it closely monitors the adherence to survey instruction and requirements (Lyberg & Stukel, 2010). At the design stages and during the field work, when there is need for ESS input, final decisions and specifications are made by CTT (Stoop, Billiet, Koch, & Fitzgerald, 2010). Figure 4 shows the organizational structure of the ESS.

Figure 4. Organizational structure of the ESS



Source: ESS information Brochure (http://www.europeansocialsurvey.org/docs/about/ESS_brochure.pdf)

1.2 Few organizations have evaluated the infrastructure model they use. Lyberg and Stukel (2010) suggest that, given that many 3MC surveys still have very weak infrastructures for coordination and monitoring, evaluation studies on current infrastructures could eventually result in more efficient infrastructure formations.

2. Develop a sustainable quality management plan.

Rationale

A sustainable quality management plan is essential for developing planned, systematic quality assurance (Guideline 3) and quality control (Guideline 4) activities, which in turn helps ensure that the study and survey data meet client and user requirements. It also facilitates development of a quality profile (Guideline 5), which should document survey methodology, key indicators of quality, lessons learned, and recommendations for improvement.

Procedural Steps

2.1 Review available 3MC survey standards and best practices for ensuring the quality of survey processes, survey data, and documentation (such as these guidelines).

2.2 Review existing quality profiles (see Guideline 5) and lessons learned from other studies. Use standardized quality profiles and protocols to establish sustainable quality management.

2.3 Review study requirements for quality assurance (see Guideline 3) and quality control (see Guideline 4). These may be developed at the study design stage by the coordinating center, the survey organization, or both.

2.4 Review study goals and objectives, required products and deliverables, and study timeline and budget.

2.5 Review country-specific regulations and legislation relevant to conducting survey research.

2.6 Through analysis of the processes in the survey lifecycle—i.e., process analysis—identify characteristics of survey products (e.g., coded data) that could vary during the processes (e.g., verification failures) (Aitken, Hörngren, Jones, Lewis, & Zilhäo, 2003). For example,

2.6.1 Use tools to analyze a process, to determine what steps in the process need to be monitored to ensure quality, and to identify quality indicators to monitor (Aitken et al., 2003). Examples of tools used to analyze processes are:

- Cause and effect diagrams ("fishbone" diagrams).
- Flow charts.

2.7 Identify key indicators of the quality of the product(s) of the process in terms of TSE and other dimensions of quality, as well as factors such as cost, burden, and the risk of not meeting quality requirements. See Appendix A for examples of survey quality indicators as they relate to TSE and the fitness for use quality dimensions. Note that many of the indicators are constructed using paradata and other auxiliary data, such as survey time and keystroke data. See Paradata and Other Auxiliary Data for more information.

2.7.1 If possible, use such indicators to determine whether the process is stable or controlled; that is, is variation on a key indicator due to randomness alone? This can be done using paradata from similar studies the organization has conducted or is conducting, or from pilot studies.

2.7.2 Define measurement and reporting requirements for use during quality assurance (see Guideline 3) and quality control (see Guideline 4), and determine who would be responsible for ensuring that quality assurance and quality control activities are carried out.

2.7.3 Assess whether these requirements can be met through current procedures and systems, and with currently collected paradata; and if not, develop a process improvement plan.

2.7.4 Create cost/error tradeoff decision rules about how to alter the features of the study design if the goals are not met. For example, consider using responsive design, where researchers continually monitor selected paradata to inform the error-cost tradeoff in real-time, as the basis for altering design features during the course of data collection or for subsequent waves (see Paradata and Other Auxiliary Data).

2.8 Use quality planning tools (see Appendix B) to help determine what performance analyses and assessments should be used. Such tools include:

2.8.1 A cost-benefit analysis of potential quality management procedures and activities: evaluating their benefits in relation to the cost of performing them relative to overall study costs.

2.8.2 Benchmarking: comparing planned activities against those of similar studies, and the outcomes of those activities, to form a basis for performance measurement.

2.8.3 Statistical analysis of factors that may influence indicators of process or product quality.

2.9 Develop a quality assurance plan, which could include (see Appendix B):

2.9.1 The process improvement plan.

2.9.2 Performance and product quality baselines.

2.9.3 Process checklists.

2.9.4 A training plan.

2.9.5 Recommended performance analyses and assessments (e.g., quality assurance procedures for verifying interviews and evaluating interviewer performance).

2.9.6 Required process quality audits, reviews, and inspections (e.g., a review of tapes of interviews to assess interviewer performance).

2.10 Develop a quality control plan for continuous monitoring of processes to ensure that they are stable and that products are meeting requirements (see Aitken et al. (2003), Guideline 4, and Appendix B). Such a plan could include:

2.10.1 The process improvement plan.

2.10.2 Performance and product quality baselines.
2.10.3 Quality indicators identified in process analysis and planning for responsive design.
2.10.4 Performance analyses and assessments to use to monitor processes.
2.10.5 Tools to use to monitor processes and product quality (e.g., Pareto charts and statistical process control charts).
2.10.6 Reports to prepare on performance measurement, such as interviewer training certification. Interviewer behavior can be monitored and evaluated using paradata and other auxiliary data. See the Paradata and other Auxiliary Data chapter for more information.

2.11 Develop procedures to ensure that throughout the survey lifecycle all documentation, reports, and files related to quality planning and assurance, quality monitoring and control, and process improvement are retained. This facilitates preparing a quality profile for users of the disseminated survey data (see Guideline 5 and Data Dissemination).

2.12 Develop procedures for updating the quality management plan as needed during the survey lifecycle.

Lessons Learned

2.1 There are many quality management methodologies that survey organizations may use that focus on the three levels of quality: product, process, and organization (for example, Total Quality Management (TQM)). Discussion of such methodologies is beyond the scope of this chapter, but experience has shown that they can help organizations manage for quality.

2.2 Developing a quality management plan alone does not necessarily guarantee quality. Other project management practices may also affect quality. Many survey organizations and statistical agencies have recognized the value of also adhering to professional project management guidelines, such as those of the Project Management Institute (PMI) (Project Management Institute, 2004) and the International Project Management Association (IPMA). Many have certified project managers and follow professional project management best practices that may affect quality, schedule, and costs, such as developing risk management and communication plans. As with a quality management plan, these can be critical to ensuring the quality of processes and survey data. See Study Management for further discussion of the study management process.

3. **Perform quality assurance activities.**

Rationale

Quality assurance is defined by the planned procedures and activities (see Guideline 2) that an organization uses to ensure that the study meets process and product quality requirements. It specifies ways in which quality can be measured.

Procedural Steps

3.1 For each stage in the survey lifecycle, perform quality assurance activities as outlined in the quality management plan. Appendix B includes quality assurance inputs and activities at each stage of the survey lifecycle. Some examples include:
  - 3.1.1 Certification by the coordinating center that an organization's study design and quality standards meet study standards (see Study Design and Organizational Structure).
  - 3.1.2 Pretest consent protocols and forms to ensure comprehension (see Ethical Considerations).

3.2 Perform performance and product quality assessments. Examples are:
  - 3.2.1 Certification of interviewers after training (rate of certification, rate of certification after follow-up training, etc.); that is, based on evaluation of interviews (taped or monitored), determination that the interviewer is ready to work on the study. See Interviewer Recruitment, Selection, and Training for further discussion.
  - 3.2.2 Verification of coded questionnaires (rate of verification failures).

3.3 Generate indicators of quality for each assessment, based on baselines established in quality planning (Guideline 2), and create reports on performance and quality assessments, which can be used for both quality monitoring and control (see Guideline 4), and documentation in a quality profile (see Guideline 5).

3.4 Perform quality audits at key points in the survey lifecycle. These generally are structured independent reviews to determine whether activities comply with study and organizational policies and procedures for managing quality. They are intended to identify inefficiencies in processes, and to make recommendations for reducing the cost of quality management and increasing the quality of processes and products. In 3MC surveys, these generally would be done by each individual survey organization, or an independent local auditor.

3.5 Provide documentation for the following aspects of quality assurance:
  3.5.1 Performance and quality assessments.
  3.5.2 Recommended corrective actions and corrective actions taken.
  3.5.3 Updates to baselines.
  3.5.4 Changes to the quality assurance plan.

4. Perform quality control activities.

Rationale

To ensure that standards and requirements are met, it is necessary to monitor study processes and the products produced against predetermined baselines and requirements, and to continuously evaluate whether processes are stable and quality requirements are being met (Biemer & Lyberg, 2003; Lyberg & Stukel, 2010). This may lead to recommendations for preventing or minimizing error or inefficiencies, updates to the quality management plan (see Guideline 2), and suggestions for improving standards and best practices. The result is continuous process, through improved quality assurance (see Guideline 3) and improved quality monitoring and control improvement (Biemer & Lyberg, 2003; Lyberg & Stukel, 2010; Morganstein & Marker, 1997).

As indicated in Figure 3, quality control is closely linked to quality assurance, and the outputs of each feed into the other.

Procedural Steps

4.1 Perform quality monitoring and control activities as outlined in the quality management plan, such as (see Appendix A for examples):
  4.1.1 Monitor process quality indicators (see Guideline 2).
  4.1.2 Analyze and report on results of quality assurance activities, such as interviewer training certification, data entry verification, checking that a process met specifications, etc.
  4.1.3 In accordance with the quality management plan (see Guideline 2), generate charts and graphs to monitor processes. Examples of such tools are (Aitken et al., 2003):
   - Pareto chart
   - Statistical process control chart

4.2 Perform process analysis if quality requirements are not being met (see Guideline 2). For example, the European Social Survey (ESS) closely monitors the survey process and collects paradata and other types of auxiliary data using contact forms.

4.3 Determine whether there is a need to:
- 4.3.1 Recommend corrective actions.
- 4.3.2 Modify the process improvement plan.
- 4.3.3 Modify the quality management plan.

4.4 Provide documentation for the following aspects of quality control:
- 4.4.1 Performance and quality assessments.
- 4.4.2 Recommended corrective actions and corrective actions taken.
- 4.4.3 Updates to baselines.
- 4.4.4 Changes to the quality management and quality assurance plans.

Lessons learned

4.1 Some organizations have used quality control techniques to monitor survey data collection processes and adapt study designs when quality goals are not met. This is known as adaptive or responsive survey design (Groves & Heeringa, 2006). For example, the National Survey of Family Growth (NSFG) uses paradata collected by interviewers to make judgments about the likelihood that individual sample cases will become respondents. By building predictive response propensity models using paradata, it is possible to estimate the probability that the next call on a sample case will produce an interview. Such information is used in a responsive design framework to reduce nonresponse in NSFG (Groves & Heeringa, 2006). See also Data Collection: General Considerations for further discussion on responsive design.

4.2 The European Social Survey (ESS) has developed a solid and continuously improving machinery for planning and implementing the survey (Pennell et al., 2017). It closely monitors the survey process, collects various types of paradata and other auxiliary data using contact forms (Stoop et al., 2010), and documents the paradata for each wave of the survey. As mentioned by Pennell et al. (2017), recent discussions on quality control of the ESS have concerned the issue of more timely interventions when countries have implementation problems.

5. Create a quality profile

Rationale

A quality profile (also known as a quality report) synthesizes information from other sources, documenting survey methodology used throughout the survey lifecycle, providing indicators of process and data quality (sampling and nonsampling errors), corrective actions taken, lessons learned, and recommendations for improvement and further research. It provides the user all information available to help assess data quality in terms of fitness for intended use, total survey error, and other factors discussed in this set of guidelines. See Defeo & Juran (2010) for an example of guidelines for such reports, Eurostat (2003a), Eurostat (2003b), and United States Bureau of the Census (1998) for examples of quality profiles, and Appendix A for examples from other sections of these guidelines.

Procedural Steps

5.1 Document procedures and methodology used for key stages or processes in the lifecycle (see Appendix B). For example, for sample design this would include:
- 5.1.1 Time dimension of design (e.g., one time cross sectional, fixed or rotating panel)
- 5.1.2 Target and survey population definitions, including inclusion/exclusion criteria.
- 5.1.3 Sampling frame(s) descriptions.
- 5.1.4 Maps and protocol used in field listing.
- 5.1.5 Description of all stages of selection, including sample sizes, stratification, clustering, oversampling and number of replicates fielded at each stage.
- 5.1.6 Documentation of procedures to determine probabilities of selection and weights for each stage of selection.
- 5.1.7 Tables of the precision of the estimates of key survey statistics.
- 5.1.8 (If necessary), descriptions of substitution procedures.

For each process documented, this should include
- 5.1.9 Quality assurance procedures.
- 5.1.10 Quality control procedures.
- 5.1.11 Corrective actions taken.

5.2 Provide key indicators of quality for all dimensions of quality (see Defeo & Juran (2010) and Appendix A), some of which can be collected during data collections, others afterwards. They include:
- 5.2.1 Comparability.
- 5.2.2 Coherence.

5.2.3 Relevance.
5.2.4 Accuracy (see Quality Framework), including:
- Measurement error
- Processing error
- Coverage error
- Sampling error
- Nonresponse error
- Adjustment error

5.2.5 Timeliness and punctuality.
5.2.6 Accessibility.
5.2.7 Interpretability.

5.3 Document lessons learned and make recommendations for improvement in studies of the same design, and, if possible, make recommendations for methodological research that could inform design of similar studies in the future. Such information would be useful for the study's coordinating center and national survey agencies, as well as for researchers and organizations interested in conducting similar studies.

Lessons learned

5.1 The ESS provides a well-documented quality reports, which includes fieldwork procedures, fieldwork results and analysis of different error sources: https://www.europeansocialsurvey.org/docs/round6/methods/ESS6_quality_report.pdf

5.2 In 3MC surveys, a quality profile is often created for each study location. For example, the Survey of Health, Ageing, and Retirement in Europe (SHARE) Compliance Profiles documented a set of quality control indicators, on which each country is evaluated (Malter and Börsch-Supan, 2014). Alcser, Benson, and Guyer (2011) discuss pilot testing the SHARE Quality Profile.

Appendix A

The following table lists recommendations from other sections in these guidelines that are related to the dimensions of quality. Also included are examples of indicators of quality adapted from Eurostat's standard quality indicators (Eurostat, 2005).

Quality Dimension	Guidelines
Comparability To ensure as much as possible, that observed data from different countries or cultures are comparable (equivalent). Indicators: Time ▪ The differences, if any, in concepts and methods of measurements between last and previous reference period ▪ A description of the differences, including an assessment of their effect on the estimates Geographical ▪ All differences between local practices and national standards (if such standards exist) ▪ An assessment of the effect of each reported difference on the estimates Domains ▪ A description of the differences in concepts and methods across study countries (e.g., in classifications, statistical methodology, statistical population, methods of data manipulation, etc.) ▪ An assessment of the magnitude of the effect of each difference	Establish minimum criteria for inclusion in a cross-national survey dataset, if applicable, as follows: Minimize the amount of undue intrusion by ensuring comparable standards when appropriate (based on differences in local survey contexts) for informed consent and resistance aversion effort, as well as other potentially coercive measures such as large respondent incentives (see Ethical Considerations and Data Collection: General Considerations). Define comparable target populations and verify that the sampling frames provide adequate coverage to enable the desired level of generalization (see Sample Design). Minimize the amount of measurement error attributable to survey instrument design, including error resulting from context effects, as much as possible (see Questionnaire Design, Instrument Technical Design and Paradata and Other Auxiliary Data). Minimize or account for the impact of language differences resulting from potential translations (see Questionnaire Design, Translation and Adaptation). Minimize the effect interviewer attributes have on the data through appropriate recruitment, selection, and case assignment; minimize the effect that interviewer behavior has on the data through formal training (see Interviewer Recruitment, Selection, and Training and Paradata and Other Auxiliary Data). Identify potential sources of unexpected error by implementing pretests of translated instruments or instruments fielded in different cultural contexts (see Pretesting and Paradata and Other Auxiliary Data). Reduce the error associated with nonresponse as much as possible (see Data Collection: General Considerations for a discussion of nonresponse bias and see Paradata and Other Auxiliary Data for nonresponse error reduction (Section A)). Minimize the effect that coder error has on the data through appropriate coder training (see Data Processing and Statistical Adjustment).

Quality Dimension	Guidelines
(Comparability)	If possible, provide a crosswalk between survey instruments fielded at different times or for different purposes, but using the same questions, to facilitate analysis and post-survey quality review (see Data Harmonization).
Coherence To ensure that the data can be combined with other statistical information for various, secondary purposes. Indicators: ▪ A description of every pair of statistics (statistical unit, indicator, domain, and breakdown) for the survey(s) that should be coherent ▪ A description of any of the differences that are not fully explained by the accuracy component. ▪ A description of the reported lack of coherence for specific statistics	Create a clear, concise description of all survey implementation procedures to assist secondary users. Study Design and Organizational Structure lists topics which should be included in the study documentation; there are also documentation guidelines within each set of guidelines for each stage of the survey lifecycle. Provide data files in all the major statistical software packages and test all thoroughly before they are made available for dissemination (see Sample Design and Data Dissemination). Designate resources to provide user support and training for secondary researchers (see Data Dissemination). See Data Harmonization for a discussion of the creation of common measures of key economic, political, social, and health indicators.
Relevance To ensure that the data meet the needs of the client or users. Indicators: ▪ A description of clients and users ▪ A description of users' needs (by main groups of users) ▪ An assessment of user satisfaction	Clearly state the study's goals and objectives (see Study Design and Organizational Structure). Conduct a competitive bidding process to select the most qualified survey organization within each country or location (see Tenders, Bids, and Contracts). While designing the questionnaire, ensure all survey questions are relevant to the study objectives (see Questionnaire Design). Construct the data file with a data dictionary of all variables in the selected element data file, with all variable names and an accompanying description which are relevant to the study objectives (see Instrument Technical Design).
Accuracy To ensure that the data describe the phenomena they were designed to measure. This can be assessed in terms of Mean Square Error (MSE).	Pretest all the versions of the survey instrument to ensure that they adequately convey the intended research questions and measure the intended attitudes, values, reported facts and/or behaviors (see Pretesting).

Quality Dimension	Guidelines
(Accuracy) Indicators: Measurement error: ▪ A description of the methods used to assess measurement errors (any field tests, reinterviews, split sample experiments, or cognitive laboratory results, etc.) ▪ A description of the methods used to reduce measurement errors ▪ Average interview duration ▪ An assessment of the effect of measurement errors on accuracy Processing Error: ▪ A description of the methods used to reduce processing errors ▪ A description of the editing system ▪ The rate of failed edits for specific variables. ▪ The error rate of data entry for specific variables and a description of estimation methodology ▪ The error rate of coding for specific variables and a description of the methodology followed for their estimation ▪ A description of confidentiality rules and the amount of data affected by confidentiality treatment Coverage error: ▪ A description of the sampling frame ▪ Rates of over-coverage, under-coverage, and misclassification broken down according to the sampling stratification ▪ A description of the main misclassification and under- and over-coverage problems encountered in collecting the data ▪ A description of the methods used to process the coverage deficiencies ▪ Coefficients of variation of estimates and a description of the method used to compute them (including software) ▪ An assessment of resulting bias due to the estimation method	In order to reliably project from the sample to the larger population with known levels of certainty/precision, use probability sampling (see Sample Design). Provide a report on each variable in the dataset of selected elements to check correct overall sample size and within stratum sample size, distribution of the sample elements by other specific groups such as census enumeration areas, extreme values, nonsensical values, and missing data (see Sample Design). If possible, assess validity of survey estimates by looking at the differences between the study estimates and any available “true” or gold standard values (see Paradata and Other Auxiliary Data). Use paradata to study and reduce different types of survey errors (See Paradata and Other Auxiliary Data). See Statistical Analysis for discussion of models most often used in 3MC analyses and their relevant assumptions.

Quality Dimension	Guidelines
(Accuracy) Sampling error: ▪ Type of sample design (stratified, clustered, etc.) ▪ Sampling unit at each stage of sampling ▪ Stratification and sub-stratification criteria ▪ Selection schemes ▪ Sample distribution over time ▪ The effective sample size ▪ Coefficients of variation of estimates and a description of the method used to compute them (including software) ▪ An assessment of resulting bias due to the estimation method Nonresponse error: ▪ Unit nonresponse rate ▪ Identification and description of the main reasons for nonresponse (e.g., non-contact, refusal, unable to respond, non-eligible, other nonresponse) ▪ A description of the methods used for minimising nonresponse ▪ Item nonresponse rates for variables ▪ A description of the methods used for imputation and/or weighting for nonresponse ▪ Variance change due to imputation ▪ An assessment of resulting bias due to nonresponse Model assumptions error: ▪ A description of the models used in the production of the survey's statistics ▪ A description of assumptions used on which the model relies ▪ A description of any remaining (unaccounted for) bias and variability which could affect the statistics	

Quality Dimension	Guidelines
Timeliness and punctuality To ensure that the data are available for analysis when they are needed. Indicators: ▪ The legal deadline imposed on respondents ▪ The date the questionnaires were sent out ▪ Starting and finishing dates of fieldwork ▪ Dates of processing ▪ Dates of quality checks ▪ The dates the advance and detailed results were calculated and disseminated ▪ If data is transmitted later than required by regulation or contract, the average delay in days or months in the transmission of results with reference to the legal deadline ▪ If data are transmitted later than required by regulation or contract, the reasons for the late delivery and actions taken or planned for the improving timeliness	Time data collection activities appropriately (see Data Collection: General Considerations, Pretesting, and Paradata and Other Auxiliary Data). Create a study timeline, production milestones, and deliverables with due dates (see Study Design and Operational Structure).

Quality Dimension	Guidelines
Accessibility To ensure that the data can easily be obtained and analyzed by users. Indicators: ▪ A description of how to locate any publication(s) based on analysis of the data ▪ Information on what results are sent to reporting units included in the survey ▪ Information on the dissemination scheme for the results ▪ A list of variables required but not available for reporting ▪ Reasons why variables are not available	Save all data files and computer syntax from the preferred statistical software package needed during sample design process in safe and well labeled folders for future reference and use (see Sample Design). Document how paradata are collected and the steps used to construct the paradata-based indicators (see Paradata and Other Auxiliary Data) Establish procedures early in the survey lifecycle to insure that all important files are preserved (see Data Dissemination). Test archived files periodically to verify user accessibility (see Data Dissemination). Create digitized versions of all project materials whenever feasible (see Data Dissemination). Produce and implement procedures to distribute restricted-use files, if applicable (see Data Dissemination).
Interpretability To ensure that supplementary metadata and paradata are available to analysts. Indicator: • A copy of any methodological documents relating to the statistics provided	At the data processing stage of the study, create a codebook that provides question-level metadata matched to variables in the dataset. Metadata include variable names, labels, and data types, as well as basic study documentation, question text, universes (the characteristics of respondents who were asked the question), the number of respondents who answered the question, and response frequencies or statistics (see Sample Design, Data Processing and Statistical Adjustment and Paradata and Other Auxiliary Data). Collect and make available process data collected during data collection, such as timestamps, keystrokes, and mouse actions (“paradata”) (see Instrument Technical Design and Paradata and Other Auxiliary Data).

Appendix B

The following table summarizes recommended elements of process quality management relevant to each chapter in these guidelines. These are meant to reflect quality management at two levels: (1) the central coordinating center; and (2) the individual country's research organization level. It is not meant to convey that all elements listed should be part of a study's design, but to provide examples and to help guide the development of specifications for quality management for a study.

If possible, the study's quality profile (quality report) would include a summary of each organization's performance, based on standardized quality indicators. It also would include lessons learned and recommendations for improvement.

Where possible, examples are taken from specific sections of these guidelines, based on the stages of the survey lifecycle. Not all stages of the lifecycle have specific measures for monitoring and controlling quality. Even without clear individual rates or measures of quality, however, there often may be reports on quality assurance activities that facilitate assessing quality.

Guidelines Chapter	Quality Planning and Assurance – Inputs and Activities	Quality Monitoring and Control –Measures and Reports	Elements of Quality Profile
Study Design and Operational Structure	**<u>Inputs</u>** ▪ Study goals and objectives ▪ Country-specific legislation on conducting survey research ▪ Leadership, roles, and responsibilities ▪ Timeline ▪ Deliverables ▪ Quality standards ▪ Budget **<u>Activities</u>** ▪ Create framework and structure of responsibilities and tasks ▪ Arrange regular meetings of working group	▪ Monitor paradata-derived quality measure indicators ▪ Monitor budget, costs, and timeline for each country	▪ Study goals and objectives ▪ Documentation and formatting requirements ▪ All study implementation procedures ▪ Documentation of modifications to study protocol ▪ Summary of each organization's performance

Guidelines Chapter	Quality Planning and Assurance – Inputs and Activities	Quality Monitoring and Control –Measures and Reports	Elements of Quality Profile
	and team leaders ▪ Develop communication flowchart ▪ Determining the study's quality standards, then implement them throughout the research process ▪ Develop quality management plan and identify quality profile elements ▪ Assess quality indicators (i.e., paradata-derived indicators) at each stage, and finally make appropriate changes to repeat the cycle of Plan-Do-Check-Act ▪ Implement a certification process to check study design and quality standards ▪ Consider site visits to all countries to monitor or support the implementation of quality standards ▪ Monitor costs in order to avoid overruns ▪ If and where possible, incorporate methodological research		
Ethical Considerations	**Inputs** ▪ Standards for ethical and scientific conduct ▪ Local and national human subject regulations and legislation ▪ Ethical guidelines in project management and human resource management ▪ Voluntary informed consent protocol and procedures ▪ Procedures for ethics training of project staff ▪ Comprehensive plan for protection of confidentiality **Activities** ▪ Review and apply ethical standards, best practices, and relevant regulations and legislation in designing study and collecting and disseminating survey data ▪ Develop and apply knowledge of local	▪ Report on staff completion of ethics training ▪ Review the implementation of informed consent procedures (percent of cases reviewed, percent of cases failing to follow procedures, actions taken, etc.) ▪ Report on interview falsification (percent of cases reviewed, percent of reviewed cases falsified, subsequent actions taken, etc.) ▪ Report on any actual or potential breaches of confidentiality, security, or other adverse event, including any resulting changes to study protocol ▪ Report on any failures of statistical disclosure control	▪ Description of voluntary consent and confidentiality procedures ▪ Copies of materials provided to respondents as part of informed consent process ▪ Summary of respondent burden assessment ▪ Description of ethics training for project staff ▪ Summary of ethics committee review ▪ Summary of review of recorded interviews regarding the implementation of informed consent procedures ▪ Summary of falsification findings ▪ Summary of any reported actual or potential breaches of confidentiality ▪ Description of disclosure analysis

Guidelines Chapter	Quality Planning and Assurance – Inputs and Activities	Quality Monitoring and Control –Measures and Reports	Elements of Quality Profile
(Ethical Considerations in Surveys)	customs and norms relevant for designing culturally-sensitive survey protocols ▪ Pretest consent protocol and forms to ensure comprehension ▪ Translate and adapt consent protocols and forms according to best practices for translation ▪ Assess respondent burden (overall and by subgroup, if appropriate) ▪ Train project staff on ethics ▪ Have project staff sign pledge of confidentiality ▪ Complete ethics review submission and maintain documentation of submission materials ▪ Review recorded interviews and monitoring to assure adherence to informed consent procedures ▪ Monitor implementation of confidentiality protocols and procedures ▪ Perform audits to determine adherence to confidentiality protocols and procedures ▪ Securely store signed pledges of confidentiality and consent forms ▪ Maintain records of all ethics review committee correspondence ▪ Recontact a sample of cases for each interviewer to verify that screening and interview procedures were appropriately followed ▪ Conduct verification to detect possible interview falsification ▪ Use analyses of paradata to identify unusual variable distributions for one or more interviewers compared to the overall distribution ▪ Conduct disclosure analysis		methods and summary of findings

Guidelines Chapter	Quality Planning and Assurance – Inputs and Activities	Quality Monitoring and Control –Measures and Reports	Elements of Quality Profile
	▪ Investigate any deviation from ethical protocols and take appropriate action to address the situation		
Tenders, Bids, and Contracts	**Inputs** ▪ Type of contract offered ▪ Study specifications ▪ Minimum quality requirements and evaluation criteria for bids **Activities** ▪ Prepare tender based on study specifications ▪ Conduct competitive bidding process within each country ▪ Evaluate bids and select a survey organization in each country ▪ Consider re-releasing the tender if no bidding survey organization can meet the requested quality standards ▪ Define progress approval points throughout the research process ▪ Develop a quality management plan ▪ Develop quality control and quality assurance procedures	▪ Report on evaluation scores of bidding organizations	▪ Summary of process of evaluating and selecting bidding organizations
Sample Design (Sample Design)	**Inputs** ▪ Target and survey population descriptions ▪ Sampling frame(s), definitions, including definitions of strata and sampling units, and any updating of the frame that was needed ▪ Desired level of precision overall and for specific subgroups ▪ Sample size based on specified levels of precision ▪ Selection procedure(s) and estimates of probabilities of selection at each stage ▪ Field listing standard procedures and	▪ Estimate coverage error ▪ Report on percentage of duplicate and ineligible sampling units on the sampling frame(s) ▪ Produce tables/charts of paradata indicators that serve as proxies of survey costs and errors ▪ Alter the survey design during data collection to minimize costs and errors in a responsive design framework ▪ Produce frequency tables for key variables from the frame of sampling	▪ Time dimension of design (e.g., one time cross sectional, fixed or rotating panel) ▪ Target and survey population definitions, including inclusion/ exclusion criteria ▪ Sampling frame(s) descriptions ▪ Examples of maps and protocol used in field listing ▪ Description of all stages of selection, including sample sizes, stratification, clustering, oversampling and number

Guidelines Chapter	Quality Planning and Assurance – Inputs and Activities	Quality Monitoring and Control –Measures and Reports	Elements of Quality Profile
	minimum requirements of field listers ▪ Unique, sample identification codes for each selected sampling unit ▪ Data dictionary of selected elements and sampling units with descriptive and distinct variable names and labels **Activities** ▪ Produce, update and/or clean sample frame(s), as needed ▪ Calculate sample size ▪ Implement selection procedure(s) ▪ Create a unique, sample identification code for each selected element or unit ▪ Arrange regular meetings of working group, project manager and sampling statistician ▪ Conduct responsive design plans to minimize survey costs and errors	units	of replicates fielded at each stage ▪ Documentation of procedures to determine probabilities of selection and weights for each stage of selection ▪ Tables of the precision of the estimates of key survey statistics ▪ (If necessary), descriptions of substitution procedures

Questionnaire Design	**<u>Inputs</u>** ▪ Research question ▪ Review of literature and any relevant studies to identify useful material ▪ Documentation templates ▪ Documentation of origin of any existing questions or materials to be considered for re-use **<u>Activities</u>** ▪ Create cross-cultural and cross-competence development team, providing briefing, training, and tools as relevant ▪ Determine design approach ▪ Create analysis plan relating constructs, indicators and question topics ▪ Implement design steps ▪ Determine appropriate methods to assess the quality of questions ▪ When possible, use wording experiments to decide between different candidate question wordings ▪ Version control procedures are necessary whenever a source questionnaire is modified across time	▪ Description of the questionnaire design procedures	▪ Report on modifications made to questions at different stages ▪ Document different versions of questionnaires if applicable
Adaptation	**<u>Inputs</u>** ▪ Source questionnaires and any materials which might be adapted ▪ Translated questionnaires and any materials which might be adapted ▪ Documentation templates as relevant ▪ Guidelines on adaptation goals and more common forms ▪ Briefing and training of team as necessary ▪ Delivery schedule and required outputs		

	Activities ▪ Determine stage(s) at which adaptation is possible ▪ Create adaptation team with skills suited for whichever stage(s) are envisaged ▪ Make adaptation proposals with documented justifications ▪ Conduct external review of adaptation proposals and their documentation ▪ Test adaptations for targeted population(s) and revise as relevant ▪ Adjudicate/sign-off on adaptation decisions and finalize documentation		
Translation	**Inputs** ▪ Source questionnaire and any material to be translated ▪ Guidelines and stipulations on procedures to be followed and on outputs required (e.g., need for documentation on decisions) ▪ Templates for translation development, as relevant ▪ Delivery schedule including any further refinements proposed that relate to translation (procedures such as language harmonization, adaptation, pretesting and any required adjudication steps ▪ Procedure to monitor performance as appropriate **Activities** ▪ Create translation team, briefing, training and monitoring as relevant. ▪ Produce draft translations, checking translator output at an early stage of production ▪ Maintain documentation at each stage ▪ Review and adjudicate translations	▪ Draft translation review report	▪ Documentation of translation review process

	▪ Pretest translations Repeat any translation refinement step as needed		
Instrument Technical Design	**Inputs** ▪ Instrument specification guidelines ▪ Comprehensive design evaluation plan, including goals, evaluation techniques, and timeline ▪ Quality assurance metrics (e.g., questionnaire and item timings, review of computer-assisted application audit trails, behavior/event codes) **Activities** ▪ Provide clear instrument specifications and/or data dictionary ▪ Provide culture or language-specific adaptations of design specifications ▪ Develop instrument evaluation procedures Perform and report on design assessments ▪ Review quality assurance metrics reports ▪ Make recommendations for improvement	▪ Collect and report on quality metrics or measures, such as: ▪ Questionnaire length and section and item timings ▪ Audit trails for computerized applications ▪ Behavior codes or event codes based on audio or video recordings of pretests or usability tests ▪ Qualitative analysis of cognitive and usability testing (see Pretesting) ▪ Heuristic evaluation or expert review	▪ Instrument specification guidelines ▪ Procedures for design evaluation ▪ Results of design evaluations ▪ Documentation and results of quality assurance and quality monitoring and control
Interviewer Recruitment and Training	**Inputs** ▪ Recruitment and training timeline ▪ Minimum standards for employment ▪ Study-specific requirements (e.g., gender, language, etc.) ▪ Assessment tests ▪ Minimum interviewer requirements checklist ▪ Criteria for dismissal or follow-up training ▪ Standard certification procedures **Activities** ▪ Establish a checklist of minimum interviewer candidate requirements ▪ Train trainers before they train	▪ Report on training attendance ▪ Report on candidate training certification (including rates) ▪ Report on follow-up training certification (including rates)	▪ Employment criteria ▪ General and study-specific training documentation ▪ Certification procedures ▪ Certification rates for training and follow-up training

	interviewers ▪ Complete checklist during candidate screening ▪ Take attendance during training ▪ At the end of basic interviewer training, evaluate the knowledge of the interviewer candidates ▪ Certify candidates ▪ Dismiss or retrain candidates who fail certification ▪ Maintain written records of results of candidates' certification tests ▪ Track the cost and success rates of different recruitment avenues ▪ Survey interviewer candidates to determine what improvements could be made to the recruitment process ▪ Debrief interviewer trainees to determine how training could be improved		
Pretesting	**Inputs** ▪ Pretesting plan, including pretest goals, evaluation techniques, timeline, and budget ▪ Standard procedures for staff training **Activities** ▪ Provide staff training and certification ▪ Examine the findings of each pretesting technique used and identify the causes of the any problems discovered ▪ Review results from a pilot study, if conducted ▪ Review recordings of focus groups and cognitive interviews for staff errors ▪ Provide retraining as necessary ▪ Test for inter-coder reliability if appropriate ▪ Coordinate the documentation of the pretest across participating countries	▪ Monitor costs and timeline ▪ Monitor staff error rates ▪ Test inter-coder reliability	▪ Pretest procedures documentation ▪ Pretest training documentation ▪ Pretest findings, change recommendations, and changes made ▪ Staff error rates

Data Collection	**Inputs** ▪ Target outcome rates (e.g., response, refusal, noncontact), and completion rates ▪ Target hours per interview ▪ Recontact or reinterview respondents ▪ Percentage of interviewer cases to be verified ▪ Verification questions ▪ Verification of case disposition codes and selected responses ▪ Interviewer performance checklist ▪ Criteria for interviewer dismissal or supplementary training **Activities** ▪ Establish a sample management system ▪ Review paper coversheets and/or questionnaires ▪ Dismiss or retrain interviewers with substandard performance ▪ Collect paradata needed for statistical adjustment	▪ Overall, by key respondent groups and by interviewer, report on: ▪ Screening rates ▪ Eligibility rates ▪ Response rates ▪ Refusal rates ▪ Noncontact rates ▪ Completion rates ▪ Hours per interview ▪ Number of completed interviews ▪ Report on interviewer performance outcomes ▪ Develop a responsive design based on cost/error tradeoffs	▪ Documentation of mode(s) of data collection and the protocol for determining mode(s) to use ▪ Documentation of the sample management system ▪ Study materials ▪ Screening/respondent selection procedures ▪ Number of completed interviews, overall and by mode ▪ Documentation of proxy interview protocol ▪ Documentation of respondent incentives, and interviewer incentive protocol ▪ Documentation of techniques to maximize response (e.g., prenotification, recontact, and refusal conversion protocol) ▪ Outcome rates, overall and by key respondent groups ▪ Dates of data collection ▪ Interviewer monitoring procedures and outcomes ▪ Verification form(s) and outcomes ▪ Any descriptions and outcomes of validation study (e.g., administrative record check against survey data)
Data Harmonization	**Inputs** ▪ Standard codebook specifications ▪ Standard procedures for collecting and producing national data files ▪ Comprehensive plan for harmonization of cross-cultural data files ▪ Procedures to judge the quality of the harmonized outputs ▪ Procedures for testing harmonized files with knowledgeable users ▪ Procedures to modify and update harmonized datasets after public release, if applicable	▪ Report on analytic results ▪ Report on user tests	▪ Documentation of specification and procedures standards ▪ Documentation of conversion and harmonization decisions ▪ Results of user tests

	Activities ▪ Create cross-cultural monitoring team ▪ Periodically review analytic results to allow for changes in harmonization rules ▪ Review end-user test results ▪ Make recommendations for harmonization process improvement		
Data Processing and Statistical Adjustment	**Inputs** ▪ Percent of manually entered questionnaires to be verified ▪ Criteria for data entry staff dismissal or supplementary training ▪ Items to be coded ▪ Coding protocol (manual or automatic) ▪ Percent of manually coded cases to be check coded ▪ Minimum acceptable inter-coder reliability ▪ Data editing protocol ▪ Appropriate statistical software ▪ Appropriate statistical adjustments (e.g., imputation, weights) ▪ Appropriate standard error estimation ▪ Quality control procedures for calculation of statistical adjustments and variance estimation **Activities** ▪ Train data entry and data coding staff ▪ Verify data accuracy ▪ Develop coding scheme(s) ▪ Assess inter-coder reliability ▪ Use data entry tools to perform keying quality checks ▪ Check outliers ▪ Edit data ▪ Continually monitor coding activities ▪ Monitor editing using some key	▪ Report on data entry accuracy rate ▪ Test inter-coder reliability Key process statistics for editing ▪ Edit failure rate ▪ Recontact rate ▪ Correction rate	Data processing ▪ Data coding and data entry training documentation ▪ Evaluation protocol for data coding and data entry staff and outcomes ▪ Items that were coded or re-coded ▪ Coding reliability ▪ Data entry verification protocol and outcomes ▪ Data editing protocol Statistical adjustment ▪ Rationale for assigning sample identification numbers ▪ Calculation of outcome rates (e.g., response, refusal, noncontact), weighted and unweighted ▪ Standard error estimates ▪ Percent item missing data Where applicable: ▪ Imputation method(s) ▪ Generation of weight(s) ▪ Trimming of weight(s) ▪ Scaling of weight(s) ▪ Adjustment(s) for differential nonresponse ▪ Poststratification adjustment(s)

	process statistics ▪ Remove any identifying information from the production data ▪ When possible, use paradata for post-survey adjustments ▪ Assign a second sampling statistician to check the post-survey adjustment methodology and the statistical software syntax of the survey's primary sampling statistician		
Data Dissemination	**Inputs** ▪ a quality compliance protocol ▪ Procedures for testing accessibility of archives with knowledgeable users ▪ Procedures for digitized preservation of files ▪ procedures for assessing disclosure risk to respondents ▪ procedures for distributing restricted-use files if applicable ▪ Procedures for testing files with major statistical packages **Activities** ▪ Create electronic versions of all files ▪ Provide data files in all major statistical software packages ▪ Designate resources to provide user support and training for secondary researchers ▪ Review results of user tests	▪ Data archive test reports	▪ Description and classification of target users and their needs ▪ Results of user satisfaction assessments ▪ Summary of conditions of access to data, accompanying documentation, and user feedback ▪ Distribution reports (dataset requests, Web hits, downloads, etc.)
Paradata	**Inputs** ▪ Study goals and objectives ▪ Survey mode and available paradata ▪ Protocols to collect different types of paradata ▪ Instructions on how to construct paradata-derived indicators ▪ Procedures to conduct responsive design using paradata ▪ Plans on how to use paradata for	▪ Paradata collection report ▪ Documentation on the construction of paradata-derived indicators ▪ Document the coding procedure, if there is any, for interviewer-generated paradata ▪ Document clearly how paradata is used to monitor and intervene the data collection process.	▪ Documentation on how paradata can be linked to main survey data ▪ Documentation on the use of paradata in a responsive design ▪ Documentation on the use of paradata in studying different types of errors

	analyzing different types of errors **Activities** ▪ For computer-generated paradata, develop procedures to make sure all programming work as designed ▪ For interviewer-generated paradata, develop clear protocols about how to record paradata for interviewers, and protocols for coders on how to code interviewer-generated paradata in the dataset ▪ Develop quality examination procedures for different types of paradata ▪ Monitor the process of using paradata for analysis ▪ Monitor the process of using paradata in responsive designs		

References

Aitken, A., Hörngren, J., Jones, N., Lewis, D., & Zilhäo, M. J. (2003). *Handbook on improving quality by analysis of process variables.* Luxembourg: Eurostat. Retrieved April 25, 2016, from http://ec.europa.eu/eurostat/documents/64157/4373903/06-Handbook-on-improving-quality-by-analysis-of-process-variables.pdf/b0006e09-1708-4f8e-97e0-6a54d840b92b

Alcser, K., Benson, G., & Guyer, H. (2011). Pilot testing the SHARE Quality Profile. In M. Schröder (Ed.), *Retrospective data collection in the Survey of Health, Ageing and Retirement in Europe. SHARELIFE Methodology* (pp. 37-43). Mannheim, Germany: MEA. Retrieved June 14. 2016, from http://www.share-project.org/uploads/tx_sharepublications/FRB-Methodology_2011_Ch.5.pdf

Anderson, R., Kasper, J., Frankel, M., & Associates (Eds.). (1979). *Total survey error: Applications to improve health surveys*. San Francisco: Jossey-Bass.

Australian Bureau of Statistics. (2009). *Draft report on quality assurance frameworks*. Retrieved May 7, 2010, from http://unstats.un.org/unsd/dnss/qaf/qafreport.htm

Biemer, P. P. (2014). *Toward a total error framework for Big Data*. Presentation at the American Association for Public Opinion Research (AAPOR) 69th Annual Conference.

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

Brackstone, G. (1999). Managing data quality in a statistical agency. *Survey Methodology, 25*(2), 1-23.

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

Cochran, W. G. (1977). *Sampling techniques*. New York, NY: John Wiley & Sons.

Couper, M. P. (1998). *Measuring survey quality in a CASIC environment*. Paper presented at the Proceedings of American Statistical Association: Survey Research Methods Section. Retrieved March 27, 2010, from http://www.amstat.org/sections/srms/proceedings/papers/1998_006.pdf

Defeo, J. A., & Juran, J. M. (2010). *Juran's quality handbook: The complete guide to performance excellence* (6th ed.). New York, NY: McGraw-Hill.

Eurostat. (2003a). *Methodological documents – Definition of quality in statistics* (Report of the Working Group Assessment of Quality in Statistics, item 4.2). Luxembourg: Eurostat.

Eurostat. (2003b). *Methodological documents – Standard report* (Report of the Working Group Assessment of Quality in Statistics, item 4.2B). Luxembourg: Eurostat. Retrieved Jan 21, 2016, from http://www.unece.org/fileadmin/DAM/stats/documents/2000/11/metis/crp.2.e.pdf

Eurostat. (2005). *Standard quality indicators*. Report to Metadata Working Group 2005, Doc. ESTAT/02/Quality/2005/9/Quality Indicators. Luxembourg: Eurostat. Retrieved May 7, 2010, from https://circabc.europa.eu/webdav/CircaBC/ESTAT/prodcom/Library/25-Quality/quality_reports/reference_documents/STANDARD_QUALITY_INDICATORS.pdf

Groves, R. M. (1989). Survey errors and survey costs. New York, NY: John Wiley & Sons.

Groves, R. M., Fowler, F. J. Jr., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2009). *Survey methodology* (2nd ed.). Hoboken, NJ: John Wiley & Sons.

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*, 439–457.

International Monetary Fund (IMF). (2003). *Data quality assessment framework.* Retrieved May 7, 2010 from http://dsbb.imf.org/images/pdfs/dqrs_Genframework.pdf

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*. Washington D.C.: U.S. National Center for Education Statistics.

Lyberg, L. (2012). Survey quality. *Survey Methodology*, *38*(2), 107–130. Retrieved from www.statcan.gc.ca/pub/12-001-x/12-001-x2012002-eng.htm

Lyberg, L. E., & Biemer, P. P. (2008). Quality assurance and quality control in surveys. In E. D. de Leeuw, J. J. Hox, & D. A. Dillman (Eds.), *International handbook of survey methodology*. New York/London: Lawrence Erlbaum Associates.

Lyberg, L. E., Biemer, P. P., Collins, M., de Leeuw, E. D., Dippo, C., Schwarz, N., & Trewin, D. (Eds.). (1997). *Survey measurement and process quality*. New York, NY: John Wiley & Sons.

Lyberg, L. E., Bergdahl, M., Blanc, M., Booleman, M., Grünewald, W., Haworth, M., … van Brakel, R. (2001). *Summary report from the Leadership Group (LEG) on quality*. Retrieved March 1, 2010, from http://www.istat.it/it/files/2011/11/LEG-on-quality.pdf

Malter, F., and Börsch-Supan, A. (2014). SHARE compliance profiles – Wave 5. Retrieved from http://www.share-project.org/fileadmin/pdf_documentation/SHARE_Wave5_ComplianceProfiles_v11.pdf

Mneimneh, Z., Lyberg, L., Sharma, S., Vyes, M., & Sathe, D. B. (2017). Case studies in monitoring interviewers' behaviors in international and multinational surveys. In T. P. Johnson, B-E. Pennell, I. A. L. Stoop, & B. Dorer (Eds.), *Advances in comparative survey methods: Multicultural, multinational and multiregional contexts (3MC).* New York, NY: John Wiley & Sons.

Morganstein, D., & Marker, D. A. (1997). Continuous quality improvement in statistical agencies. In L. E. Lyberg, P. P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz, & D. Trewin (Eds.), *Survey measurement and process quality* (pp. 475-500). New York: John Wiley & Sons.

Murray, T. S., Kirsch, I. S., & Jenkins, L. B. (1998). *Adult literacy in OECD countries: Technical report on the first International Adult Literacy Survey.* US Government Printing Office, Superintendent of Documents, Mail Stop: SSOP, Washington, DC 20402-9328. Web site: https://nces.ed.Gov

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. West (Eds.), *Total Survey Error in Practice*. New York, NY: John Wiley & Sons.

Project Management Institute. (2004). *A guide to the project management body of knowledge (PMBOK® guide)* (3rd ed.). Newtown Square, PA: Project Management Institute, Inc.

Smith, T. W. (2011a). Refining the total survey error perspective. *International Journal of Public Opinion Research*, (23), 464-484.

Smith, T. W. (2011b). The report of the international workshop on using multi-level data from sample frames , auxiliary databases , paradata and related sources to detect and adjust for nonresponse bias in surveys. *International Journal of Public Opinion Research*, *23*(3), 389-402.

Smith, T. W. (2017). Improving cross-national/cultural comparability using the total survey error paradigm. In T. P. Johnson, B-E. Pennell, I. A. L. Stoop, & B. Dorer (Eds.), *Advances in comparative survey methods: Multicultural, multinational and multiregional contexts (3MC).* New York, NY: John Wiley & Sons.

Statistics Canada. (2002). *Statistics Canada quality assurance framework*. Ottawa: Statistics Canada. Retrieved May 7, 2010, from http://www.statcan.gc.ca/pub/12-586-x/12-586-x2002001-eng.pdf

Stoop, I., Billiet, J., Koch, A., & Fitzgerald, R. (2010). *Improving survey response: Lessons learned from the European Social Survey*. New York: John Wiley & Sons.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge University Press.

United States Bureau of the Census. (1998). *Survey of income and program participation: SIPP quality profile* (Working Paper No. 230) (3rd ed.). Washington, DC: U.S. Bureau of the Census. Retrieved March 27, 2010, from http://www.census.gov/sipp/workpapr/wp230.pdf

Tupek, A. R. (2006). Definition of data quality. *Census Bureau Methodology & Standards Council. Census Bureau*, *6*.

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).

Further Reading

Alwin, D. F. (2016). Survey data quality and measurement precision. In C. Wolf, D. Joye, T. W. Smith, & Y. Fu (Eds.), *The SAGE handbook of survey methodology*. Thousand Oaks, CA: Sage.

Biemer, P. P. (2016). Total survey error: A paradigm for survey methodology. In C. Wolf, D. Joye, T. W. Smith, & Y. Fu (Eds.), *The SAGE handbook of survey methodology*. Thousand Oaks, CA: Sage.

Lyberg, L. E., & Weisberg, H. F. (2016). Total survey error paradigm in practice. In C. Wolf, D. Joye, T. W. Smith, & Y. Fu (Eds.), *The SAGE Handbook of Survey Methodology*. Thousand Oaks, CA: Sage.

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. West (Eds.), *Total survey error in practice*. New York, NY: John Wiley & Sons.

Smith, T. (2016). Survey standards. In C. Wolf, D. Joye, T. W. Smith, & Y. Fu (Eds.), *The SAGE handbook of survey methodology*. Thousand Oaks, CA: Sage.

Thiessen, V., & Blasius, J. (2016). A fresh look at survey data quality. In C. Wolf, D. Joye, T. W. Smith, & Y. Fu (Eds.), *The SAGE handbook of survey methodology*. Thousand Oaks, CA: Sage.

Ethical Considerations

Julie de Jong, Kristen Cibelli Hibben, and Steve Pennell, 2016
(2010 Version: Kirsten Alcser, Christopher Antoun, Ashley Bowers, Judi Clemens, and Christina Lien)

Introduction

These guidelines focus on ethical concerns arising from the use of human subjects for research in multinational, multicultural, or multiregional surveys, which we refer to as "3MC" surveys. The World Health Organization defines human subject research as the "…systematic collection or analysis of data…in which human beings (i) are exposed to manipulation, intervention, observation, or other interaction with investigators either directly or through alteration of their environment, or (ii) become individually identifiable through investigators' collection, preparation, or use of biological material or medical or other records" (World Health Organization, 2009).

There is no lack of source material on ethical guidelines for human subject research (see Singer (2008) for a review). International efforts to protect the rights of human subjects involved in research are predominately rooted in the ethical principles established by the Declaration of Helsinki. The Declaration of Helsinki (World Medical Association, 1964), originally adopted by the World Medical Association in 1964 and most recently revised in 2008, defines the ethical responsibilities of physicians to their patients and to the subjects of biomedical research. It asserts the principle of informed consent from research subjects and the precedence of individual subjects' well-being over any anticipated benefits of the research to science and society. The principles in the Declaration of Helsinki have been extended to include social science human subject research.

The Belmont Report, issued in 1979 by the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, underlies regulation of human subjects research in the United States. It advances three key fundamental ethical principles for the conduct of all research involving human subjects: respect for persons, beneficence, and justice (United States, 1978). Application of these principles requires careful consideration of the selection of research subjects, informed consent, and an assessment of potential risks and benefits to research subjects and to society. The Belmont Report has influenced research ethics in many parts of the world.

Professional organizations, such as the American Association for Public Opinion Research (AAPOR), the World Association for Public Opinion Research (WAPOR), the European Society for Market Research (ESOMAR), and the International Statistical Institute (ISI), have also developed ethical codes and guidelines for their members. The ethical codes of these professional organizations define the norms and responsibilities for survey researchers in relation to respondents, as well as to clients or sponsors, the public, and other researchers (Singer, 2008).

In addition to these self-regulatory measures, many countries have legislation in place that affects human subject research (e.g., data protection legislation and requirements for ethics review boards). Whether working in familiar surroundings or in new contexts, researchers must make sure they are informed about, and comply with, relevant legislation. When working in other countries or locations, researchers may need to comply not only with local requirements, pertaining to the place where they are collecting data, but also with their own country's requirements. A compilation of laws, regulations and guidelines from over 100 countries has been prepared by the US Office for Human Research Protections: http://www.hhs.gov/ohrp/international/index.html.

As might be expected, there is considerable overlap in the principles contained in the various ethics codes, professional association guidelines, and government regulations. This section attempts to consolidate their common elements, as well as to highlight concerns particular to 3MC surveys, including cross-national variation in laws and regulations relevant to human subject research and cultural differences that affect the conduct of ethical research across cultures.

Beyond professional codes of ethics, is useful to consider the ethical framework or philosophy that guides the research, engendering sensitivity toward research participants (Hesse-Biber, 2010; Patton, 2002) and their social contexts (Markham, 2005). And, it is important to note that the way in which researchers, as well as survey interviewers, perceive ethics, is situated in their own cultural context. Hesse-Biber (2010) suggests that choosing a research problem is itself an ethical decision.

It is important to recognize that researchers may confront tradeoffs between ethical principles and that there is no one ethical principle that overrides all others. For example, maintaining sensitivity to cultural differences by having other family members present during the interview may conflict with ethical obligations to protect confidentiality and to minimize error in respondent reporting. For further information on the ethical principles presented here, please see the listing of ethics codes, declarations, guidelines, and other resources for researchers conducting 3MC human subject research that is provided in Further Reading.

Guidelines

Goal: To ensure that all members of participating research teams follow widely accepted standards for ethical, professional, and scientific conduct from the design of the study through implementation, dissemination, and reporting.

1. Observe professional standards for managing and conducting scientifically-rigorous research at all stages of the study.

Rationale

Researchers have a responsibility not only to protect participants from the risks of harm associated with participation in the research but also to adhere to ethical management practices and to conduct research that meets the scientific standards of their field. The reader is referred here to other sections in these Guidelines which provide useful guidance on meeting scientific standards for the design, implementation, analysis, and documentation of 3MC surveys.

Procedural steps

1.1 Understand and adhere to the best practices of survey methodology

- 1.1.1 Clearly and objectively lay out the study's major research questions for internal use in guiding the development of the study.
- 1.1.2 Ensure that a survey is the most appropriate method to use to answer the research questions.
- 1.1.3 Adhere to ethical business practices in bidding, contracting, and project management. These include the following:
  - Honestly describing the organization's expertise in a bid.
  - Disclosing if a survey project is being carried out on behalf of multiple clients or is using subcontractors.
  - Meeting contractual obligations.
  - Ensuring agreement by both parties on any changes to contractual obligations.
  - Maintaining good relations between the coordinating center and research organizations involved in the study.
  - For additional detail, see Tenders, Bids, and Contracts.
- 1.1.4 Disclose sources of financial support or relevant relationships that have the appearance of or potential to constitute a conflict of interest.
- 1.1.5 Fulfill ethical responsibilities to employees (e.g., fair hiring practices, an objective performance evaluation process, and a commitment to employee safety). See Guideline 1 of Data Collection: Face-To-Face Surveys for guidance on the survey organization's responsibility to protect the well-being and safety of its interviewing staff.
- 1.1.6 Train staff on the importance of ethics and scientific rigor in research involving human subjects, as discussed in the remainder of this chapter.

1.1.7 Ensure that interviewers are aware of their ethical and responsibilities (e.g., in the United States, interviewers are obligated to report evidence of child abuse).

1.1.8 Instruct interviewers on the limits of their ethical responsibilities (e.g., when they should provide information about local health resources or contact a clinical psychologist or social worker assigned to the project, rather than attempting to provide medical assistance or mental health support services themselves).

1.1.9 Equip staff involved in design, data collection, and analysis with appropriate skills to perform scientifically rigorous research.

1.1.10 Follow best practices in survey design, data collection, and post-survey processing as described in the following chapters:

- Survey Design and Organizational Structure
- Survey Quality
- Tenders, Bids, and Contracts
- Sample Design
- Questionnaire Design
- Adaptation
- Translation
- Instrument Technical Design
- Interviewer Recruitment, Selection, and Training
- Pretesting
- Paradata and Other Auxiliary Data
- Data Collection
- Data Harmonization
- Data Processing and Statistical Adjustment
- Data Dissemination

1.1.11 Employ appropriate tools and methods of analysis.

1.1.12 Make interpretations of research results that are consistent with the data.

1.1.13 Be clear and honest about how much confidence can be placed in the conclusions drawn from the data.

1.1.14 Report research findings, even if they are not in line with the researcher's hypothesis.

1.1.15 Monitor possible ethics violations, such as interviewer falsification or plagiarism, during the design, data collection, and analysis phases.

1.1.16 Consider both cost and error implications of decisions that are made in the design, implementation, and analysis phases of the research study and the relationship that these decisions have with ethical considerations.

1.1.17 When possible, conduct methodological studies to inform understanding of the cost and quality implications of survey design decisions for the benefit of future studies and the scientific research community. Most of the methodological research on ethics and other survey design considerations has been conducted in Western cultures. Additional research is needed in non-Western societies.

1.2 Understand and adhere to relevant professional codes of ethics regarding survey research.

1.2.1 In the United States, the primary organization representing survey researchers is the American Organization for Public Opinion Research (AAPOR). AAPOR obligates members to adhere to its code of ethics (AAPOR, 2015)

1.2.2 There are two international survey professional organizations, each of whom prescribes principles of ethical practices for organization members:

- World Association for Public Opinion Research (2011)
- European Society for Market Research (2008)

1.2.3 See Smith (2007) for a more exhaustive list of existing professional and trade associations and codes of standards

1.3 Observe general standards of scientific conduct as well as standards mandated by study countries themselves.

1.3.1 Countries have different methods to assess adherence to ethical standards.

- In the United States, institutions generally have an Institutional Review Board (IRB) which assesses the protocols proposed for protection of human subjects and is approved by the U.S. Department of Health and Human Services (http://www.hhs.gov).
- Institutions in many other countries are subject to country-specific regulations as well. See the 2017 edition of the International Compilation of Human Research Standards for laws, regulations, and guidelines on human subjects protection in over 100 countries as well as from a number of international and regional organizations: http://www.hhs.gov/ohrp/international/index.html.

1.3.2 When developing an ethical protocol for a cross-cultural survey, consider using the International Organization for Standardization (ISO) standards catalog on the vocabulary for market, opinion, and social research: (ISO, 2016).

1.3.3 Do not engage in scientific misconduct, including:

- Revealing the identity of research participants.
- Generalizing results beyond the study's target population, or otherwise misrepresenting the sample design used to select respondents.

- Plagiarism, falsification, or fabrication in proposing, performing, reviewing research, or in reporting research results.
- Fundraising, selling, or canvassing under the guise of research.

2. Respect and safeguard the rights of free will, privacy, confidentiality, and well-being of research participants, and minimize the burden of study participation to the greatest extent possible, adhering to both ethical and legal obligations toward participants.

Rationale

The social researcher's responsibility to respect the human rights of study participants is universally prescribed in ethics codes and guidelines such as the Declaration of Helsinki (World Medical Association, 1964) and the Belmont Report (United States, 1978), and monitored by ethics review boards in countries where such boards exist. In addition, the collection of accurate data depends upon the cooperation of respondents: individuals are more likely to agree to participate in a study and to give complete, accurate information if they feel that they can trust the research organization. Finally, a positive experience with regard to the research interaction encourages participation in future research.

Procedural steps

2.1 Observe the principles embodied in the Belmont Report. The Belmont Report is based upon three unifying principles for using any human subjects for research: Respect for persons, beneficence, and justice. The United States Department of Health and Human Services uses these three principles to form the basis of their regulations to protect human subjects. The principles are used below to organize the different aspects of ethical obligations that researchers must consider.

2.1.1 Respect for persons: *Protecting the autonomy of all people and treating them with courtesy and respect and allowing for informed consent. Researchers must be truthful and not engage in fraudulent claims*.

- Encourage participation in the research study only in ways that avoid personal harassment, while recognizing appropriate ways to minimize non-response through acceptable means of contact (see Guideline 3). This may include limiting the number of times that an interviewer visits a household to attempt to obtain sample member participation.

- Be respectful and honest with survey respondents (e.g., be honest about the length of the interview, any benefits being offered, and the purpose of the study).
  - Bulmer (2008) points out that it is not always possible "to be completely open to all participants" (p. 154) without overwhelming the listener. Furthermore, the definition of honesty and the way in which honesty is expressed vary according to culture (Berry, Poortinga, Segall, & Dasen, 2002).
- Adapt the study protocol as needed to protect the rights of vulnerable populations -- that is, populations with diminished autonomy resulting from age, cognitive impairment, or imprisonment, such as children, the elderly, prisoners, the mentally impaired, and members of economically and otherwise disadvantaged groups. Use special consent procedures (e.g., obtaining consent from a parent or family member) or other appropriate study modifications. See Guideline 3 for further information about obtaining informed consent.

2.1.2 Beneficence: *The philosophy of "do no harm" while maximizing benefits for the research project and minimizing risks to the research subjects*

- Use existing data whenever possible; do not collect new data unnecessarily.
- Keep respondent burden as low as possible (Bradburn, 1978) by ensuring that each question in the survey maps to a specific research goal, balancing the need for information against the effort that is required to complete additional questions, asking questions in a way that is easy for respondents to answer (see Converse & Presser (1986), Dillman, Smyth, & Christian (2009), and Fowler (1995) for guidance), and, if sensitive or otherwise demanding information is required, devising ways to help respondents provide it without undue burden. For example, part of the interview could be self-administered if there is concern that respondents might be uncomfortable providing responses to an interviewer. See Guideline 4 in Data Collection: Face-to-Face Surveys and Guideline 3 in Data Collection: Telephone Surveys as well as Data Collection: Self-Administered Surveys for a discussion of self-administered modes of data collection.
- Determine whether asking respondents to provide information on specific topics could bring harm or political repercussions to them and do not include questions on those topics.

- Alternatively, disclose the probability and magnitude of a risk of harm and let competent adult participants decide whether to provide the information. Respect for persons means allowing people to choose for themselves while providing extra protection to those with limited autonomy.
- If the information gathered by sensitive questions is necessary for the research goals, consider constructing a series of questions to define a latent construct, rather than asking a direct question.

- Consider carefully whether the requested information may be seen as private, threatening or embarrassing by the population interviewed, and implement techniques to minimize unease.
    - In mental health studies, provisions are often made to provide suitable support for respondents or interviewers who experience emotional distress (for example, some form of emotional or psychological support service or provision of a resource list). A resource list may be made available to all participants, not only those who demonstrate emotional distress. In addition, interviewers in these studies should complete specialized training on how to handle interviewing on sensitive topics.
    - Also, recognize that cultures differ in what topics can be discussed and how they can be discussed.
- Some study designs permit the use of a proxy interview, which is an interview with someone other than the person about whom the survey information is sought, such as the parent or spouse. If the study design allows for a proxy interview, then consider the sensitivity of the requested information and assess whether it would be appropriate to ask a person other than the respondent for sensitive information in a proxy interview.
    - If proxy interviews are used, create and adhere to a clearly defined set of rules concerning who can serve as a proxy respondent**.** Consider whether the use of a proxy interview requires the consent of the target respondent. If the target respondent has indicated any unwillingness to provide information, do not gather the information from a proxy instead. Take care not to affect the relationship between the proxy and the target respondent.

2.1.3 Justice: *Ensuring reasonable, non-exploitative, and well-considered procedures are administered fairly—the fair* *distribution of costs and benefits to potential research participants—and equally.*

- Do not exclude minority groups, native populations, or aboriginal peoples in the sample, unless it is appropriate to do so. Examples include exclusion of respondents living in certain areas of a country because of heightened security concerns and increased risk to interviewers, exclusion of respondents living in very remote areas because of budget constraints, and exclusion of respondents because of language barriers and the prohibitive cost of additional translation and administration of the survey.

2.2 In addition to ethical obligations, consider the legal obligations to research participants. These obligations will differ depending on the country of the researchers, the country of the research participants, and the country from which the source of funding originates.

2.2.1 In the United States, the legal foundation for protection of human subjects of research, including survey respondents, is the Research Act of 1974 (P.L. 93-348, July 12, 1974). This Act led to the development of Regulations of the Protection of Human Subjects of Research, which require universities and other institutions receiving federal funds to establish Institutional Review Boards (IRBs) to safeguard the rights of research volunteers. See the following website at the United States Office of the Federal Register: https://www.federalregister.gov/articles/2015/09/08/2015-21756/federal-policy-for-the-protection-of-human-subjects for upcoming revisions to the U.S. federal policy for the protection of human subjects.

2.2.2 In Canada, all research involving human subjects must adhere to the ethics policy put forth by the Panel on Research Ethics (Government of Canada, 2014).

2.2.3 In Australia, human subjects research is regulated by the Human Research Ethics Committees(http://www.health.gov.au/).

2.2.4 While human subjects research is not regulated by a single entity in the European Union, the EU does have regulations designed to safeguard the confidentiality of personal data (http://www.coe.int).

2.2.5 Many other countries around the world have similar research ethics policies and regulations. It is an individual researcher's responsibility to identify the appropriate policies in the respective country or countries.

2.3 If appropriate, obtain a Certificate of Confidentiality (CoC) or other legal document for protection from the requirement to release the identity of a respondent in a legal proceeding. In the U.S., CoCs are issued by the National Institutes for Health (NIH) and other Department of Health and Human Services (DHHS) agencies and are generally issued for data that are sensitive, such as mental health or sexual or illegal behavior. Certificates of confidentiality are only issued for research projects that:

2.3.1 Collect personally identifiable, sensitive information.
2.3.2 Are approved by an Institutional Review Board (IRB) operating under a Federal Wide assurance (FWA) issued by the DHHS Office of Human Research Protections (OHRP) or with the approval of the FDA.
2.3.3 Are on a topic that is within the HHS health related research mission.
2.3.4 May receive federal funding (not required but issuance is at the discretion of the issuing agency).
2.3.5 Store research data in the United States.
2.3.6 Are allowable under federal regulations.
2.3.7 Make clear to respondents the extent to which confidentiality is protected.

Lessons learned

2.1 When determining a survey project schedule, leave ample time to procure human subjects approval from the necessary institution(s). Depending on the agencies involved, obtaining approval can take many months and delay start dates. This can interfere with the comparative nature of a 3MC survey, if, for example, one country has the necessary ethics approval to begin fieldwork but another country cannot begin until several months later.

2.2 Risk-benefit analyses will differ depending on the research topic. Trauma-focused research, such as that in disaster or conflict regions, is particularly challenging and provides a useful example of special considerations necessary when conducting research on certain topics and/or in certain locations. In trauma-focused research, it is crucial that both respondents and interviewers be adequately protected from both psychological and physical harm, and the definition of “vulnerable population” may need to be expanded in trauma-focused research. “In the context of a disaster, the goals of the research, the benefits from participation, along with auspices and affiliations, must be made clear to potential respondents. That is, any link (or lack thereof) between participating in research and receipt of aid or other benefits should be explicit, and all study materials should clearly state the purpose of the research and list all affiliations” (Pennell et al., 2014, p.124). Recognition of opportunity for increased respondent and interviewer burden is important as well.

2.2.1 Researchers must consider whether respondents are capable of providing voluntary consent in the aftermath of a disaster, necessitating carefully designed consent procedures, based on the context and location of the study.

2.2.2 High profile events can lead to increased respondent burden, with multiple survey requests to individuals from numerous organizations.

2.2.3 After the tsunami in the Indian Ocean in 2004, researchers warned that victims in Sri Lanka may feel pressured to comply with survey requests because of a presumed link of survey participation with humanitarian aid, and appealed to the international community to adequately address the issue of respondent protection in extenuating circumstances (Sumathipala & Siribaddana, 2005). Similarly, interviewers should not deceive respondents about benefits to participation. In a study in India, dishonest interviewers were believed when they told respondents that survey participation would result in new schools, roads, and an electricity supply (Armer & Grimshaw, 1973).

2.2.4 Refusal rates were high in a Haitian study after the 2010 earthquake because respondents had already been interviewed several times and never received the aid or assistance they had expected (Andre & Lusk, 2011).

2.2.5 Interviewer safety and security is crucial in a disaster or conflict zone, and an adequate understanding of current conditions is necessary before fieldwork begins. Additionally, if recruited locally, interviewers themselves may be struggling with the aftermath of the disaster or conflict, as was the case in research on populations after Hurricanes Katrina and Rita in the United States in 2005 (Richardson et al., 2009). In such cases, interviewers should be offered the same mental health referral services that they offer to respondents in need.

2.2.6 See Pennell et al. (2014), Mneimneh et al. (2014), and other chapters in Tourangeau et al. (2014) for further details to consider when conducting research in these populations.

2.3 It is a well-known concern that sharing certain information with the respondent, such as specific research questions the study aims to address, can produce undesirable bias. In such cases, it may be desirable to omit certain specific information, while at the same time, sharing information that is only truthful, and without any deception.

2.3.1 In some contexts, such as surveying in areas of armed conflict, “Researchers need to give careful thought to how the study is introduced in any scripted material and how it is presented by interviewers... From a measurement perspective, affiliating the study with a political party or even an aid agency may influence respondents' answer affecting the validity of the data" (Mneimneh et al., 2014, p. 142-143; see also Mneimneh et al., 2008).

2.3.2 Project Camelot was a U.S. Department of Defense research study designed to evaluate the Chilean masses' potential for revolutionary political action, and to determine the most effective means of counteracting that action. Participating Chilean social scientists were not told that the U.S. Department of Defense was funding the project and would ultimately receive the data. When Chilean researchers learned the facts, the study was cancelled. The image of the U.S. funders and U.S. research suffered greatly (Armer & Grimshaw, 1973).

2.3.3 In a 3MC study in the Middle East conducted by researchers in the U.S. in collaboration with national partners in study site countries, researchers were concerned that respondents would be reluctant to participate if they knew that the study was affiliated with a U.S. institution. Therefore, researchers obtained permission from their university's IRB to omit reference to the IRB in the consent documents, and interviewers introduced the study to respondents as being conducted by the study country partner. However, all participating project members in the study country research organizations and academic institutions had full knowledge of the U.S. collaboration (personal communication, de Jong, 2015).

2.3.4 Beginning in 1961, psychologist Stanley Milgram conducted a study at Yale University (Milgram, 1965). Test subjects were told that they were part of an experiment on punishment and memory, and that they would act as "teachers." The "teacher" subjects were instructed by the experimenter to administer an electric shock to a "learner" if the latter failed to perform as required. Unbeknownst to the subject, the "learner" was one of the research team and deliberately gave many incorrect answers. The subject was ordered by the experimenter to give higher and higher intensity shocks to correct this poor performance. Although in fact no shocks were administered, the majority of subjects believed that they were actually administering electric shocks to the "learner." As a result, subjects experienced distress and tension during the experiment; several even had seizures. The unethical Milgram study was highly criticized after the event, and became a landmark in the effort to develop ethical guidelines for social science research (Groves et al., 2009). However, while attacked from an ethics perspective, the Milgram study made a major contribution to research on obedience in social psychology. This study illustrates how it can be a challenge to balance the goals of science and ethical considerations.

2.4 3MC studies may involve the use of field research methods beyond the survey interview. Participant observation is a field research technique that involves the researcher becoming a trusted, yet temporary, participant in the community under study (Singleton Jr. & Straits, 2005). This temporary membership may lead to feelings of abandonment on the part of the participants. Possible solutions include maintaining honesty with the participants and community as well as providing the researched community with a final copy of the research results in the community's native language (Punch, 1994).

2.5 Regarding respondent burden and privacy, the duration and location of interviews varies among established 3MC surveys.

2.5.1 The Afrobarometer Survey lasts approximately one hour and is usually administered in the respondent's home, although other locations are sometimes used (Afrobarometer Survey). Similarly, the Asian Barometer interview is completed in the respondent's home or workplace (Asian Barometer).

2.5.2 The basic face-to-face portion of the European Social Survey (ESS) takes approximately 60 minutes and is conducted in the respondent's home (http://www.europeansocialsurvey.org/index.php?option=com_content&view=article&id=23&Itemid=318).

2.5.3 The International Social Survey Programme (ISSP) questionnaire consists of 60 questions, not including demographics, and takes approximately 15 minutes to complete (http://www.issp.org/).

2.5.4 The length of the Living Standard Measurement Study Survey (LSMS) varies across participating countries, depending upon the number of modules administered (LSMS, 1996).

2.5.5 The Survey of Health, Ageing and Retirement in Europe (SHARE) is completed in the respondent's home; it takes approximately 80 minutes to administer to a single-family household, and 120 minutes to administer to a multi-family household (http://www.share-project.org).

2.5.6 The average length of the interview for the World Mental Health Survey varies across participating countries, ranging from 49 minutes as a computer-assisted interview in Italy to 210 minutes as a paper-and-pencil interview in South Africa; most interviews are administered in the respondent's home, but in some countries, they are conducted in the respondent's place of employment, group quarters, cafes, libraries, or the office of the research organization (Kessler, Ustun, & World Health Organization, 2008).

3. **Obtain informed consent from every research participant.**

Rationale

Informed consent is an important component of the ethical principle *respect for persons* and is mechanized through the idea that all people deserve the right to exercise their autonomy and agency to make a choice to participate. Informed consent must provide the respondent with enough information about the project and its risks and benefits to make an informed choice

Procedural steps

3.1 Develop the necessary documentation to obtain either oral or written voluntary informed consent (Groves et al., 2009). In implementing the consent process, provide the following information and adhere to the following principles.

3.1.1 Information to provide (in oral or written form, as appropriate):

- A clear identification of the research firm.
- A brief description of the survey or examples of questions or topic areas that can be easily understood by research participants (Patton, 2002).
- A description of the role of the respondent in the study, including the expected duration of the respondent's participation (i.e., what the respondent is being asked to do).
- A clear indication that participation is voluntary and that the information provided will be held in a confidential manner**,** unless there are special circumstances in which respondents have waived confidentiality. For example, disclosure of harm to self or others may trigger a breach of confidentiality, and such an exception should be noted in the informed consent document.
- A clear indication of the use of any electronic equipment (e.g., taping, recording, photographing) and/or one-way viewing rooms.
- A clear description of any benefits and risks associated with participation.
- A clear indication that a respondent's contact information will be held for possible future contact if there is anticipation of a second wave of data collection in the future.
- Contact information for a study investigator or other research team member whom respondents can contact (provided or available on request).

- Contact information for a review board member whom respondents can contact if the study has been reviewed by an ethics review board. If consent is obtained orally, the interviewer can provide a paper document with relevant contact details to the respondent.
- See Appendix A for examples of both oral and written requests for informed consent.

3.1.2 Principles to follow when developing materials to obtain voluntary informed consent:

- Do not use coercion through force or threats.
- Do not use excessive or disproportionate influence to recruit research participants. Whether a practice is defined as coercive or not may vary by culture, population, and study.
    - For example, large monetary payments that are given to participants may be considered to be too great to refuse, particularly in resource-poor populations (Pennell et al., 2014). Always take into account the local context, particularly when surveying vulnerable populations, and discuss any planned incentives with study country collaborators. See Guideline 3 in Data Collection: General Considerations for further discussion on appropriate use of incentives.
- Consider what medium "best protect[s] the human subject" (Markham, 2005, p. 814).
- Respect the rights of individuals to refuse to be interviewed, to refuse part of the interview, and to terminate an interview in progress. Whether or not follow-up with individuals who initially refuse the survey request is appropriate may vary by culture, population, and study. The right of individuals to refuse participation in any and all part of the interviewer is an important part of the concept of *respect for persons*
- Respect the right of individuals to refuse to answer any question in the interview.
- Consent information should be conveyed in a format that is easy for respondents to understand, with language suitable for the general public. Consider the literacy level of the intended population. Written formats that may be appropriate include a document with narrative text, a list of Frequently Asked Questions (FAQs), and a brochure format. Samples of these formats can be found in Appendix A and from the American Association for Public Opinion Research (2010).
- Protect rights to privacy of study participants. This should include a careful review of government privacy laws and

regulations, which could vary on the type of data and persons that are covered and the definition of an "identifiable" case (Benson, 2007).

3.2 Obtain and document consent. Whether consent is obtained in oral or written form depends on a number of factors, including government laws and regulations, risk of harm for respondents revealing sensitive information, the mode of data collection, the type of information requested, and cultural norms. For example, in mail surveys, consent may be implied (that is, not explicitly obtained in oral or written form) if the respondent chooses to fill out the questionnaire and mail it back.

3.2.1 Obtain oral or written informed consent from all adult research participants.

3.2.2 Obtain oral or written informed consent from a parent or responsible adult before interviewing children or young people. Minors cannot consent to participate in research but can give their oral or written assent after obtaining parental permission.

3.2.3 Avoid making inaccurate or overly restrictive statements (e.g., the data will only be shared with the research team) if the data will be archived and shared with the research community (Groves et al., 2009). It is difficult to foresee all possible future uses of survey data.

3.2.4 Develop protocol for use in the field to monitor that informed consent is received for each completed survey interview. For example, interviewers can be provided with a checklist of items to complete at each interview, with the list including obtaining informed consent.

Lessons learned

3.1 Obtaining informed consent and assent may be simple and straightforward in one location but require multiple steps in another. In Western cultures, simple parental consent may suffice when studying minors. In Mali, on the other hand, a medical research team that wanted to study children under 9 years of age who had been exposed to malaria first discussed the study with a group of village elders. Next, they convened focus group discussions with the heads of extended families. Then, they held similar discussions with mothers whose children might become part of the malaria study. Finally, they obtained the consent of the individual families involved (Doumbo, 2005).

3.2 When obtaining informed consent for a study that has been reviewed by an ethics review board, contact information for a review board member whom respondents can contact may not be useful. For example, contact information for a U.S. university review board may be irrelevant for the rural population in a country or context where actually contacting the U.S. IRB is not realistic due to language, access, or other issues. In such cases, it is more relevant to provide contact information for a local, within-country entity whom the respondent could more realistically contact with any questions or concerns.

3.3 The American Association for Public Opinion Research (AAPOR) also has a number of examples of consent forms for review. See http://www.aapor.org/Standards-Ethics/Institutional-Review-Boards/Consent.aspx

4. Develop protocol for interviewers and other project members to use to protect respondent identifying details and survey data.

Rationale

Protection of respondent identity and data is a crucial element of the concept *beneficence*; that is, protecting respondents from harm, and, specifically, harm stemming from disclosure of survey responses. Protection of respondent confidentiality is achieved through appropriate interviewer training as well as data processing, storage, and dissemination procedures.

Procedural steps

4.1 Provide appropriate training to interviewing staff about ethical standards and study specific procedures to protect human subjects. Interviewers are often the first (and only) member of the research team with whom the respondent has contact. It is crucial for interviewers to understand the responsibility they have in protecting the identity and data of the respondent, as well as adequately conveying the respondent's rights with regard to the research process.

4.1.1 To the extent allowed by law or regulations, train staff to keep confidential both identifying material (e.g., respondent names, addresses, and telephone numbers) and all information given by respondents.

4.1.2 Conduct staff training on the concepts of respect for persons, beneficence, and justice and the steps all project team members must take to ensure compliance with ethical standards. Consider requiring staff to complete an online ethics course, such as the ethical training course offered by the Collaborative Institutional Training Initiative: http://www.citiprogram.org.

4.1.3 Discuss with staff the protocols that will be used to detect data falsification and of the negative contribution to the integrity of the research process.

4.1.4 Require staff to sign a pledge of confidentiality or to provide assurance in some form that they will maintain confidentiality (see Appendix B for an example of a pledge of confidentiality). It is important to note that preserving confidentiality takes on even greater significance if local interviewers are working in areas where they may be acquainted with sample members prior to the interview request.

4.1.5 Include discussion of ethical standards in any interviewer training refresher courses conducted during the field collection period.

4.2 Separate personally identifiable information (PII) from the respondent data. PII minimally includes name, address, telephone number, and identification number(s) (including an identification number assigned by a government agency such as a social security number in the United States or a national registration identity card number in the United Kingdom), but may include other information including biometric data.

4.3 Keep secure and confidential any data source which links survey responses to identifiable respondents.

4.4 Use information gained through the research activity for study-related purposes only.

4.5 Adhere to government laws and regulations on storage, retention, and dissemination of survey data.

4.6 Limit access to confidential data to project staff members who have pledged to maintain confidentiality and have been trained on appropriate use of study data.

4.7 If disclosing survey data to outside parties, require all subcontractors, consultants, and third parties to enter into an agreement to maintain respondent confidentiality. This agreement should include an explicit statement that the outside party cannot use contact information or any other information to recontact the respondent for any reason not directly related to the study (e.g., data cannot be used to approach respondents for a different study or for marketing purposes).

4.8 Report any breach of confidentiality in accordance with ethics review board policies and government regulations.
   4.8.1 Establish specific protocols for interviewers to report breaches of confidentiality and provide protocols in interviewer training. Interviewer training should include examples of anticipated breaches of confidentiality (such as reporting of abuse witnessed within the household), as well as discussion about the use of common sense, based on what interviewers know about the survey, to determine whether a breach has occurred or when a breach may be necessary.
   4.8.2 Establish specific protocols which dictate how the principal investigator and study personnel must report any breach of confidentiality to the IRB overseeing the project.
   4.8.3. If the data collection mode involves any form of technology, take the appropriate steps to secure electronic data and train interviewers accordingly (see Data Collection: Face-to-Face Surveys, Guideline 3). Loss or theft of equipment containing confidential survey data is a breach of confidentiality and should be reported to the IRB overseeing the project.

Lessons learned

4.1 Circumstances leading to a necessary and intentional breach of confidentiality can, in some cases, be anticipated and in other cases be unexpected.
   4.1.1 In the United States, interviewers may be mandated to report suspicions of child abuse or neglect that are witnessed during the research process, depending on individual state laws. If state and/or local laws apply, researchers should clearly explain interviewers' responsibilities during the training process and document any such breaches of confidentiality accordingly.
   4.1.2 In the course of the data collection period, unexpected events can arise that also necessitate a confidentiality breach. During the production period of a survey in a South Asian country, political activists burglarized the local data collection firm and stole several laptops which contained survey data files. Fortunately, the data files were securely encrypted and did not contain any identifying information. Nevertheless, the incidence was reported to the IRB overseeing the project.

5. Develop procedures and obtain voluntary informed reconsent for any additional data collection activities

Rationale

It is becoming increasingly common for survey research to include additional measurement modes beyond the survey questionnaire, including collection of biomeasures in addition to linkages to other data sources, such as government registries (e.g., U.S. Social Security Administration data) or social media data (e.g., Twitter activity). After the survey questionnaire is complete, a second consent procedure—that is, a reconsent—is administered for the secondary data collection.

Procedural steps

5.1 Consider whether a secondary data collection will be administered. For further discussion on secondary data, see the chapters on Biomeasures and Paradata and Other Auxiliary Data.

5.2 Develop oral or written reconsent documentation, which should address the same principles as the primary consent procedures outlined in Guideline 3.

5.3 Provide reconsent-specific training to interviewers.

5.4 Obtain reconsent from research participants prior to collecting secondary data or performing any linkage to a respondent's secondary data source.

5.5 Protect data obtained from secondary data collection equally to that obtained from the survey questionnaire.

Lessons learned

5.1 In a secondary data collection, the World Mental Health Survey in Saudi Arabia successfully collected saliva from respondents, from which DNA was extracted for analyses. Because there would be a cost born by the study for DNA processing, the reconsent form explicitly stated that the respondents would bear no extra cost as a result of participation in this study. Respondents were given the option to receive a general summary of the study results and to receive the results of the study that pertain specifically to the respondent. Additionally, respondents were asked a series of questions regarding consent to the potential use of any leftover saliva samples in the future.

6. Maintain sensitivity to cultural and social differences

Rationale

Designing study protocols that are sensitive to cultural traditions and norms is vital to building trust and gaining cooperation. Being respectful of cultural norms and customs also leaves individual participants with a positive impression of the research community. Beyond the individual level, it may forestall negative political and social consequences. Finally, participation in social science and health studies may promote awareness of research issues in the community.

Procedural steps

6.1 Consider a medium of data collection that is “appropriate for participants,” rather than only a form convenient for researchers (Markham, 2005, p. 812).

6.2 Do not exclude minority groups, native populations, or aboriginal peoples in the sample, unless it is appropriate to do so. Document any necessary exclusions.

6.3 Identify ethnic or religious power structures in the areas in which data collection will occur and approach study participants in accordance with the cultural traditions and norms of the ethnic or religious groups (e.g., through the head of the family or a local leader).

6.4 Involve other individuals or groups in the consent decision-making process as appropriate (e.g., older family members or local leaders).

6.5 Observe local customs in planning for and conducting the interview (e.g., giving advance notice before arriving, dressing in a culturally appropriate manner, removing one’s shoes inside the house, partaking of refreshment, sending a thank-you note).

6.6 Be flexible when implementing consent procedures (e.g., obtaining permission to accept oral consent in place of a written form, if literacy is an issue).

6.7 Present study materials in a form that can be understood by the respondent (e.g., in the respondent’s native language or orally rather than written if literacy is an issue). Avoid the use of technical language or jargon.

6.8 Observe cultural norms when assigning interviewers to sample elements (e.g., matching female interviewers with female respondents, if matching is culturally appropriate).

6.9 Attempt to conduct interviews in settings that afford as much privacy as possible while still respecting cultural norms. See Guideline 4 in Data Collection: Face-to-Face Surveys.

6.10 Identify the level or degree of sensitivity for different question topics during preliminary fieldwork, observations, and pretesting, since sensitive topics often vary among cultures and societies (Lee, 1993).

6.11 Consider cultural traditions and norms when deciding whether to offer respondent incentives and determining what type of incentives would be most appropriate. See Guideline 3 in Data Collection: General Considerations for more on incentives).

6.12 Determine whether it is appropriate to follow up with persons who initially refuse the survey request and develop follow-up protocols in accordance with cultural traditions and norms.

6.12.1 Interviewer training should specify the definition of a "hard refusal" from a respondent, how many contact and call-back attempts are permissible, and appropriate methods to address respondent concerns.

6.12.2 In the case of panel surveys, develop protocol to specify whether to contact in latter wave(s) those respondents who refused to participate in former wave(s).

6.13 Do not over generalize. Fine, Tuck, and Zeller-Berkman (2008) caution against generalizing in a way that implies "universality and sameness" (p. 159). These authors also recommend care in how language is interpreted or what the semantics of language is assumed to mean.

Lessons learned

6.1 As with other aspects of research, we cannot assume that "one size fits all" when implementing a study protocol with regard to ethics.

6.1.1 There may be different levels of requirements for privacy in different cultures. In a study involving 11-year-old boys in India, in-home interviews tended to include relatives and neighbors. At times the interviewers had to use considerable tact to discourage members of the audience from interjecting their own answers to the questions being asked (Armer & Grimshaw, 1973).

6.1.2 In some cultures, it may be necessary to gain approval from authority figures within a community (gatekeepers). In a fertility study in Guatemala, interviewers were effectively barred from a rural municipality by the single act of a local priest. The priest warned his parishioners against the "red urbanites who would prevent women from having children," as he described the researchers (Amaro & Gehlert Mata, 1968).

6.1.3 Respondents in some cultures may be reluctant to provide written documentation of consent. Researchers in Mali found that documenting the consent process with a signed paper was a challenge. At first, villagers were opposed to signing any document, because they strongly believed that their word should be sufficient. In addition, participants found the legal language difficult to understand. It took very careful explanation and patience to overcome this resistance (Doumbo, 2005). Oral consent may also be necessary because of literacy limitations, to which sensitivity should be applied.

6.1.4 Sensitivity of topic can vary widely across countries and discussions with study country collaborators to identify sensitive topics are imperative. In some religiously conservative countries, such as Egypt, it is considered inappropriate to ask general questions about religion, such as whether the respondent believes in God. In a 3MC survey in the Middle East, researchers prefaced the item about belief in God with the statement: "Please keep in mind that we ask the next set of questions because we will compare the results from COUNTRY with many other countries." This phrase diffused some of the sensitivity surrounding the topic (de Jong & Young-DeMarco, forthcoming).

6.2 Transformative researchers who empower community members to work with researchers for social change see respect as involving critical study of "cultural norms of interaction in diverse communities across cultural groups" (Mertens et al., 2010, p. 196). Quality, or the degree to which findings reflect participants' perspectives, relates to the "degree of collaboration" (Patton, 2002, p. 269) between researchers and the researched. Collaboration allows "the meanings and diffusion of knowledge" (Battiste, 2008, p. 500) to be in the hands of local participant groups. "It is vital that Indigenous peoples have direct input into developing and defining research practices and projects related to them" (Battiste, 2008, p. 503).

7. Report research findings and methods and provide appropriate access to study data.

Rationale

Professional social science organizations generally agree that their members should report findings to benefit the widest possible community. From this, it follows that data collection agencies should provide full information to allow readers and data users to assess both methodology and results. Dissemination of results and research reports also increases public confidence and alerts potential users to limits of accuracy and reliability, avoiding misinterpretation of findings. In addition, sharing documentation on study methods can assist other researchers in making informed choices about research design and implementation in future studies. While providing access to study data and methods is advantageous for the reasons outlined here, researchers must also assess the risk of a breach of confidentiality and address this concern when preparing data for dissemination.

Procedural steps

7.1 Report findings as completely, widely, and objectively as possible, while also protecting participants' confidentiality. While the full reporting of results is an important ethical obligation, it is also important to consider the negative impact that reporting unfavorable findings about a specific ethnic, religious, or other social group may have on members of that group.

7.2 Make available as much of the study's methods, results, and raw data as possible, within the bounds of protecting participants' confidentiality, in order to permit others to evaluate the study and to replicate the findings.

7.3 Evaluate the risk of a breach of confidentiality and implement appropriate techniques to protect the confidentiality of the data, including de-identification of publicly available datasets to the greatest extent possible (see Data Dissemination for a detailed discussion.

7.4 Provide a summary report of the study methodology and findings. See Appendix C for a checklist of items to include in the summary report.

7.5 Provide a copy of the findings to all researchers and organizations that were involved in the study.

7.6 Provide a copy of the de-identified dataset(s) and documentation to a trusted national data repository. See Data Dissemination for further details about dissemination and documentation.

7.7 Provide safe, sustainable storage of the datasets and documentation.

7.8 Adhere to government laws and agreements that address disclosure of survey data both within and across borders.

7.9 If an error is discovered after publication of the results, make an effort to correct the error using an erratum document that describes the error and its likely effect on study results, and provide an additional variable or other means along with appropriate documentation for analysts to identify the corrected value(s).

7.10 Make an effort to respond to specific written requests for additional items pertaining to the publicly released findings (National Council on Public Polls, 2006).

Lessons learned

7.1 There are useful examples of efforts to fully document study methods and provide survey data from 3MC surveys to a wide community of users. In part or whole, their approach and templates can serve as models for other studies.
  7.1.1 The European Social Survey website provides comprehensive information on study methodology and access to data for any registered user. Registration is free and easy to complete.
  7.1.2 The World Mental Health Survey Initiative used a standardized web-based survey instrument to collect information on study methodology from participating countries.
  7.1.3 It is important to be aware that some national standards require that raw and de-identified datasets be stored for a minimum time period (e.g., 10 years is the German National Science Foundation standard for empirical data).

8. Institute and follow appropriate quality control procedures.

Rationale

Development and implementation of quality control procedures is necessary to ensure that the procedures that have been developed to meet standards for ethical research are being carried out appropriately. If a failure to meet these standards is detected, protocols should be in place to remedy the failure. In addition, monitoring of procedures related to the ethical conduct of the study should inform efforts to improve quality and cost-effectiveness at all stages of the life cycle process. For a more in-depth discussion of the survey quality framework, see Survey Quality.

Procedural steps

8.1 Pretest consent protocol and forms to ensure comprehension.

8.2 Translate and adapt consent protocols and forms according to best practices (see Translation and Adaptation).

8.3 Consider reviewing recorded interviews and monitoring live interviews when possible to assure adherence to informed consent procedures.

8.4 Monitor implementation of confidentiality protocols and procedures, including, but not limited to performing audits to determine adherence to these protocols and procedures.

8.5 Securely store signed pledges of confidentiality and consent forms.

8.6 Recontact a sample of cases for each interviewer to verify that screening and interview procedures were appropriately followed. (see Guideline 5 of Data Collection: General Considerations for additional information.

8.7 Use analyses of paradata (e.g., identification of question-level timings that are unusually short or long and identification of unusual variable distributions for one or more interviewers compared to the overall distribution (Murphy, Baxter, Eyerman, Cunningham, & Kennet, 2004; Schäfer, Schräpler, Müller, & Wagner, 2004). For a detailed description of the use of paradata to assess survey quality, see Paradata and Other Auxiliary Data.

8.8 Conduct disclosure analysis. (see Data Dissemination for more details.

8.9 Investigate any deviation from ethical protocols and take appropriate action to address the situation.

Lessons learned

8.1 Sometimes a small group of interviewers can have a large impact on the quality of survey estimates. In a mental health survey of six European countries, the prevalence rates of mental health disorders were unusually low among German respondents. Experienced German interviewers were suspected of skipping screening questions that lead to a more extensive set of follow-up items in order to complete interviews more quickly. Even though only a small group of interviewers had prior interviewing experience, they conducted a sizeable percent of the total number of interviews and the responses that they solicited were very different. In general, positive responses screened respondents into more extensive sections on mental health disorders. Only 14.5% of screening questions administered by the interviewers with prior interviewing experience were positive, while 44.7% of screening questions administered by interviewers without prior experience were positive (Matschinger, Bernert, & Angermeyer, 2005).

9. Consider whether there are any other ethical issues resulting from design decisions, particularly when technology is used.

Rationale

With the continual expansion of technology into survey research, a number of other ethical issues have arisen, such as the accuracy of data obtained through web surveys, whether consent for the capture of certain types of paradata should be obtained, and how social media might be used in the survey process. We highlight several more common ethical concerns here, but advise researchers to review their study design to identify any such issues, especially with regards to technology. Ethical questions such as these will continue to arise and evolve, as there is further innovation in technology.

Procedural steps

9.1 If the survey design includes the use of a web survey administered without any interaction with an interviewer, consider the following:
- 9.1.1 Data collected through a web survey should be encrypted adequately for protection against a security breach.
- 9.1.2 If minor children are not included in the study design (and there is no assent process in place for their inclusion), then there should be a procedure to verify that all respondents to the survey are indeed consenting adults.
- 9.1.3 See Data Collection: Self-Administered Surveys for further considerations when using a web survey.

9.2 If the survey design includes automated or interviewer-recorded capture of paradata or other auxiliary data, consider whether reconsent should be obtained from the respondent and/or whether the respondent should be informed of the paradata capture.

9.3 Be aware of the potential uses of social media. The inclusion of social media in the survey process has increased in recent years. If the study design includes the use of social media, researchers should consider whether usage violates the principles of beneficence, justice, and respect for persons. Possible uses of social media (e.g., Facebook, Twitter, Instagram, etc.) include:

9.3.1 Use of social media profiles to screen respondents from a sample frame to identify those having particular attributes of interest.

9.3.2 Use of social media profiles to track respondents in a panel survey.

9.3.3 Use of survey respondents' publically available social media data, both profile information and Facebook updates, Twitter tweets, etc., to augment data from survey questionnaires.

9.3.4 Use of social media to identify questions, domains, and concepts in populations of interest during study design development.

Lessons learned

9.1 The American Association for Public Opinion Research (AAPOR) has published the *Social Media Task Force Report*, outlining the most recent considerations on emerging technologies in public opinion research (Murphy et al., 2014).

10. Document materials and procedures related to the ethical conduct of the study and ethics committee reviews.

Rationale

In research that involves human subjects, it is critical to maintain documentation of materials that were used to inform potential participants about study participation and subsequently record consent, in case there is ever a question of ethics violations or a request for additional information from an ethics review board. In addition, documentation of all survey procedures including those related to the ethical conduct of the study is a key element of high quality scientific research.

Procedural steps

10.1 Consider what ethical concerns may pertain to the issue of selecting research findings to publish (Hesse-Biber, 2010).

10.2 Maintain a copy of the following documents:
- 10.2.1 Scripts, letters, fact sheets, and any other materials provided to respondents to give them information they need to make an informed decision about participation.
- 10.2.2 Consent form templates and protocols.
- 10.2.3 Translated or adapted consent form templates and protocols.
- 10.2.4 Individual consent information for each respondent, stored in a safe environment separate from survey data.
- 10.2.5 Confidentiality procedures and protocols.
- 10.2.6 Pledge(s) of confidentiality completed by staff.
- 10.2.7 Records of completion of any specialized staff training on ethics.
- 10.2.8 The original submission to the Ethics review board, requests for modification to study protocol, and routine renewal material (see Appendix D for a checklist of materials to include in an ethics review board submission).
- 10.2.9 Ethics review board correspondence (e.g., letters of approval).
- 10.2.10 Documentation of any other ethical review required by study sponsors, individual study countries, etc.
- 10.2.11 Any correspondence between study staff or ethics review board members/staff and respondents regarding an ethical issue or concern.
- 10.2.12 Reports of quality control activities (e.g., documentation of verification activities).

10.3 Provide a copy of the following documents to any central coordinating organization:
- 10.3.1 Translated or adapted consent form templates and protocols.
- 10.3.2 Ethics review board original submission and requests for modification to study protocol.
- 10.3.3 Ethics review board correspondence (e.g., letters of approval).
- 10.3.4 Reports of quality control activities (e.g., documentation of verification activities).

Appendix A

Study brochure

The following is a sample study brochure that can be mailed or handed to respondents to provide general information about the study purpose and protocol and to address frequently asked questions.

What is this project about?

The Chicago Healthy Neighborhoods Study (CHNS) is a research study funded by the US National Office for Health to determine the impact of the quality of life in Chicago neighborhoods on the health of adults living there.

The information gathered from this study will help us better understand why there are social, economic, and racial/ethnic differences in the health of Chicagoans and how these differences affect Chicagoans' lives. With data from this study, effective approaches can be developed to improve the health and lives of all Chicagoans.

Who is asked to participate?

The CHNS is one of the largest surveys, done in a major American city, studying the relationship of the quality of people's lives and the neighborhood in which they live to their health. About 4,500 adults will participate in this important study.

Households are randomly selected using a scientific sampling procedure. Once a household is selected, an interviewer visits the house and makes a listing of all residents. One adult is randomly selected from all eligible residents. Only the selected individual may participate. Each person who is asked to participate has been carefully selected to represent fellow Chicagoans like them.

Is participation voluntary?

Yes. Participation in this project is voluntary. Project participants may choose not to answer any or all of the questions. However, each participant has been carefully selected and thus cooperation from each potential participant is critical to the success of this research.

How will the interviews be conducted?

Interviews will be conducted in the participant's home or at another location by a professional University of West Chicago Survey Research Center interviewer. The interviewer will ask questions and record answers using a laptop computer. Participants will be provided with $20 as a token of appreciation for their participation in this project.

What kinds of questions will I be asked?

The interview includes a wide range of questions about work and family life, health, and social and physical characteristics of neighborhoods in which study participants live. There are no right or wrong answers. Most participants find the interview to be an enjoyable experience.

How will the data be used?

The data collected will help researchers and government policy makers better understand social, economic, and racial/ethnic differences in the health of adults living in Chicago, so that effective approaches can be developed to improve the health and lives of all Chicagoans. Data from this study will only be reported in summary form. Participants' individual identities and answers to questions will remain strictly confidential.

Who is funding the project?

Funding for CHNS comes from the US National Office for Health (NOH).

The University of West Chicago's Survey Research Center will conduct the interviews for this study. A University of West Chicago interviewer will greet you at your home. For security reasons, you may want to ask the interviewer to reveal his/her identification badge. UWC employees will gladly comply with your request.

We thank you for your interest in this project!

Chicago Healthy Neighborhoods Study

Project Leader

Christopher Antoun, Ph.D., Survey Research Center (SRC) & Department of Urban Health, University of West Chicago

Senior Investigators

Benjamin Duffey, Ph.D., Department of Urban Health, University of West Chicago

Hyun Jung Lee, Ph.D., Department of Urban Health & SRC, University of West Chicago

Emily Blasczyk, Ph.D., Department of Psychology & SRC, University of West Chicago

Mason Flounder, Ph.D., Department of Sociology, Northwestern University

Yuchieh Lin, M.D., Mental Health Research Institute & Department of Psychiatry, University of West Chicago

William Jones, M.D., Department of Psychiatry, University of West Chicago

Consultants

CHICAGO HEALTHY NEIGHBORHOODS STUDY

If you have any questions, please contact the project team toll-free at:

1-800-733-7373

**University of West Chicago
Survey Research Center**

Appendix B

Pledge of confidentiality to safeguard respondent privacy

This pledge to maintain respondent privacy is used by the Institute for Social Research at the University of Michigan. The form is signed by all staff members, and fulfillment of the pledge is a requirement of employment.

I have read the Institute for Social Research Policy on Safeguarding Respondent Privacy, and pledge that I will strictly comply with that Policy. Specifically:

- I will not reveal the name, address, telephone number, or other identifying information of any respondent (or family member of a respondent or other informant) to any person other than an employee directly connected to the study in which the respondent is participating.
- I will not reveal the contents or substance of the responses of any identifiable respondent or informant to any person other than an employee directly connected to the study in which the respondent is participating, except as authorized by the project director or authorized designate.
- I will not contact any respondent (or family member, employer, other person connected to a respondent or informant) except as authorized by the project director or authorized designate.
- I will not release a dataset (including for unrestricted public use or for other unrestricted uses) except in accordance with authorization, policies and procedures established by ISR and the Center with which I am affiliated.
- I will take all necessary precautions to avoid unintended disclosure of confidential information, including securing of paper and electronic records, computers, user IDs and passwords.

I agree that compliance with this Pledge and the underlying Policy is: 1) a condition of my employment (if I am an employee of ISR), and 2) a condition of continuing collaboration and association with ISR (if I am an affiliate of ISR). I understand that violation of this Policy and Pledge may result in disciplinary action, up to and including termination of employment or severance of any relationship with ISR and the applicable research project.

If I supervise affiliates who have access to ISR respondent data (other than unrestricted public release datasets), I will ensure that those affiliates adhere to the same standards of protection of ISR respondent privacy, anonymity, and confidentiality, as required by this Pledge and the associated Policy.

Signature: ______________________________ Date: ____________________
Typed or printed name: ____

Appendix C

Checklist of items to include in summary report of study methodology and findings

3.1	The purpose of the study
3.2	Who sponsored the survey and who conducted it
3.3	A copy of ethics review board approval (if appropriate)
3.4	A copy of the informed consent form or script
3.5	A definition of the population under study and a description of the sampling frame
3.6	A description of the sampling and survey designs
3.7	Sample sizes and, where appropriate, eligibility criteria, screening procedures, and response rates. A summary of the disposition of sample elements should be included, in order for the user to calculate a response rate should one not be included in the report or a different one desired.
3.8	Method, location, and dates of data collection
3.9	A copy of questionnaire, interviewer instructions, and any visual aids used in the interview
3.10	A detailed description of results that are based on anything less than the total sample, including the size of the sample and inclusion/exclusion criteria
3.11	A full description of the weighting (if appropriate) and estimation procedures used for all results that are reported
3.12	The major findings
3.13	A description of the precision of the findings, including estimates of sampling error

Primary Source: American Association for Public Opinion Research. (2005). *Standards for minimal disclosure.* Retrieved April 5, 2010, from http://www.aapor.org/Disclosure_Standards.htm

Appendix D

Checklist of materials to be provided to an ethics review board

General Study Information, including: • Financial sponsorship • Key personnel • Performance sites • Study dates • Study abstract/summary • Research design (including specific aims, background/prior research, methodology, analysis plan, etc.) • Benefits to subjects from participation • Risks to subjects • Recruitment methods and description of subject population • Informed consent procedures • Data confidentiality provisions • Conflicts of interest
Discussion of Special Considerations, for example: • Procedures used to obtain consent to interview minors or other populations that require special consent (e.g., if interviewing minors, describe procedures for obtaining parental consent and include child assent and parental consent forms/oral protocols). • Compensation and costs involved in participation for study subjects • Procedures for handling biological samples, such as blood or saliva • Proposal to conduct genetic typing/analysis from biological samples • Considerations in conducting epidemiological or public health research • Use of deception • Use of internet/email for research • Consent procedures for audio or video recording of interviews • International research considerations • Protocols for viewing of images or listening to recorded material

• Secondary data analysis
Forms, including: • Copy of the grant/contract application • Consent protocols/scripts/forms • Copy of the questionnaire *Other forms (as appropriate):* • Cognitive interview protocol • Focus group moderator guide • Recruitment flyers or emails • Study brochure/fact sheet • Letter(s) to be sent to respondents • Data use agreement (for use of secondary data from third party sources) • Documentation of review from other ethics review boards • Documentation of training in research ethics for study staff

Checklist developed based on material available from the University of Michigan Health Sciences and Behavioral Sciences Institutional Review Boards (http://www.irb.umich.edu/).

Appendix E: Sample of Oral Consent Form Used for a Survey in Tunisia

Hello. I am from [SURVEY DATA COLLECTION COMPANY NAME] and am working in collaboration with [UNIVERSITY IN TUNISIA]. We are carrying out academic research in Tunisia on what people value in life. This research will interview a nationally representative sample of the population in Tunisia. Your home address has been selected randomly as part of a representative sample of the people living in Tunisia.

We are seeking your permission to ask your opinion on topics such as development, beliefs about families, politics, media use, corruption, and various other attributes of individual and family life. For example, we might ask you how optimistic you feel these days, or about how important you think democracy is when discussing attributes of a good government. Please be assured that there is no right or wrong answer to any of these questions. Your help is extremely important because it will contribute to a better understanding of what people around the world believe and want out of life.

Your answers will be kept completely confidential. Your identifying information will be kept in a separate, secure location from your survey responses and will be linked only by an arbitrary identification number. We believe there is no risk to you for taking part in this study. Any answers you give will be combined with the responses of all other participants. This means that no one will be able to trace the identities of any of our individual participants. The results of this research will be used for academic purposes only and will be disseminated in scholarly journals and presentations. This research may be indirectly beneficial to you because it contributes to the development of the social sciences and to public policy. However, you will experience no direct benefits from participating in this study.

This interview will take about an hour or so and I want to assure you that it is completely voluntary and confidential. If we should come to any question that you do not want to answer, please let me know and we will go on to the next question. There is no penalty for not participating or for refusing to answer any question. You may stop the interview at any time.

We may contact you in the future about an opportunity to participate in a follow-up discussion about some of the same topics raised in the questions I'll be asking you today. Again, participation in any subsequent interview would be completely voluntary and confidential.

This research protocol and informed consent document has been reviewed and approved by Eastern Michigan University Human Subjects Review Committee for use from _________ to _________ (date). If you have questions about the approval process, please contact (PRINCIPAL INVESTIGATOR NAME) (PHONE, EMAIL) or the Eastern Michigan University Human Subjects Review Committee (PHONE).

By verbally stating "I agree," you are indicating that you are at least 18 years of age, you have had this consent form read to you, your questions have been answered, and you voluntarily agree to participate. If you agree, please state "I agree."

If you want to know more about the study, you can call (NAME) or (NAME) [SURVEY DATA COLLECTION COMPANY NAME] at (PHONE).

Sample of Written Consent Form Used for a Survey in Tunisia

Hello. I am from [SURVEY DATA COLLECTION COMPANY NAME] and am working in collaboration with [UNIVERSITY IN TUNISIA]. We are carrying out academic research in Tunisia on what people value in life. This research will interview a nationally representative sample of the population in Tunisia. Your home address has been selected randomly as part of a representative sample of the people living in Tunisia.

We are seeking your permission to ask your opinion on topics such as development, beliefs about families, politics, media use, corruption, and various other attributes of individual and family life. For example, we might ask you how optimistic you feel these days, or about how important you think democracy is when discussing attributes of a good government. Please be assured that there is no right or wrong answer to any of these questions. Your help is extremely important because it will contribute to a better understanding of what people around the world believe and want out of life.

Your answers will be kept completely confidential. Your identifying information will be kept in a separate, secure location from your survey responses and will be linked only by an arbitrary identification number. We believe there is no risk to you for taking part in this study. Any answers you give will be combined with the responses of all other participants. This means that no one will be able to trace the identities of any of our individual participants. The results of this research will be used for academic purposes only and will be disseminated in scholarly journals and presentations. This research may be indirectly beneficial to you because it contributes to the development of the social sciences and to public policy. However, you will experience no direct benefits from participating in this study.

This interview will take about an hour or so and I want to assure you that it is completely voluntary and confidential. If we should come to any question that you do not want to answer, please let me know and we will go on to the next question. There is no penalty for not participating or for refusing to answer any question. You may stop the interview at any time.

We may contact you in the future about an opportunity to participate in a follow-up discussion about some of the same topics raised in the questions I'll be asking you today. Again, participation in any subsequent interview would be completely voluntary and confidential.

This research protocol and informed consent document has been reviewed and approved by Eastern Michigan University Human Subjects Review Committee for use from _________ to _________ (date). If you have questions about the approval process, please contact (PRINCIPAL INVESTIGATOR NAME) (PHONE, EMAIL) or the Eastern Michigan University Human Subjects Review Committee (PHONE).

If you want to know more about the study, you can call (NAME) or (NAME) of [SURVEY DATA COLLECTION COMPANY NAME] at (PHONE).

Consent

The nature and purpose of this research have been sufficiently explained and I give my consent to participate in the interview.

Name (PLEASE PRINT CLEARLY): ____

Signature: ______________________________________**Date:** _____________________

References

Afrobarometer Survey. Retrieved June 1, 2016, from http://www.afrobarometer.org/sites/default/files/survey_manuals/ab_r6_survey_manual_en.pdf

Amaro, N., & Gehlert Mata, C. (1968). *Encuesta sobre el condicionamiento socio-cultural de la fecundidad en áreas marginales urbanas-metropolitanas, ladino-rurales e indígenas-tradicionales: Primera parte, marco teórico y metodológico* [*Survey on socio-cultural conditioning of fertility in urban marginal areas-metropolitan, rural and indigenous-ladino-traditional: Part One, theoretical and methodological framework*] (2nd ed.). Guatemala: Instituto Centroamericano de Población y Familia.

American Association for Public Opinion Research (AAPOR). (2010). *Examples of consent forms*. Retrieved June 1, 2016, from http://www.aapor.org/Standards-Ethics/Institutional-Review-Boards/Samples/Consent-Form-Examples.aspx

American Association for Public Opinion Research (AAPOR). (2015). AAPOR Code of Ethics. Retrieved May 12, 2016, from http://www.aapor.org/Standards-Ethics/AAPOR-Code-of-Ethics.aspx.

Armer, M., & Grimshaw, A. D. (1973). *Comparative social research*: *Methodological problems and strategies*. New York, NY: John Wiley & Sons.

Andre, R., & Lusk, J. L. (2011). *What do Haitians need after the earthquake?* 2011 Annual Meeting, February 5-8, 2011, Corpus Christi, Texas 98631, Southern Agricultural Economics Association.

Asian Barometer. Retrieved July 1, 2010, from http://www.asianbarometer.org/intro/program-overview

Battiste, M. (2008). Research ethics for protecting indigenous knowledge and heritage: institutional and research responsibilities. In N. K. Denzin, Y. Lincoln, & L. T. Smith (Eds.), *Handbook of critical and indigenous methodologies* (pp. 497-510). Los Angeles, CA: Sage.

Benson, G. (2007). *The impact of privacy laws on survey data quality: A European case study.* Paper presented at the Council for Marketing and Opinion Research, Atlanta, GA.

Berry, J. W., Poortinga, Y. H., Segall, M. H., & Dasen, P. R. (2002). *Cross-cultural psychology: Research and applications* (2nd ed.). Cambridge University Press.

Bradburn, N. M. (1978). Respondent burden. *Proceedings of the section on survey research methods* (pp. 35-40). Alexandria, VA: American Statistical Association.

Bulmer, M. (2008). The ethics of social research. In N. Gilbert (Ed.), *Researching social life* (3rd ed., pp. 145-161). London, UK: Sage.

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

De Jong, J. & Young-DeMarco, L. (Forthcoming). Best practices: Lessons from a Middle East survey research program. In M. Gelfand & M. Moaddel (Eds.), *The Arab spring and changes in values and political actions in the Middle East: Explorations of visions and perspective*. Oxford, U. K.: Oxford University Press.

Dillman, D. A., Smyth, J. D., & Christian, L. M. (2009). *Internet, mail, and mixed-mode surveys: The tailored design method* (3rd ed.). New York, NY: John Wiley & Sons.

Doumbo, O. K. (2005). It takes a village: Medical research and ethics in Mali. *Science, 307*(5710), 679–681. Retrieved May 13, 2008, from http://www.sciencemag.org/cgi/content/full/307/5710/679

Fine, M., Tuck, E., & Zeller-Berkman, S. (2008). Do you believe in Geneva? Methods and ethics at the global-local nexus. In N. K. Denzin, Y. Lincoln, & L. T. Smith (Eds.), *Handbook of critical and indigenous methodologies* (pp. 157-180). Los Angeles, CA: Sage.

Fowler, F. J. Jr. (1995). *Improving survey questions: Design and evaluation* (*Vol. 38, Applied social research methods series*). Thousand Oaks, CA: Sage Publications.

Government of Canada. (2014). Tri-council policy statement: Ethical conduct for research involving humans. Retrieved from http://www.pre.ethics.gc.ca/eng/policy-politique/initiatives/tcps2-eptc2/Default/

Groves, R. M., Fowler, F. J. Jr., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2009). Principles and practices related to ethical research. In ,R. M. Groves , F. J. Fowler Jr., M. P. Couper, J. M. Lepkowski, E. Singer, & R. Tourangeau (Eds.), *Survey methodology* (2nd ed.) (pp. 371-400). Hoboken, NJ: John Wiley & Sons.

Hesse-Biber, S. N., & Leavy, P. (2010). *The practice of qualitative research*. Thousand Oaks, CA; Sage Publications.

ICC/ESOMAR. (2008). ICC/ESOMAR International code on market and social research. Retrieved May 2016, from https://www.esomar.org/uploads/public/knowledge-and-standards/codes-and-guidelines/ICCESOMAR_Code_English_.pdf .

International Organization for Standardization (ISO). (2016). ISO Standards Catalogue. Retrieved May 2016, from http://www.iso.org/iso/home/store/catalogue_tc/catalogue_tc_browse.htm?commid=324087&published=on.

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

Lee, R. M. (1993). *Doing research on sensitive topics*. London, UK: Sage Publications.

Living Standard Measurement Study Survey. (1996). Working paper 126: A manual for planning and implementing the Living Standard Measurement Study Survey. Retrieved July 10, 2010, from http://www-wds.worldbank.org/external/default/WDSContentServer/WDSP/IB/2000/02/24/000009265_3961219093409/Rendered/PDF/multi_page.pdf

Markham, A. N. (2005). The methods, politics, and ethics of representation in online ethnography. In N. K. Denzin & Y. S. Lincoln (Eds.), *The Sage handbook of qualitative research* (Vol. 3rd, pp. 793-820). Thousand Oaks, CA: Sage.

Matschinger, H., Bernert, S., & Angermeyer, M. C. (2005). An analysis of interviewer effects on screening questions in a computer-assisted personal mental health interview. *Journal of Official Statistics, 21*(4), 657-674.

Mertens, D. M., Bledsoe, K. L., Sullivan, M., & Wilson, A. (2010). Utilization of mixed methods for transformative purposes. In A. Tashakkori & C. Teddlie (Eds.), *SAGE handbook of mixed methods in social & behavioral research* (2nd ed., pp. 193-214). Thousand Oaks, CA: Sage.

Milgram, S. (1965). Some conditions of obedience and disobedience to authority. *Human Relations, 18*(1), 57-76.

Mneimneh, Z.,Karam, E.G., Karam, A.N., Tabet, C.C., Fayyad, J., Melhem, N., et al. (2008). Survey design and operation in areas of war conflict: The Lebanon wars surveys experience. Paper presented at International Conference on Survey Methods in Multinational, Multiregional, and Multicultural Contexts, Berlin, Germany.

Mneimneh, Z., Axinn, W.G., Ghimire, D., Cibelli, K.L., Alkaisy, M.S. (2014). Conducting surveys in areas of armed conflict. In R.

Tourangeau, B. Edwards., T, P, Johnson, K. M. Wolter, & N. Bates (Eds.), *Hard-to-survey populations* (pp. 134-156). Cambridge: Cambridge University Press.

Murphy, J., Baxter, R. K., Eyerman, J., Cunningham, D., & Kennet, J. (2004). *A system for detecting interviewer falsification*. Paper presented at the annual meeting of the American Association for Public Opinion Research, Phoenix, Arizona.

Murphy, J., Link, M. W., Childs, J. H., Tesfaye, C. L., Dean, E., Stern, M., ... & Harwood, P. (2014). Social Media in Public Opinion Research: Report of the AAPOR Task Force on Emerging Technologies in Public Opinion Research. Retrieved May 12, 2016, from http://www.aapor.org/AAPOR_Main/media/MainSiteFiles/AAPOR_Social_Media_Report_FNL.pdf

National Council on Public Polls. (2006). *Principles of disclosure.* Retrieved January 29, 2010, from http://www.ncpp.org/files/NCPP%20POD%202006.pdf

Patton, M. Q. (2002). *Qualitative research and evaluation methods* (Vol. 3rd). Thousand Oaks, CA: Sage.

Pennell, B. E., Deshmukh, Y., Kelley, J., Maher, P., Wagner, J., & Tomlin, D. (2014). Disaster research: Surveying displaced populations. In R. Tourangeau, B. Edwards, T. P. Johnson, K. M. Wolter, & N. Bates (Eds.), *Hard-to-survey populations* (pp. 111-133). Cambridge: Cambridge University Press.

Punch, M. (1994). Politics and ethics in qualitative research. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (pp. 83-95). Thousand Oaks, CA: Sage Publications.

Richardson, R. C., Plummer, C. A., Barthelemy, J. J., & Cain, D. S. (2009). Research after natural disasters: Recommendations and lessons learned. *Journal of Community Engagement and Scholarship*, *2*(1), 3-11.

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.* Paper presented at the European Conference on Quality and Methodology in Official Statistics, Mainz, Germany. Retrieved April 13, 2010, from http://www.diw.de/documents/publikationen/73/diw_01.c.42515.de/dp441.pdf

Singer, E. (2008). Ethical issues in surveys. In E. D. de Leeuw, J. J. Hox, & D. A. Dillman (Eds.), *International handbook of survey methodology* (pp. 78-96). New York/London: Lawrence Erlbaum Associates.

Singleton Jr., R. A., & Straits, B. C. (2005). *Approaches to social research* (4th edition). New York, NY: Oxford University Press.

Smith, T. (2007). 42 codes of ethics and standards in survey research. In W. Donsbach & M. Traugott (Eds.), *The SAGE handbook of public opinion research* (pp. 459-468). London: SAGE Publications Ltd.

Sumathipala, A. & Siribaddana, S. (2005). Research and clinical ethics after the tsunami: Sri Lanka. *Lancet*, *366,* 1418–1420.

Tourangeau, R., Edwards, B., Johnson, T. P., Wolter, K. M., & Bates, N. (Eds.) (2014). *Hard-to-survey populations*. Cambridge: Cambridge University Press.

United States. (1978). *The Belmont report: Ethical principles and guidelines for the protection of human subjects of research*. Bethesda, MD: The Commission.

WAPOR. (2011). WAPOR Code of Ethics. Retrieved May 12, 2016, from http://wapor.org/wapor-code-of-ethics/.

World Health Organization (WHO). (2009). *Research leadership development training grants guidance for developing a research proposal*. Geneva: WHO. Retrieved October 7, 2009, from http://www.who.int/rpc/research_ethics/guide_rp/en/

World Medical Association (WMA). (1964). *Declaration of Helsinki: Ethical principles for medical research involving human subjects* (adopted amended 29th WMA General Assembly, Tokyo, Japan, October 18th WMA General Assembly, Helsinki, Finland, June 1964; 1975; 35th WMA General Assembly, Venice, Italy, October 1983; 41st WMA General Assembly, Hong Kong, September 1989; 48th WMA General Assembly, Somerset West, Republic of South Africa, October 1996; and 52nd WMA General Assembly, Edinburgh, Scotland, October 2000. Notes of Clarification added by the WMA General Assembly, Washington 2002 and Toyko 2004). Ferney-Voltaire, France: WMA. Retrieved April 13, 2010, from http://www.wma.net/en/30publications/10policies/b3/

Further Reading

American Anthropological Association. (1998). Code of Ethics of the American Anthropological Association. approved June 1998. Arlington, Va.: *American Anthropological Association.*

American Association for Public Opinion Research (AAPOR). (2005). *AAPOR code of professional ethics & practices*. Lenexa, KS: AAPOR. Retrieved May 16, 2016, from http://www.aapor.org/Standards-Ethics/AAPOR-Code-of-Ethics/AAPOR_Code_Accepted_Version_11302015.aspx

American Statistical Association (ASA). (1999). *Ethical guidelines for statistical practice*. Alexandria, VA: ASA. Retrieved May 5, 2010, from http://www.amstat.org/about/ethicalguidelines.cfm

Bewaji, J. A. I. (2004). Ethics and morality in Yoruba culture. In K. Wiredu (Ed.), *A companion to African philosophy* (pp. 396-403). Malden, MA: Blackwell Publishing.

Bulmer, M. (1998). The problem of exporting social survey research. *American Behavioral Scientist, 42*, 153-167.

Carapico, S. (2006). No easy answers: The ethics of field research in the Arab world. *Political Science and Politics*, *36*(3), 429-431.

Chilisa, B. (2009). Indigenous African-centered ethics: Contesting and complementing dominant models. In D. M. Mertens & P. E. Ginsberg (Eds.), *The handbook of social research ethics* (pp. 407-425). Thousand Oaks, CA: Sage.

Christakis, N. A. (1992). Ethics are local: Engaging cross-cultural variation in the ethics for clinical research. *Social Science & Medicine*, *35*(9), 1079-1091.

Christians, C. G. (2005). Ethics and politics in qualitative research. In N. K. Denzin & Y. S. Lincoln (Eds.), *The Sage handbook of qualitative research* (Vol. 3rd, pp. 139-164). Thousand Oaks, CA: Sage.

Christians, C. G. (2008a). The ethics of universal being. In S. J. A. Ward & H. Wasserman (Eds.), *Media ethics beyond borders: A global perspective* (pp. 6-23). Johannesburg, South Africa: Heinemann.

Council for International Organizations of Medical Sciences (CIOMS). (2002). *International ethical guidelines for biomedical research involving human subjects*. Geneva: CIOMS.

Council of American Survey Research Organizations (CASRO). (1997-2008). *CASRO code of standards and ethics for survey research*. Port Jefferson, NY: CASRO. Retrieved November 13, 2009, from https://c.ymcdn.com/sites/www.casro.org/resource/resmgr/casro_code_of_standards.pdf

Couper, M. P., & de Leeuw, E. D. (2003). Nonresponse in cross-cultural and cross-national surveys. In J. A. Harkness, F. J. R. van de Vijver, &

P. P. Mohler (Eds.), *Cross-cultural survey methods* (pp. 155-177). Hoboken, NJ: John Wiley & Sons.

Dawson, L., & Kass, N. E. (2005). View of US researchers about informed consent in international collaborative research. *Social Science & Medicine, 61*, 1211-1222.

Dunbar Jr., C., Rodriguez, D., & Parker, L. (2001). Race, subjectivity and the interview process. In J. F. Gubrium & J. A. Holstein (Eds.), *Handbook of interview research* (pp. 279-298). Thousand Oaks, CA: Sage.

European Union (EU). (1995). *Directive 95/46/EC of the European parliament and the council on the protection of individuals with regard to the processing of personal data and on the free movement of such data. official journal of the European communities*, 281, 31-50. Retrieved April 14, 2010, from http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31995L0046:EN:HTML

Fluehr-Lobban, C. (2003). *Ethics and the profession of anthropology: Dialogue for ethically conscious practice* (2nd ed.). Walnut Creek, CA: AltaMira Press.

Freed-Taylor, M. (1994). Ethical considerations in European cross-national research. *International Social Science Journal, 142*, 523-532. Retrieved May 13, 2008, from http://www.unesco.org/most/ethissj.htm

Gatson, S. N. (2011). The methods, politics, and ethics of representation in online ethnography. In N. K. Denzin & Y. S. Lincoln (Eds.), *The SAGE handbook on qualitative research* (4th ed., pp. 513-528). Sage.

Goodhand, J. (2000). Research in conflict zones: Ethics and accountability. *Forced Migration Review, 8*(4), 12-16.

Graffigna, G., Bosio, A. C., & Olson, K. (2010). How do ethics assessments frame results of comparative qualitative research? A theory of technique approach. *International Journal of Social Research Methodology, 13*(4), 341-355.

Hershfield, A. F., Rohling, N. G., Kerr, G. B., & Hursh-Cesar, G. (1983). Fieldwork in rural areas. In M. Bulmer & D. P. Warwick (Eds.), *Social research in developing countries* (pp. 241-252). New York, NY: John Wiley & Sons.

ICC/ESOMAR. (2007). *ICC/ESOMAR international code on market and social research practice*. Amsterdam, The Netherlands: ESOMAR. Retrieved May 13, 2008, from https://www.esomar.org/uploads/public/knowledge-and-standards/codes-and-guidelines/ICCESOMAR_Code_English_.pdf

International Statistical Institute (ISI). (2009). *International Statistical Institute declaration on professional ethics*. The Hague, The Netherlands: ISI. Retrieved March 1, 2010, from http://isi.cbs.nl/ethics0index.htm

Joe, K., Raben, F., & Phillips, A. (2016) Ethical issues of survey and market research. In C. Wolf, D. Joye, T.W. Smith, & Y. Fu (Eds.), *The SAGE Handbook of Survey Methodology*. Thousand Oaks, CA: Sage.

Johnson, D., & Bullock, M. (2009). The ethics of data archiving: Issues from four perspectives. In D. M. Mertens & P. E. Ginsberg (Eds.), *The handbook of social research ethics* (pp. 214-228). Thousand Oaks, CA: Sage.

Kass, N. E., Hyder, A. A., Ajuwon, A., Appiah-Poku, J., Barsdorf, N., Elsayed, D. E., Mokhachane, M., et al. (2007). The structure and function of research ethics committees in Africa: A case study. *PLoS Medicine, 4*(1). doi:10.1371/journal.pmed.0040003

Kearl, B., Gafsi, S., Mbilinyi, M., Nabila, J., & Norman, D. (Eds.). (1976). *Field data collection in the social sciences: Experiences in Africa and the Middle East*. New York, NY: Agricultural Development Council.

Lieban, R. W. (1990). Medical anthropology and the comparative study of medical ethics. In G. Weisz (Ed.), *Social science perspectives on medical ethics* (pp. 221-240). Dordrecht, The Netherlands: Kluwer Academic.

Marshall, P. A. (2001). The relevance of culture for informed consent in U.S.-funded international health research. In *Ethical and policy issues in international research: Clinical trials in developing countries, Vol. II: Commissioned papers and staff analysis* (Chapter C). Bethesda, MD: National Bioethics Advisory Commission. Retrieved June 9, 2008, from http://bioethics.georgetown.edu/nbac/clinical/Vol2.pdf

Mertens, D. M. (2003). Mixed methods and the politics of human research: The transformative-emancipatory perspective. In A. Tashakkori & C. Teddlie (Eds.), *Handbook of mixed methods in social & behavioral research* (pp. 135-166). Thousand Oaks, CA: Sage.

Mertens, D. M., & Ginsberg, P. E. (2009a). *The handbook of social research ethics*. Thousand Oaks, CA: Sage.

Mertens, D. M., & Ginsberg, P. E. (2009b). Ethics within diverse cultural groups. In D. M. Mertens & P. E. Ginsberg (Eds.), *The handbook of social research ethics* (pp. 387-390). Thousand Oaks, CA: Sage.

Mneimneh, Z., Pennell, B.E., Kelley, J., & Cibelli Hibben, K. (2016). Surveys in societies in turmoil. In C. Wolf, D. Joye, T.W. Smith, & Y. Fu (Eds.), *The SAGE* handbook *of survey methodology*. Thousand Oaks, CA: Sage.

Moewaka Barnes, H., McCreanor, T., Edwards, S., & Borell, B. (2009). Epistemological domination: Social science research ethics in Aotearoa. In D. M. Mertens & P. E. Ginsberg (Eds.), *The handbook of social research ethics* (pp. 442-457). Thousand Oaks, CA: Sage.

Mohler, P. Ph., & Uher, R. (2003). Documenting comparative surveys for secondary analysis. In J. A. Harkness, F. J. R. Van de Vijver, & P. P. Mohler (Eds.), *Cross-cultural survey methods* (pp. 311-327). Hoboken, NJ: John Wiley & Sons.

Mohler, P. Ph., Pennell, B-E., & Hubbard, F. (2008). Survey documentation: Toward professional knowledge management in sample surveys. In E. D. de Leeuw, J. J. Hox, & D. A. Dillman (Eds.), *International handbook of survey methodology.* New York/London: Lawrence Erlbaum Associates.

National Bioethics Advisory Committee (NBAC). (2001). *Ethical and policy issues in international research: Clinical trials in developing countries, Vol. 1: Report and recommendations of the National Bioethics Advisory Commission submitted to the President of the United States, April 18, 2001*. Bethesda, Maryland: NBAC. Retrieved May 13, 2008, from http://bioethics.georgetown.edu/nbac/pubs.html

National Commission for Protection of Human Subjects of Biomedical and Behavioral Research. (1978). *The Belmont Report: Ethical principles and guidelines for the protection of human subjects of research* (DHEW Publication No. (OS) 78-0012). Washington, DC: Government Printing Office.

Office for Human Research Protections. (1993). *Institutional review board guidebook.* Retrieved May 13, 2016, from http://wayback.archive-it.org/org-745/20150930181805/http://www.hhs.gov/ohrp/archive/irb/irb_guidebook.htm

Ong, A., & Collier, S. J. (2005). *Global assemblages: Technology, politics, and ethics as anthropological problems.* Oxford, UK: Blackwell.

Organisation for Economic Co-operation and Development. (1981). *Guidelines on the protection of privacy and transborder flows of personal data*. Retrieved June 11, 2016, from http://www.oecd.org/sti/ieconomy/oecdguidelinesontheprotectionofprivacyandtransborderflowsofpersonaldata.htm

Sugarman, J., Popkin, B., Fortney, J., & Rivera, R. (2001). International perspectives on protecting human research subjects. In *Ethical and policy issues in international research: Clinical trials in developing countries, Vol. II: Commissioned Papers and Staff Analysis* (Chapter E). Bethesda, MD: National Bioethics Advisory Commission.

Thomas, V. G. (2008). Critical Race Theory: Ethics and dimensions of diversity in research. In D. M. Mertens & P. E. Ginsberg (Eds.), *The handbook of social research ethics* (pp. 54-68). Thousand Oaks, CA: Sage.

Twumasi, P. A. (2001). *Social research in rural communities* (2nd ed.). Accra, Ghana: Ghana Universities Press

United States Department of Commerce. (2000). Safe Harbor overview. Washington, DC: United States Department of Commerce. Retrieved May 5, 2010, from http://www.export.gov/safeharbor/index.asp

United States Department of Health and Human Services Office for Human Research Protections. (2005). US code of federal regulations, Title 45, Part 46. Washington, DC: United States Department of Health and Human Services Office for Human Research Protections. Retrieved May 13, 2008, from http://www.hhs.gov/ohrp/sites/default/files/ohrp/policy/ohrpregulations.pdf

United States Department of Health and Human Services Office for Human Research Protections. (2010). *Office for Human Research Protections (OHRP): International issues*. Washington, DC: United States Department of Health and Human Services. Retrieved May 13, 2008, from http://www.hhs.gov/ohrp/international/

United States Government Printing Office. (1949-1953). *Trials of war criminals before the Nuremberg Military Tribunals under Control Council Law No. 10*. Nuremberg, October 1946 - April 1949 (The Nuremberg Code). Washington, DC: U.S. Government Printing Office. Retrieved May 13, 2008, from https://www.loc.gov/rr/frd/Military_Law/pdf/NT_war-criminals_Vol-II.pdf

Vargas, L. A., & Montoya, M. E. (2009). Involving minors in research: Ethics and law within multicultural settings. In D. M. Mertens & P. E. Ginsberg (Eds.), *The handbook of social research ethics* (pp. 489-506). Thousand Oaks, CA: Sage.

Warwick, D. P. (1983). The politics and ethics of field research. In M. Bulmer & D. P. Warwick (Eds.), *Social research in developing countries* (pp. 315-330). New York, NY: John Wiley & Sons.

World Association for Public Opinion Research (WAPOR). (2011). *WAPOR code of ethics*. Retrieved May 9, 2016, from http://wapor.org/wapor-code-of-ethics/

World Health Organization (WHO). (2000). *Operational guidelines for ethics committees that review biomedical research* (Document TDR/PRD/ETHICS/2000.I). Geneva: WHO. Retrieved May 13, 2016, from http://www.who.int/tdr/publications/documents/ethics.pdf

World Psychiatric Association (WPA). (2008). *Declaration of Madrid* (adopted WPA General Assembly, Madrid, Spain, August 1996; and enhanced by the WPA General Assemblies in Hamburg, Germany on August 8, 1999, in Yokohama, Japan, on August 26, 2002, and in Cairo, Egypt, on September 12, 2005). Geneva: WPA. Retrieved May 5, 2016, from http://www.wpanet.org/detail.php?section_id=5&content_id=48