The SAGE Handbook of
Survey Methodology

SAGE was founded in 1965 by Sara Miller McCune to support the dissemination of usable knowledge by publishing innovative and high-quality research and teaching content. Today, we publish over 900 journals, including those of more than 400 learned societies, more than 800 new books per year, and a growing range of library products including archives, data, case studies, reports, and video. SAGE remains majority-owned by our founder, and after Sara's lifetime will become owned by a charitable trust that secures our continued independence.

Los Angeles | London | New Delhi | Singapore | Washington DC | Melbourne

The SAGE Handbook of
Survey Methodology

Edited by
Christof Wolf, Dominique Joye,
Tom W. Smith and Yang-chih Fu

SAGE reference
Los Angeles | London | New Delhi | Singapore | Washington DC | Melbourne

Los Angeles | London | New Delhi
Singapore | Washington DC | Melbourne

SAGE Publications Ltd
1 Oliver's Yard
55 City Road
London EC1Y 1SP

SAGE Publications Inc.
2455 Teller Road
Thousand Oaks, California 91320

SAGE Publications India Pvt Ltd
B 1/I 1 Mohan Cooperative Industrial Area
Mathura Road
New Delhi 110 044

SAGE Publications Asia-Pacific Pte Ltd
3 Church Street
#10-04 Samsung Hub
Singapore 049483

Editor: Mila Steele
Editorial Assistant: Mathew Oldfield
Production Editor: Sushant Nailwal
Copyeditor: David Hemsley
Proofreader: Sunrise Setting Ltd.
Indexer: Avril Ehrlich
Marketing Manager: Sally Ransom
Cover Design: Wendy Scott
Typeset by Cenveo Publisher Services
Printed and bound by CPI Group (UK) Ltd, Croydon, CR0 4YY

At SAGE we take sustainability seriously. Most of our products are printed in the UK using FSC papers and boards. When we print overseas we ensure sustainable papers are used as measured by the PREPS grading system. We undertake an annual audit to monitor our sustainability.

Library of Congress Control Number: 2015960279

British Library Cataloguing in Publication data

A catalogue record for this book is available from the British Library

ISBN 978-1-4462-8266-3

33

Harmonizing Survey Questions Between Cultures and Over Time

Christof Wolf, Silke L. Schneider, Dorothée Behr and Dominique Joye

INTRODUCTION

Comparing societies across time or space is an important research approach in the social sciences. This approach allows studying how the collective context, such as economic conditions, laws, educational systems or welfare state institutions, shapes the values, attitudes, behaviors and life chances of individuals. Obviously, the validity of this kind of research depends on the quality of the underlying data. If data are gathered using a survey, than sampling, survey mode, question wording, translation, field-work practice or coding all affect the quality and comparability of the resulting data (some of these issues are discussed in Chapters 4, 12, 19, 20 and 23 in this handbook). In this chapter, we focus on strategies to obtain comparative *measures* across different contexts, e.g. countries or societies at different points in time. These strategies – also referred to as harmonization approaches – focus on questionnaire development, including translation, as well as on the processing of the resulting data. We present different approaches to harmonizing measures and discuss several ways to assure and assess comparability of the resulting data.

Questions about the comparability of survey data arise in many different situations. For example, if we aim to analyze social change in the United States and for that purpose draw on the cxurrently available data from the General Social Survey (GSS) waves 1972 to 2014, we may wonder whether the data, or more specifically the variables we are interested in, are comparable over time. However, if we are interested in studying a topic over time that is not covered by the GSS or any other cumulative data set we have to search for different surveys covering the years we wish to study that contain the variables of interest. Should we find such data, then we can attempt to 'harmonize' them, i.e. render them comparable. Obviously, such an exercise rests on the assumption that the variables being harmonized refer to the same theoretical construct.

Ideally, the questionnaire items used in the different surveys should be identical and their meaning should not have changed across *time*. Going international now, a similar assumption has to be made when working with cross-national surveys, like the International Social Survey Programme (ISSP) or the European Social Survey (ESS), which aim to produce variables with identical meaning and understanding across *countries*.

In this chapter, we will look at harmonization mainly from the angle of cross-national surveys, even though much of what we write is also applicable to comparisons over time using monocultural surveys. Our presentation in this chapter rests on a simple model of measurement, as depicted in Figure 33.1 (for more on measurement please see Chapters 14 and 16 in this handbook). Measurement should begin by carefully defining a theoretical concept. Based on this definition, one or more empirical indicators should be selected that are observable and valid representations of the concept. Based on these indicators, a target variable should be defined, i.e. a definition of the variable that should result after data collection and potentially (re-)coding of the data. Only then, the questionnaire item(s) have to be formulated which cover the empirical indicator(s) and which either capture the data directly as intended for the target variable or which can be recoded to do so.

In comparative research, in addition to validity and reliability, the challenge of creating comparative measures has to be met. Broadly speaking, we consider measures to be comparable if similarities or differences in measurements over time or across countries reflect similarities or differences in the measured trait and cannot be attributed to method, i.e. the measurement process or any other aspect of the survey. That is, to ascertain the comparability of measures we have to rule out that differences in method affect the measurement.[1] For inter-temporal analysis, we additionally have to assume that the meaning of questionnaire items does not change over time (see Smith, 2005), and for cross-cultural analysis we equally have to ensure that the meaning of (translated) items does not differ across countries. A more in-depth discussion of these issues is presented in the remainder of this chapter. We first describe common approaches to harmonizing survey questions and survey data. Then we discuss how comparability is assured by input and output harmonization approaches respectively. This is followed by a presentation of different methods to assess the quality of harmonized survey measures and a general conclusion.

HARMONIZATION APPROACHES

Survey methodologists have invested heavily in developing procedures to harmonize

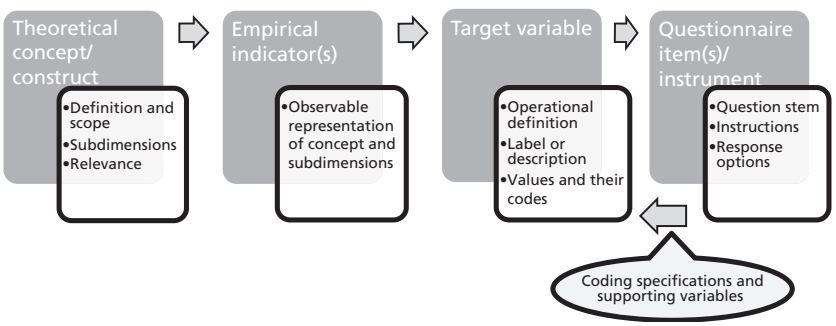

Figure 33.1 From theoretical construct to questionnaire item.

survey data, i.e. procedures aimed at ensuring comparability of survey data from different countries or time points. As Depoutot (1999: 38) puts it: 'Harmonisation is a remedial action to improve comparability'. Two main types of harmonization can be distinguished (see Figure 33.2):

- ex-ante harmonization (with the subtypes input and output harmonization);
- ex-post harmonization (which is always output harmonization).

The main distinction is whether harmonization is an aspect foreseen before data collection, i.e. *ex ante*, and thus gets its place in survey design, or if harmonization is done on pre-existing data, i.e. *ex post*. While ex post harmonization necessarily can only aim at harmonizing the 'output', using the existing measures, ex ante harmonization may aim at harmonizing measurement instruments and procedures, i.e. input harmonization, or aim at the optimal realization of a pre-defined comparable target variable, i.e. *ex-ante output harmonization*. We will describe these approaches to harmonization now in more detail (see also Ehling, 2003).

If a survey is to be conducted in several contexts and the aim is to produce comparative measures, *ex-ante input harmonization* seems to be the natural choice. In this approach to harmonization, the same questionnaire is used in all contexts. Although we restrict our discussion here to questions and questionnaires it is important to note that from the Total Survey Error perspective (e.g. Groves et al., 2009) all elements of a survey should be considered when planning a comparative survey. Ideally, all these elements should be chosen to be the same in all contexts, e.g. sampling, mode of data collection, fieldwork procedures, etc. At first sight, this seems to be clear and input harmonization straightforward to apply. However, one soon realizes that it is not always possible to follow exactly the same protocol for all elements of a comparative survey in all countries.

Let us consider a cross-cultural survey[2]: in this case, the master questionnaire usually has to be translated. Over the last 20 years, survey methodologists have developed specific translation procedures aimed at obtaining questionnaire translations that result in equivalent measurement (Harkness et al.,

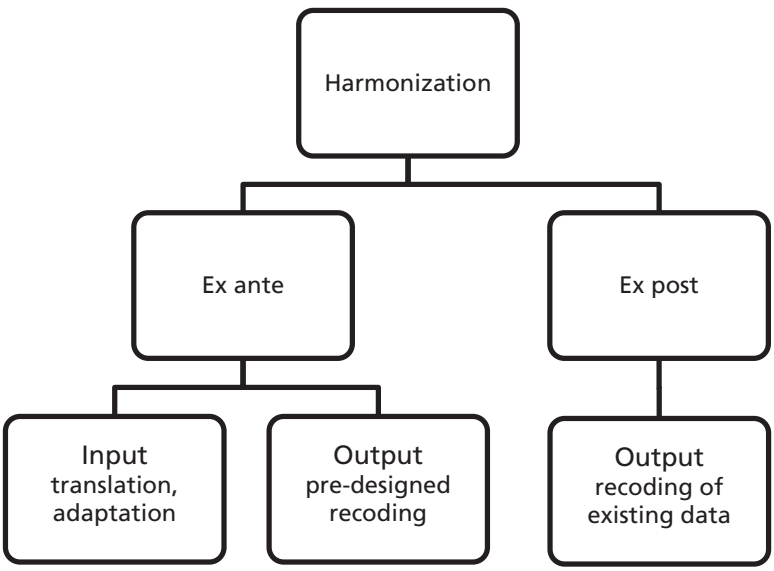

Figure 33.2 Overview of harmonization approaches.

2010; see Chapter 19 in this handbook for a more thorough discussion of this topic). This approach by and large works if the question to be translated does not refer to any issue strongly shaped by specific institutions, culture or history of a country.[3] In the latter case translation is bound to fail. Grais (1999: 54) gives an ingenious example for this (itself an illustration of the historical boundedness of the social realm):

> Mrs. Clinton is the First Lady of the United States. Who is the First Lady of France? Madame Chirac? No doubt, but Madame Jospin might also be a valid candidate given the major role played by a prime minister in France. And who is the First Lady of the United Kingdom? This is where things become complicated. Prince Philip is certainly a possible candidate, but Mrs. Blair is, too, since the Queen of England is not a president and the political role of the prime minister is certainly more important and closer to that of a president. But the notion of First Lady is made up of two concepts: 'first' and 'lady', and from this point of view Queen Elizabeth herself might be a better candidate.

As this example demonstrates, some terms cannot be easily transferred from one country/culture to another. Instead, we have to find more abstract concepts underlying such notions that are relevant in a variety of cultures, find comparable empirical indicators for them and then word questions and response options accordingly. For this example, a possibility would be to find the best term in each language to describe the concept 'spouse of the head of state'. This example also shows that an empirical referent for a concept may not necessarily be found in each country or cultural context. Just think about the Vatican where, for the time being, the existence of a spouse of the head of state is ruled out by ecclesiastical law.

Therefore, in the first step of instrument design for cross-cultural surveys, a common understanding of the theoretical concept to be measured needs to be established and documented, including a working definition and specification of the 'universe' of manifestations for this concept (or its scope). For example, with respect to the concept of 'educational attainment', survey designers need to decide whether they want to include vocational training and/or non-formal continuing education in the concept of education or whether they want to focus on schooling and academic higher education only (see also Hoffmeyer-Zlotnik and Wolf (2003) and Chapter 20 on the measurement of background variables, which are often strongly affected by national culture, history and institutions, in this handbook). The (ideally) cross-cultural survey design team needs to make sure at this point that this theoretical concept is meaningful and relevant in all cultures to be covered by the study.

In the second step, cross-culturally comparable empirical manifestations or indicators for the theoretical concept need to be specified. Theoretical arguments, prior research and knowledge of the countries in question will serve as important guidelines as to which specific indicators to choose (or whether several need to be envisaged and the responses then combined). Importantly, the cross-cultural 'portability' of the indicator(s) needs to be considered here. For example, in order to measure the theoretical concept of 'political participation', democratic countries without compulsory voting can use the indicator of 'voting', but this indicator cannot be used in non-democratic countries (where the indicator does not apply) or democratic countries where voting is mandatory (where the indicator has a different meaning). Participation in demonstrations may be an indicator that would be equally valid in both types of democratic countries but may not have the same meaning in non-democratic societies.

Finally, questionnaire items have to be chosen or designed that capture the indicators of interest. As with surveys in general, multiple (at least three) items should be envisaged whenever possible to facilitate reliability assessment and latent variable modeling, as well as assessment as to whether cross-national differences are

substantive or linguistic/methodological (Smith, 2004; see also Chapters 14, 34 and 39 in this handbook). In order to maximize standardization and thus comparability, the same questionnaire items should be used across countries whenever possible (Ask the Same Question-approach). If this is impossible, cross-cultural equivalents need to be found that equally well represent the theoretical concept across cultures (Ask a Different Question-approach). Within the Ask a Different Question-approach it is useful to further distinguish between questions for which only the response categories have to be adapted – and often later recoded into a common code frame – and questions in which also the question stem has to be adapted to different context-specific circumstances.

Input harmonization, i.e. asking the same question and using the same answering options, may not be possible with respect to all concepts and indicators of interest, such as political parties voted for, or educational qualifications obtained. An adequate approach for cases in which the question stem refers to the same concept and indicator, but response categories need to refer to country-specific manifestations, is *ex-ante output harmonization*. In this approach, one also first agrees on the theoretical concepts that should be measured, as well as comparable empirical indicators. Then, however, one defines a comparable target variable together with the values this variable can obtain, i.e. a code frame, *before* designing the country-specific questionnaire items. This additional process aims to inform the design of country-specific response categories, produced in the next step, so that they can eventually be coded into the target variable. The way how this information is collected is then left to the national teams, or the correspondence between country-specific responses and target variable codes is centrally coordinated to some degree. That is, one agrees on equivalent output measures but allows for variation of the survey questions, especially response options, used to produce them.

Surveys coordinated by Eurostat on behalf of the European Union, such as the EU Labour Force Survey, are extreme cases of *ex-ante output harmonization* in that there is only minimal central coordination of questionnaire (or survey) design. For these surveys, there exist legally binding lists of target variables containing the concept behind the variables and their categories and codes. However, countries are free to collect these data according to their needs, customs and budgets. Thus the mode of data collection varies, there is no common questionnaire, the order, the wording and the answer categories of questions are country-specific, etc. It is even left to countries if they use surveys at all or if the data is obtained from registers and administrative records. Thus, this model optimizes the flexibility and budgets of national statistical institutes at the expense of comparability. For variables such as sex, age or marital status, the country-specific variability of data collection may not affect comparability much. However, the comparability for other measures, such as status of (un-) employment or supervisory status (see Pollack et al., 2009), and in particular subjective indicators, such as self-perceived health or deprivation, will be problematic.

A contrasting example for ex-ante output harmonization, where more central coordination is used, is provided by the ISSP, which uses the *ex-ante output harmonization* approach for its 'background variables', i.e. the socio-demographic questions (a special treatment of background variables in cross-cultural surveys can be found in Chapter 20 in this handbook). For a few years now, the ISSP has offered a blueprint for this section of the questionnaire, which many members find useful.[4] Even more centrally coordinated and thus input harmonized are the ESS, where the Central Scientific Team sets up a substantial set of specifications for countries to follow, and the Survey of Health, Ageing and Retirement in Europe (SHARE), which even employs a centralized CAPI system. The examples show that strict input and *ex-ante*

output harmonization may be considered to be extreme points on a continuum of more or less input- or output-oriented harmonization strategies that may be chosen according to the variables of interest, participating countries, available resources, etc.

The second major approach – *ex-post harmonization* – is characterized by the fact that harmonization takes place only after, sometimes long after, data collection, usually in the context of a secondary data analysis project. The aim of this approach is to take existing data and build an integrated database with variables following a common definition. Typically, this is done to allow for either cross-national comparison or analysis over time (or both). The data sources used for *ex-post harmonization* are typically not produced with comparability in mind, that is to say, achieving comparability was not part of the original survey design. Usually *ex-post harmonization* is the only feasible way to obtain comparable data for the research question at hand at all. A well-known example is the Integrated Public Use Microdata Series (IPUMS) offering harmonized data from the US Census and the American Community Survey, from 1850 onwards (cf. https://www.ipums.org/).

If our aim is to obtain comparative survey data, the most important message is that there is no generally applicable best strategy to reach cross-cultural equivalence of survey measures. Instead, the preferable strategy depends on the concept and indicator we want to measure as well as the basic set-up of the survey, i.e. to what extent the questionnaire can still be modified (*ex-ante harmonization*) or not (*ex-post harmonization*). Whether a questionnaire item can be harmonized ex ante, either by translation, adaptation or ex-ante output harmonization, depends on the theoretical concept, empirical indicator and the specific national, cultural and historical conditions prevailing in the contexts studied. Successful harmonization therefore depends on expert knowledge of these circumstances.

INPUT HARMONIZATION

In this section, we discuss some of the elements that are crucial in making an input harmonization approach successful.[5] Frequently, comparative survey projects begin with the development of a master questionnaire, most often in English, which is subsequently translated into various languages. The major challenge is that the master questionnaire needs to 'get it right' for all multilingual questionnaire versions: That is, all decisions taken at the development stage – in terms of operationalization of concepts, wording of items, choice of answer scales, etc. – need to make sense for all languages and cultures (Harkness et al., 2003a). There is a drawback involved in the sense that questions may get quite general and decontextualized (and thus open to different interpretations) in order to be applicable to a diverse set of countries (Heath et al., 2005). Sometimes, some forms of permitted adaptation (replacement, addition or omission on the item level) are already anticipated in such a master questionnaire (e.g. replacement of examples), but other than that, translated versions should usually follow the master questionnaire when input harmonization is applied. Deviations from the master questionnaire, such as adding or collapsing answer categories, need to be explicitly documented as such[6] – and often require prior approval from a coordinating party in a cross-national study because they may hint at the need to output-harmonize measurement for the indicator in question, or threaten the cross-cultural comparability of the resulting variable.

Various procedures can be applied when producing a common master questionnaire for a cross-national study. Three approaches to master questionnaire development can be distinguished: sequential, parallel and simultaneous development (Harkness et al., 2003b, 2010). The sequential approach essentially does without much cross-national input during questionnaire development. A team of

more or less monocultural researchers develops a questionnaire and submits it for translation. Cross-national requirements and needs only receive attention when translation starts and data collection is often imminent. At this stage, however, it is often too late to correct for the lack of cultural appropriateness of questions (e.g. an item does not fit a country's reality) or of translatability (e.g. heavily idiomatic wording, difficult to translate). Smith (2003) and others (e.g. Jowell, 1998) reject such 'research imperialism' (Smith, 2004: 443) and instead call for parallel questionnaire development, which includes different layers of cross-national research collaboration and collective development work to prevent cultural and linguistic bias. The most direct form of parallel questionnaire development in cross-national collaboration employs a multinational drafting group to decide on concepts, indicators, target variables and items. By bringing together persons from different cultural, institutional and linguistic communities, the hegemony of a single cultural, institutional and linguistic frame of reference can be avoided. However, some form of cultural dominance may (unintentionally) re-enter with the working language in multinational drafting teams (Granda et al., 2010; Harkness et al., 2010). In the same vein, the methodological and scientific background of country representatives may influence the outcome in favor of certain research or cultural traditions and at the expense of others (Bréchon, 2009). Carefully selected research teams, respect vis-à-vis others, intercultural awareness and a common ground in terms of project goals and processes are thus crucial factors to a successful collaboration.

Collaboration between researchers can be organized differently. In small cross-national studies, country representatives may collectively contribute to each stage of development work. In larger studies, it is not uncommon to have smaller multinational core teams doing the actual development work but regularly reaching out to the entire multinational research group (see also Chapter 4 in this handbook). Bréchon (2009) compares the decision-making processes for the Eurobarometer, the ESS, the European Values Study (EVS) and the ISSP. While the Eurobarometer constitutes a survey type on its own due to its political origins – decisions are undertaken by the administrative team in charge, under the supervision of a European Commissioner – the other three surveys are academically run. The decision-making process in the ESS and the EVS is rather centralized and ultimately in the hands of a core team, even though study-wide feedback and collaboration is solicited. In the ISSP annual general meetings and formalized voting procedures provide the backbone of a highly democratic study, in which each member has the same voting power on drafting group composition or question selection, for instance (ISSP, 2012).

Apart from joint discussions, proposition of items for consideration or feedback on others' items, development work can include more empirically-driven forms of cross-national input, in particular advance translation and pretesting.

Advance translation (Dorer, 2011), also called translatability assessment (Dept, 2013), has recently been added to the toolbox of cross-national questionnaire developers; although it was long called for by the research community (Harkness and Schoua-Glusberg, 1998). Advance translation involves producing a translation of a pre-finalized master questionnaire with the explicit goal of spotting potential harmonization problems. These can then be removed or tackled in the final master questionnaire before translation starts. The rationale behind this procedure is that many problems are identified only once an actual translation is attempted. An advance translation does not have to be as thoroughly conducted as a final translation since stylistic fine-tuning for a field study is not needed.

Once a questionnaire is translated it should be subjected to a cognitive pretest (see Chapter 24 in this handbook). Cognitive pretesting assesses to what extent the target group understands the items in the intended

way. Recent years have seen the advent of cross-national cognitive interviewing studies during the development phase of a questionnaire (Fitzgerald et al., 2011; Lee, 2014; Miller et al., 2011). The unique feature of these studies is that they are conducted in a comparable fashion (usually with specifications on what needs to be kept identical across countries and where some leeway in implementation is allowed) and that they are subsequently analyzed with a cross-national perspective in mind. For instance, researchers may be interested in whether 'the system of public services' is understood similarly across countries (Fitzgerald et al., 2011) or whether respondents across countries conceptualize pain in similar ways (Miller et al., 2011). Problems resulting from obvious translation errors or from difficult or ambiguous master question wording, issues of cultural portability and general design issues may all be discovered in these cross-national interviewing studies (Fitzgerald et al., 2011). The latter three types of problems point towards problems for cross-cultural implementation and should lead to the rewording of the master items or even reconsideration of empirical indicators for the concept in question. The diverse types of problems that can be found in cross-cultural interviewing show that monolingual pretesting is usually insufficient for cross-national studies.

Despite its usefulness in detecting comparability flaws, cross-cultural cognitive interviewing may suffer from small case numbers (especially at the country level) and the organizational challenges in setting up these studies. Against this backdrop cross-national web probing has been developed. Cross-national web probing involves asking cognitive probes – similar to those typically used in cognitive interviews – in cross-national web surveys (Braun et al., 2014). For instance, Behr and Braun (2015) followed up on a questionnaire item on 'satisfaction with democracy' with a probe asking for the reasons for having chosen a certain answer value on the scale. Implemented in a cross-national context, probes such as these allow researchers to understand which societal, personal or methodological aspects influence respondents' answers. Answer patterns can then be established, coded, and assessed in terms of comparability. Among the advantages of cross-national web probing are large sample sizes, better country coverage, the possibility to quantify answer patterns across countries, probe standardization across countries, anonymity of answers, and ease of implementation. Even though web probing is affected by probe non-response or mismatching answers, sample size can compensate for these to a great extent. The usefulness of cognitive interviewing and web probing as well as their respective strengths and limitations are discussed in Meitinger and Behr (2016).

While cognitive interviewing or web probing allow for insights into the thought processes of respondents, pilot studies or field tests collect 'real' quantitative survey data and thus allow for preliminary statistical analyses, ranging from non-response distributions over means and correlations to sophisticated equivalence checks. The types and number of items representing a concept as well as available sample sizes will determine the types of analyses that are possible at the piloting stage (see also Chapter 24 in this handbook, in this regard).

A finalized master questionnaire needs to be translated following specific translation and assessment procedures. Parallel translation, team-based review approaches, thorough documentation of problems and decisions, as well as pretesting all can contribute to a measurement instrument that is as comparable as possible to the master version (see Chapter 19 in this handbook). There may still be (traces of) difference left, because connotations of words or slight shifts of meaning due to different semantic systems cannot fully be ruled out. This is but one of the reasons why Smith (2004), amongst others, calls for multiple indicators for the same concept to disentangle societal and linguistic differences.

OUTPUT HARMONIZATION

This section presents strategies for output harmonization. We first describe survey design procedures aimed at ensuring that ex-ante output harmonization results in comparable measures. Then the respective issues for ex-post output harmonization are presented.

Design of Ex-ante Output Harmonized Target Variables and Questionnaire Items

When a survey question stem can be meaningfully translated into different languages and cultures, but the empirical realizations, i.e. the number or even types of response categories vary across countries, translation and adaptation will not suffice to render measures comparable. This is true, for example, for indicators such as 'highest educational qualification obtained', 'political parties' or 'marital status'.[7] In such cases, survey organizers need to employ procedures for ex-ante output harmonization. Successful ex-ante output harmonization relies on a well-structured survey design procedure that, like in the case of input harmonization, requires the collaboration of a cross-national survey design team with expertise in cross-national measurement of the concepts in question, and local experts for the same concepts. Although ex-ante output harmonization is most well-known for its application to demographic and socio-economic variables (see Chapter 20 in this handbook) it may also be relevant to certain attitudinal and behavioral questionnaire items.

For indicators requiring *ex-ante output harmonization*, there is often a tension between validity in a specific cultural context and cross-national comparability. While drafting questionnaire items a country's or culture's specificities (such as institutions, history, legal constructs, concrete objects, symbols and products or customs in the everyday conceptualizations of quantity, space and time) may have to be taken into account in order to accurately measure the concept in question at the national level. At the same time, the cross-cultural survey design team needs to ensure that harmonization into the pre-specified target variable after data collection is possible, in order to achieve comparable measurement. To achieve 'harmonizability', it is recommended to define both the comparable target variable (see "Introduction") as well as the correspondence between country-specific response categories and target variable already during the phase of questionnaire design (i.e. ex ante), rather than only during data processing (i.e. ex post) – otherwise harmonization problems may be identified too late. However, over-adjustment towards the target variable, e.g. by implementing the target categories directly without consideration of potentially important country-specific variations, may lead to oversimplification of the country-specific measure and should thus also be avoided.

After the theoretical concept, an empirical indicator and target variables have been specified, it should now be clear whether translation and adaptation are insufficient to render a measure cross-culturally comparable. If so, it should be determined which elements of the measure require output harmonization and which elements can be input harmonized. For those elements that cannot be input harmonized – usually response categories – specific steps for *ex-ante output harmonization* need to be followed:

1 Target variable specification (adoption or development of a comparative coding scheme or classification).
2 Questionnaire item design, especially response options and their mapping to the target variable.
3 Application of harmonization recodes (after data collection).

Firstly, a comparative target variable to represent the identified concept(s) has to be specified by the survey organizers, including an explicit coding scheme for the

response values. The coding scheme can range from simple scales without official status (e.g. for harmonizing marital status or family or household type across societies, see examples in the ESS, EVS or ISSP) to multi-digit standard classifications such as the International Standard Classification of Occupations (ISCO, see International Labour Organization, 2007) or the International Standard Classification of Education (ISCED, see UNESCO, 2006; UNESCO Institute for Statistics, 2012). When the underlying classification is complex, as in the case of ISCO or ISCED, survey organizers also need to decide whether they want to use and publish the most detailed coding (as is common for ISCO) or some simplified version (as is common – but problematic, see Schneider, 2010 – for ISCED). In the latter case there is a risk that ad-hoc variations of standards are chosen, which decrease standardization across surveys and the usefulness of the variable. In general social surveys, it is useful for researchers if such target variables are rather differentiated in order to allow flexible recoding according to the requirements of the research question. It would be helpful if the possibility of transformation into the official standard is ensured at this stage, especially if official data is to be used as a reference (e.g. for checking sample representativeness or developing adjustment weights). Often survey organizers provide aggregated or otherwise derived variables (either exclusively or additionally) that can be directly used for statistical analyses. These would also benefit from cross-survey standardization.

When adopting an official standard classification then the official documentation (such as the classification document, operational manuals, glossaries and the like) should be made available to all involved in the questionnaire design and coding process. Additionally, it has proven helpful to provide country teams with a coding template and further available resources, such as dedicated short guidelines briefly explaining the concept and its relevance, the theoretical rationale behind measuring it, an explanation of each code in the comparative coding scheme of the target variable, and to include a few examples for its measurement and coding, while illustrating common errors or pitfalls. If only one example is used this single example can be too influential and country teams may remain unaware of the degrees of variation required across country-specific measurement instruments. For this reason, examples should be chosen wisely, not only using the simplest countries as cases but rather different complex ones.

In a second step, each country team develops the country-specific questionnaire item and response categories in such a way as to allow later recoding into the target variable specified in the previous step. Here it is advisable to start from questionnaire items already existing in the respective country. Newly developed or amended questionnaire items should, as usual, undergo pretesting. Especially measures that differ substantially across countries, as in the case of indicators requiring *ex-ante output harmonization*, need to be pretested in every country participating in a comparative survey. At the end of this step, there should be two core outputs: a country-specific questionnaire item (possibly with an input-harmonized question stem and country-specific response categories) and coding instructions for the data processing stage to convert the country-specific response categories to the comparative target variable. In some cases, centralized consultation and sign-off procedures, even though they are time-consuming and pose an extra burden on central and country teams, could be a useful strategy to assure comparability. Content and the purpose of the survey will influence the decision on which target variables survey organizers want to design in this way.

As a final step, after the data have been collected, the coding instructions developed in the previous step need to be executed. This is ideally a merely technical process that is then followed by a quality assessment (see Section

'Assessing the Quality of Harmonized Measures'). When coding open-ended questions, as in the case of occupations or residual 'other …' options, training those coding the information into the classification is strongly advisable. Alternatively, using a dedicated service with established expertise for this difficult task may be an option, but even then one should check the rules and procedures followed by its coders. It is also advisable to assess the degree of inter-coder reliability. For documentation purposes the measurement guidelines and templates should be made available to data users alongside the country-specific questionnaires, coding instructions and country-specific source variables.

In practice, the process described here will often be less straightforward and more iterative with feedback loops from the country-specific items to the definition of the target variable and the coding scheme. As long as this process is clearly documented, it should not jeopardize comparability of the resulting measure.

Deriving Expost Output Harmonized Measures from Existing Data

For analyses of social change based on survey data covering a long or historical time span, or when pooling specialized national surveys on topics for which no cross-national survey exists, it is usually necessary to reconcile existing data sources, i.e. to harmonize ex post (Granda et al., 2010). Prominent examples for this approach are the Luxembourg Income Study (LIS, see Smeeding et al., 1990), the different projects of the Integrated Public Use Microdata Series (IPUMS, see www.ipums.org), the Cross-National Equivalent File of household panel surveys (CNEF, see Burkhauser et al., 2000, http://cnef.ehe.osu.edu/), the International File of Immigrant Surveys (IFIS, see van Tubergen, 2004) or the International Social Mobility File (ISMF, see Ganzeboom and Treiman, 2012, http://www.harryganzeboom.nl/ismf/index.htm). Other extensive harmonization projects include the endeavor of Breen and colleagues who have combined 117 national surveys for cross-national studies of social mobility (Breen et al., 2009, 2010), or the 'Democratic Values and Protest Behavior' project jointly carried out by the Institute of Philosophy and Sociology at the Polish Academy of Sciences and the Mershon Center for International Security Studies at Ohio State University in which over 1,700 national survey files are pooled (see http://dataharmonization.org/).

The founders of CNEF, a set of harmonized household panel surveys from around the world, point to two important aspects of ex-post harmonization, which are the importance of national laws, institutions, history and culture as well as the lack of international standard instruments or coding schemes (which has however somewhat improved since 2000):

> Even the most sophisticated national surveys are unlikely to have cross-national comparability as a survey goal. Hence, while most national surveys use equivalent measures of age and gender, there is no international standard for measuring complex concepts like income, education, health or employment. Thus, researchers interested in doing cross-national work must investigate the institutions, laws and cultural patterns of a country in order to ensure that the variables they create for their analyses are equivalently defined across countries. (Burkhauser et al., 2000: 354)

What harmonizing data means in this context is nicely described by IPUMS International, an attempt to harmonize and integrate microdata from censuses:

> Integration – or 'harmonization'[8] – is the process of making data from different censuses and countries comparable. For example, most censuses ask about marital status; however, they differ both in their classification schemes (one census might recognize only a general category of 'married', while another might distinguish between civil and religious marriages) and in the numeric codes assigned to each category ('divorced' might be coded as a '4' in one census and as a '2' in another). To create an integrated variable for

> marital status we recode the marital status variable from each census into a unified coding scheme that we design. Most of this work is carried out using correspondence tables …[9]

While in principle similar issues as in *ex-ante output harmonization* have to be considered in *ex-post output harmonization*, the process is more driven by the available data than the desired target variables, reflecting general problems of secondary research (Dale et al., 2008). Therefore, *ex-post harmonization* has to live with the fact that survey questions concerning the same underlying concept or indicator may be worded (quite) differently across surveys. Also, it is impossible to change basic survey design features by which the various data sets may differ (e.g. survey mode, sample design, prevalence of proxy interviewing). For example, we know that sensitive questions – which are also not the same across cultures[10] – generate different results depending on the survey mode. Therefore, responses to such questions should be combined with caution if data were collected using different modes in different countries. This does not mean, however, that surveys carried out with the same mode can always be easily combined; after all, the given mode may work differently in different countries (e.g. due to variations in literacy levels, see Smith, 2004). The degree of comparability that can be achieved using *ex-post harmonization* is therefore almost destined to be lower than for surveys designed to be comparative from the outset.

Turning to the *ex-post harmonization* process step by step, the analyst first has to establish, across all included countries, comparable underlying theoretical concepts from existing questionnaires and data sets.[11] Databases documenting questionnaires for large numbers of surveys offering search facilities based on key concepts and keywords can be of great value in this respect.[12] Next, the analyst has to assemble the questionnaire items from the relevant questionnaires and analyze the respective variables from the data sets to find a common denominator in order to code them into one harmonized, cross-nationally comparable target variable. In the case of ex-post output harmonization one thus needs to go from questionnaire item to target variable rather than the other way round, which requires a certain degree of pragmatism.

For variables with 'natural' units measured on a ratio or interval scale such as temperature, currency or distance, re-scaling to a common standard is possible without loss of information (e.g. converting measures from the imperial to the metric system). With scales in attitudinal measures, however, the situation is different. In this case, for ex-post harmonization we especially need to consider the number of scale points, whether the scale is bipolar or unipolar, scale labeling and the availability of a 'don't know' option.

The only case in which a technical transformation of scales is rather straightforward is when all scales have similar even or uneven numbers of response options, and all are unipolar or bipolar (but not a mix of the two). In all other cases, respondents use different scales differently (method effects) and measurement error is also likely to differ, so that the responses from a 10-point scale cannot be just re-scaled to correspond to a 4-point scale. In this example, it is likely that the 10-point scale is more sensitive at the extreme ends than the 4-point scale, potentially leading to different conclusions. In a similar vein, responses from a bipolar 5-point scale cannot be equated with responses from a unipolar 5-point-scale. There are several solutions to this problem, two of which are:

- Dichotomizing items. However, this carries the cost of losing all differentiation – and thus potential for association with other variables – within the two extreme categories.
- Thinking about the items in terms of an underlying latent variable and using scaling techniques adapted to ordinal variables or using external criteria in order to estimate the position of scale values in each context (see Clogg, 1984; Mohler

et al., 1998; Mair and de Leeuw, 2010). However, latent variable modeling requires multiple (at least three) questionnaire items for each concept in each of the surveys, which may be difficult to achieve.

Also the labeling of the scale categories or end points often differ across surveys, as they necessarily do across languages in a cross-national setting: vague scale quantifiers such as 'very', 'strongly', 'pretty', 'not too much' or 'probably' combined with some adjective may not find exact equivalents across countries or they may not tap equivalent cut points in the underlying latent continuum (Mohler et al., 1998). When one scale has a midpoint and/or a 'don't know' option and the other does not, this problem is most obvious. It is therefore hard to judge which scale points are equivalent and how two different scales should be harmonized ex post. Analysis of the covariance structure between different versions of recoded scales might indicate to which degree certain common scales can be considered to produce equivalent data (see Section 'Comparability of Meaning of Multiitem Measures' below).

For categorical variables, this is also a difficult and crucial step because of the risk of losing information when trying to achieve a common coding scheme across countries or data sources: While it may be possible to code standard variables (see Chapter 20 'When Translation is not Enough: Background Variables in Comparative Surveys' in this handbook) or whatever would be ideal for the research question at hand from existing sources, this is not always the case. On the one hand, a cross-nationally identical coding can often only be achieved by aggregating categories in the different data sources, usually following the data with the least differentiation ('the lowest common denominator'), resulting in aggregation error (see also Section 'Assessing Completeness and Comparative Validity of Output-harmonized Measures' below). Thereby it may happen that relevant aspects of the concept in question are 'harmonized away'. Harmonizing data to a highly simplified scheme not based on theoretical considerations risks producing irrelevant or invalid data. Indeed, for certain concepts it may be impossible to arrive at any satisfactory ex-post harmonization that allows valid comparisons over time and/or across countries. On the other hand, existing data can sometimes be harmonized in various ways, and different research questions and theoretical backgrounds will result in differently harmonized variables. Then ex-post harmonization cannot be done 'once and for all' but data need to be re-examined for different research purposes.

To solve both problems, at least for background variables, Granda et al. (2010: 319) describe the approach of hierarchical coding, as applied, for example, in the IPUMS project (see Table 33.1). Its aim is to preserve as much information and thus validity from the original data as possible, especially if the harmonized variables are designed to allow data users to derive more specific measures later on (e.g. when data centers produce time series/longitudinal data files for the community). By using multi-digit codes, differing amounts of detail across studies can be retained while offering cross-nationally equivalent categories. The first one or two digits contain information available in all sources, while further digits provide additional information available in some data sets only. Technically, it often helps to put the response categories of existing variables into a spreadsheet next to each other to establish common boundaries between categories, and then construct the desired target variable. The result is a correspondence table which is very useful also for documentation purposes (see an example from IPUMS concerning marital status, using a hierarchical coding system, Table 33.1). The outcomes of this process are thus – in the optimal case – the assembled source variables from different data sources, their mapping or recode to the comparable coding scheme, as well as the newly constructed, detailed harmonized variable.

Table 33.1 IPUMS Integrated Coding Scheme for Marital Status, slightly simplified

Code	*Target variable*	*Survey 1*	*Survey 2*	*Survey 3*
100	SINGLE/NEVER MARRIED	0 = Single	5 = Single	5 = Single
101	Never married			
102	Single, previously in a consensual union			
103	Single, previously in a religious marriage			
200	MARRIED/IN UNION			
210	Married (not specified)			
211	Civil	7 = Only civil marriage	2 = Only civil marriage	2 = Civil marriage only
212	Religious	8 = Only religious marriage	3 = Only religious marriage	3 = Religious marriage only
213	Civil and religious	6 = Civil and religious marriage	1 = Civil and religious marriage	1 = Civil and religious marriage
214	Polygamous			
220	Consensual union	9 = Living maritally	4 = Consensual or other	4 = Other
300	SEPARATED/DIVORCED/SPOUSE ABSENT			
310	Separated or divorced			
320	Separated			
321	Legally separated	2 = Legally separated (desquitado)	7 = Legally separated (desquitado)	7 = Separated/left (desquitado)
322	De facto separated	1 = Separated	6 = Separated	6 = Separated
330	Divorced	3 = Divorced	8 = Divorced	8 = Divorced
340	Married, spouse absent (n.s.)			
350	Consensual union, spouse absent			
400	WIDOWED	4 = Widower	9 = Widower	0 = Widower
999	UNKNOWN/MISSING	5 = Don't know	0 = No declaration	9 = No answer/left blank

Source: https://international.ipums.org/international/examples/transtable_example.html

ASSESSING THE QUALITY OF HARMONIZED MEASURES

Reliability and validity are the two basic quality aims for survey measures in general. In cross-national research, comparability needs to be added to these two aspects of data quality. This does not only mean comparability in a nominal sense, but includes the requirement for reliability and validity to be comparable across countries (Smith, 2004). This section presents ways to assess the reliability, validity and comparability of harmonized measures in the quest for quality.

Assessing Process Quality

There are two basic approaches to quality assurance for harmonized survey measures: process quality and output quality assessment. To assess the *process* quality of harmonized measures – which is obviously only possible in the case of *ex-ante harmonization*, i.e. for comparative surveys – one looks for documentation of the questionnaire

design and harmonization process indicated as an output of the processes described in the Sections on Input Harmonization and Output Harmonization above.

Only a well-documented, transparent harmonization process allows researchers to check the relationship between theoretical concepts and empirical measures across countries; and in the case of *ex-ante output harmonization*, between country-specific questionnaire items, country-specific variables and harmonized variables. If, for example, it is unclear which educational qualifications are mapped to which categories of a cross-national coding framework such as ISCED, it is impossible to say whether the resulting variable can be regarded as comparable or not across countries (assuming the cross-national coding scheme in principle ensures comparability). Such black boxes do not help when trying to interpret results from statistical analyses. This has e.g. been the case with the education variables in the EU-LFS in the past, when Eurostat published 'only' general survey data quality reports (e.g. for the EU-LFS 2013: Eurostat, 2014). This has fortunately changed with the introduction of ISCED 2011 from 2014 onwards (Eurostat, 2015). Given the high 'documentation burden', lack of documentation of harmonization procedures can be expected to be the rule rather than the exception. The ISSP and ESS both provide detailed templates for documenting ex-ante output harmonization strategies. A final issue concerning documentation is that it is often difficult to find as this information typically is not published using persistent identifiers.

Based on Figure 33.1, questions to be considered in this harmonization process are:

- Is there a common understanding of the underlying theoretical concept, supported by internationally accepted definitions and scope?
- Is the underlying theoretical concept relevant in all countries studied?
- Are the indicators directly comparable across cultures? If not, how was the cross-cultural equivalence of indicators established?
- Is the translation approach and process documented? Was a suitable approach adopted?
- Are all language versions of the questionnaire and (if the survey was interviewer-administered) show cards publicly available? The same for the material given to the interviewee or the instructions to interviewers.
- For variables that were output-harmonized:
  - Is the implemented comparative coding scheme suitable and available (i.e. one that validly measures the underlying theoretical concept across countries)?
  - Is the link between country-specific and harmonized variables clear? Are country-specific variables publicly available, to check whether specifications were followed?
  - Were comparative definitions applied consistently across countries?

These questions are also useful when evaluating *ex-post* output-harmonized data, as they should be equally meticulously documented. However, often these questions are not easy to answer and require some degree of expert knowledge. Specifically for ex-post harmonization the use of indicators reflecting the degree of comparability of variables across contexts was proposed:

> … each variable is assigned a reliability code that represents the degree of cross-national comparability that the surveys permit. For example, a code of '1' indicates that the variables are completely comparable, whereas a code of '4' indicates that there is no comparable variable between the two surveys. These reliability codes are based on direct comparisons of the survey instruments as well as on knowledge of institutional differences across the countries (Burkhauser et al., 2000: 362).

Assessing Reliability

In contrast to process quality assessment, output quality assessment takes the available data and uses statistical methods to measure data quality in quantitative ways. In the remainder of this section, consistency or reliability, completeness, (comparative) validity and comparability will be looked into as quality criteria, while acknowledging that there are other criteria that are less relevant

with respect to harmonization (e.g. timeliness, confidentiality or accessibility).

Reliability of individual harmonized measures at the level of surveys or data sets can be conceptualized as data consistency. Consistency of data across data sources is a necessary but insufficient condition of comparability. Consistency of harmonized survey data across data sources can be checked by comparing descriptive statistics (mean and variation of interval level variables; distributions of categorical variables) across data sources, or with external data, such as censuses or register data containing the same variables (i.e. also coded in the same way, which can sometimes only be achieved by aggregating categories). This can be done even when original country-specific variables are unavailable. In this case, however, no in-depth interpretation is possible beyond diagnosing the degree of (in-)consistency.

For example, the chapters in Schneider (2008) tried to reconstruct the education distributions found in the European Labour Force Survey from national data sources and found many instances in which it was unclear how the harmonized data came about exactly. Ortmanns and Schneider (2015) compared education distributions, all coded in ISCED 97, across four cross-national public opinion surveys over five years using Duncan's dissimilarity index and found major discrepancies that require closer investigation.

Assessing Completeness and Comparative Validity of Output-harmonized Measures

With respect to completeness of output-harmonized data, two aspects can be distinguished: loss of cases or even of whole countries or survey waves due to insufficient 'harmonizability' of the collected data, and loss of information (i.e. variation) in harmonized variables compared to country-specific source variables. There is a trade-off between the two, i.e. analysts often have to choose between more valid harmonized variables but at the cost of losing countries where this coding cannot be achieved, or better country coverage but with fewer valid harmonized variables. Both will be presented in sequence here.

Simple measures of completeness are the proportion of respondents for whom the harmonized target variable can be derived relative to the number of respondents for whom the source variables have non-missing values. For good reasons excluding groups of cases that cannot be harmonized is uncommon as it would distort the sample. Usually, the whole sample is then excluded, e.g. a country or survey wave (or combination of both) with insufficient measurement quality. At the survey level, indicating the proportion of countries or survey waves that could be harmonized would provide a simple measure of completeness. However, harmonized data will usually be pretty complete because completeness in practice often takes priority over comparative validity when defining target variables in ex-post harmonization, reflecting a pragmatic approach to harmonization. In ex-ante harmonization, completeness should not be an issue if measurement instruments are carefully designed to allow deriving the comparative target variable.

For assessing validity in terms of the loss of information occurring from the aggregation of response categories for the purpose of output harmonization, the reference data are typically the original country-specific variables. Detailed external data could, however, also be envisaged. Two ways for analyzing loss of information can be distinguished: the comparison of the original and harmonized variables with respect to a) their *variability*, and b) their *explanatory power* relative to a criterion variable (comparative construct validation). The idea behind both is that harmonization usually entails the aggregation of categories of country-specific variables. This necessarily leads to some loss of information, which may in turn lead to aggregation error in statistical analyses, such as attenuation

of correlations or regression coefficients (as well as confounding bias in coefficients of third variables).

The questions for assessing *comparative* validity are: How much relevant information (and thus validity) is lost through harmonization, and how much does this differ across countries (or data sets)? If the loss of information differs strongly across countries, correlations with the harmonized variable will be attenuated by different degrees in different countries which invalidates cross-national comparisons of correlation and regression estimates.[13] There is a caveat though: If the country-specific source variable itself is measured with a low degree of differentiation and thus may not be particularly valid this analysis may not reveal any loss of information with respect to the harmonized variable for the affected country. Thus, it remains highly important that measures are valid within the survey country in addition to being harmonizable for cross-national comparison.

The 'pure' loss of information or aggregation error is best assessed by comparing a measure of dispersion of the harmonized variable with the same measure of dispersion of the country-specific variable. Granda et al. (2010: 323) provide the following general equation for such a quality measure:

$$Q_{X^{h_i}} = \frac{\text{disp}^{h_i}}{\text{disp}^{o_i}}$$

where disp^{h_i} is the dispersion of the harmonized variable and disp^{o_i} is the dispersion of the original (country-specific) variable, all in data set *i*. Such an analysis was e.g. conducted for educational attainment measures in Schneider (2009: Chapter 6) using the index of qualitative variation (Mueller et al., 1970) as the measure for dispersion, and Granda et al. (2010) provide an example using religious denomination.

This method is especially advisable when the harmonized variable is supposed to be used as a 'multi-purpose' variable where all information contained in the variable may be relevant for one or the other analyst. When a measure is to be evaluated with a specific theoretical background and dependent variable in mind, comparative construct validation may be the more adequate procedure. This additionally allows for distinguishing relevant and irrelevant information, given a specific hypothesized relationship.

With comparative construct validation then, the loss of information is evaluated by comparing the predictive power of one or several differently harmonized variables with the predictive power of the country-specific source variable when predicting a criterion variable in a country-by-country regression analysis. Here the criterion variable needs to be comparable across countries. The analyst can then perform sensitivity analyses, checking how much explanatory power is lost by comparing the (adjusted) determination coefficient R^2 from a regression model using the harmonized variable as a predictor compared to a regression model using the country-specific or source variable as a predictor of the criterion variable. For categorical variables, dummy indicators should be constructed from the harmonized and country-specific variables, respectively. The equation for the respective quality measure would be the same as above, just replacing disp by (adjusted) R^2. This analysis may hint at problematic harmonized variables for the specific relationship in question. Such an analysis was, for example, conducted by Kerckhoff et al. (2002) and Schneider (2010) to compare the quality of differently harmonized education variables.

Comparability of Meaning of Multi-item Measures

A number of statistical procedures are available to quantitatively test the equivalence of meaning when using multiple indicators to measure a latent construct. In comparative surveys, these methods are ideally used during the piloting stage of a comparative

survey, and not only in the analysis stage, to assess the quality of ex-ante harmonization, especially translation and adaptation. However, for ex-post harmonization, they can obviously only be used in the analysis stage and, given the lack of ex-ante procedures in this setting, are amongst the only procedures to empirically assure comparability of measurement.

The most common method used to assess equivalence of a measurement instrument today is multigroup confirmatory factor analysis (MGCFA) presented in detail in Chapter 39 in this handbook (see also the literature cited there). Depending on the type of latent variable of interest and depending on the link between the latent and observed variables other methods, for example extensions of Latent Class Analysis or Item Response Theory, may be more appropriate (for an overview see Wirth and Kolb, 2012). In principle, all these methods test whether constraining certain parameters of a measurement model to be equivalent across groups, e.g. countries, still lead to satisfactory model fit. Depending on the number and kind of parameters that are constrained to be equal, different levels of equivalence can be distinguished. As mentioned, these methods are especially useful to test and develop instruments in a comparative survey design setting.

CONCLUSION

We have distinguished two major approaches to harmonization: ex-ante and ex-post harmonization. For the former, two subtypes were introduced, namely input and ex-ante output harmonization. With respect to most concepts, input harmonization results in the highest level of comparability. However, this approach is not always feasible. If the concepts we are interested in are shaped by national institutions, histories and cultures, strict input harmonization does not work because respondents do not think about these things in the same way across countries. In these cases, we may have to consider varying numbers and types of answering categories, sometimes in addition to adaptations of question wording, and decide how these country-specific variables should be combined or recoded to a harmonized international measure, which we call target variable. That is, we have to apply ex-ante output harmonization. Most comparative surveys show a mixture of these two approaches. Input harmonization is also less feasible if a comparative survey is connected with or built on pre-existing non-comparative surveys in some or all countries. The higher the proportion of ex-ante output harmonization in a comparative survey is, the more national teams will determine its content by, for example, taking questionnaire items or data from existing sources, thereby challenging comparability.

The concrete harmonization strategy followed by a comparative survey, as other aspects of design, affects its other quality dimensions, such as timeliness, response burden or cost. Thus, the concrete mixture of input and ex-ante output harmonization reflects the preferences and constraints of those responsible for the survey with respect to several relevant criteria. It should also be noted that the approach taken for harmonization should correspond to decisions on other design features and standardization measures applied in a comparative survey (for the latter see Lynn, 2003), in accordance with its aims.

Whether theoretical concepts are appropriate across contexts and whether survey measures can be harmonized, either by translation or adaptation of questionnaire items or by recoding during data processing, depends on the specific national, cultural and historical conditions prevailing in the contexts in which we want to run a survey. The same is true for ex-post harmonization. Which target variables the harmonized data file will contain depends not only on the availability of data in the source files but also on our theoretical premises, our research questions and our perspective on the concrete historical and

national context. Thus, harmonization – both ex ante and ex post –is not just a 'mechanical' task of menial recoding work, but depends on expert knowledge of concrete historical, political, social and other national conditions met in the countries and times of interest, and it is not free of normative judgment (see also Chapter 20 in this handbook).

It is therefore paramount that any harmonization project is done in close collaboration of scholars from or at least with in-depth knowledge about the cultural contexts of interest, as well as the theoretical concepts to be studied. The multi-cultural perspective established by such a group helps to prevent resulting measures that are culturally biased, invalid or irrelevant in some country. For the same reason, all steps of a harmonization process should be closely documented. This has the additional advantage that successful solutions from one project may be carried over to other projects thereby supporting cumulative, comparative research.

NOTES

1 Definitions of measurement equivalence are stricter and rely on statistical properties of measures (see Milfont and Fischer, 2010; Steenkamp and Baumgartner, 1998; van Deth, 1998).
2 Most readers probably think about a cross-national survey, but sometimes the cultural heterogeneity within a country (or another relevant target population) is so large that even within the same national context a cross-cultural approach has to be followed.
3 Differences of distributions also often need to be taken into account when adapting a questionnaire to different contexts. Think about a question on religious affiliation: In India, the answering categories should include 'Hindu', while this would not have to be the case in a continental European country.
4 See http://www.gesis.org/fileadmin/upload/dienstleistung/daten/umfragedaten/issp/members/codinginfo/BV_questionnaire_for_issp2014.pdf
5 Other, more general aspects of process quality are of great importance, too, but not covered here (see for example Lyberg and Biemer, 2008; Lyberg and Stukel, 2010).
6 See, for example, for the ESS: http://www.europeansocialsurvey.org/data/deviations_index.html, or for the Comparative Study of Electoral Systems: http://www.cses.org/datacenter/module4/data/cses4_codebook_part2_variables.txt
7 This assumes that in the countries surveyed there are political parties or regulations defining different categories of marital status or there is a formal educational system; assumptions that might be wrong.
8 Since comparative surveys by design and thus ex-ante harmonization are a more recent, extremely costly and thus special endeavor, the term 'harmonization' is often used synonymously with the term 'ex-post harmonization'.
9 https://international.ipums.org/international-action/faq, What are integrated variables?
10 The issue of sensitive questions should also be taken into consideration during input harmonization; truly multi-cultural questionnaire design groups, advance translation and cognitive pretesting can help to identify such questions.
11 This would be much easier if the intended theoretical concepts were routinely documented alongside survey questionnaires, which is not (yet) common practice but highly recommended.
12 Such databases are offered by the Council for European Social Science Data Archives (CESSDA; http://www.cessda.net/catalogue/), the German survey data archive at GESIS – Leibniz Institute for the Social Sciences (http://zacat.gesis.org/), the UK data service (http://nesstar.ukdataservice.ac.uk/) or the Inter-university Consortium for Political and Social Research (http://www.icpsr.umich.edu/icpsrweb/ICPSR/).
13 If, conceptually, the analyst is only interested in certain aspects of a concept (reflected in single categories of a variable), loss of information may not be a concern. For example, when looking at the effects of being divorced, there is no need to keep all possible distinctions on the variable of marital status as predictor variables. In the end, the analyst needs to think carefully about the theoretical concept and target variable and subject the desired variable to quality checks.

RECOMMENDED READINGS

There are surprisingly few texts discussing general approaches to the harmonization of survey data (but see Granda et al., 2010). Some more papers detail the path taken to harmonization in specific surveys, in particular the International Social Survey Programme

(Scholz, 2005), the European Social Survey (Kolsrud and Kalgraff Skjak, 2005), the Cross-National Equivalent File (Lillard, 2013) or household surveys in European official statistics (Ehling, 2003; Körner and Meyer, 2005). Then there are those contributions focusing on a specific variable, such as education (Schneider, 2009), occupation (Ganzeboom and Treiman, 2003) or income (Canberra Group, 2011). Finally, we can recommend the practical advice gained from different projects at the Minnesota Population Center that can be found on the web at http://www.ipums.org/.

REFERENCES

Behr, D., and Braun, M. (2015). Satisfaction with the way democracy works: how respondents across countries understand the question. In P. B. Sztabinski, H. Domanski, and F. Sztabinski (eds), *Hopes and Anxieties: Six Waves of the European Social Survey* (pp. 121–138). Frankfurt/Main: Lang.

Braun, M., Behr, D., Kaczmirek, K., and Bandilla, W. (2014). Evaluating cross-national item equivalence with probing questions in web surveys. In U. Engel, B. Jann, P. Lynn, A. Scherpenzeel, and P. Sturgis (eds), *Improving Survey Methods: Lessons From Recent Research* (pp. 184–200). New York: Routledge.

Bréchon, P. (2009). A breakthrough in comparative social research: the ISSP compared with the Eurobarometer, EVS and ESS surveys. In M. Haller, R. Jowell, and T.W. Smith (eds), *The International Social Survey Programme, 1984–2009* (pp. 28–44). London: Routledge.

Breen, R., Luijkx, R., Müller, W., and Pollak, R. (2009). Nonpersistent inequality in educational attainment: evidence from eight European countries. *American Journal of Sociology*, 114(5), 1475–1521.

Breen, R., Luijkx, R., Müller, W., and Pollak, R. (2010). Long-term trends in educational inequality in Europe: class inequalities and gender differences. *European Sociological Review*, 26(1), 31–48. doi:10.1093/esr/jcp001

Burkhauser, R. V., Butrica, B. A., Daly, M. C., and Lillard, D. R. (2000). The cross-national equivalent file: a product of cross-national research. In I. Becker, N. Ott, and G. Rolf (eds), *Soziale Sicherung in einer dynamischen Gesellschaft*. Festschrift für Richard Hauser zum 65. Geburtstag (pp. 354–376). Frankfurt/Main: Campus.

Canberra Group (2011). *Handbook on Household Income Statistics* (2nd edn). Geneva: United Nations.

Clogg, C. C. (1984). Some statistical models for analyzing why surveys disagree. In C.F. Turner and E. Martin (eds), *Surveying Subjective Phenomena* (Vol. 2, pp. 319–366). New York: Russell Sage Foundation.

Dale, A., Wathan, J., and Higgins, V. (2008). Secondary analysis of quantitative data sources. In P. Alasuutari, L. Bickman, and J. Brannen (eds), *The SAGE Handbook of Social Research Methods* (pp. 520–535). London: Sage.

Depoutot, R. (1999). Quality definition and evaluation. In European Commission (ed.), *The Future of European Social Statistics: Harmonisation of Social Statistics and Quality* (pp. 31–50). Luxembourg: Office of Official Publications of the European Communities.

Dept, S. (2013). Translatability assessment of draft questionnaire items. *Paper presented at the conference of the European Survey Research Association* (ESRA), Ljubljana, SI. Unpublished.

Dorer, B. (2011). Advance translation in the 5th round of the European Social Survey (ESS). *FORS Working Paper Series 2011, 4.*

Ehling, M. (2003). Harmonising data in official statistics: development, procedures, and data quality. In J. H. P. Hoffmeyer-Zlotnik and C. Wolf (eds), *Advances in Cross-National Comparison. A European Working Book for Demographic and Socio-Economic Variables* (pp. 17–31). New York: Kluwer Academic/Plenum Publishers.

Eurostat (2014). Quality report of the European Union Labour Force Survey 2013. *Eurostat Statistical Working Papers*. Luxembourg: Publications Office of the European Union.

Eurostat (2015). ISCED mappings 2014. https://circabc.europa.eu/w/browse/51bfe88f-cb68-4316-9092-0513c940882d [accessed 2015/12/29].

Fitzgerald, R., Widdop, S., Gray, M., and Collins, D. (2011). Identifying sources of error in cross-national questionnaires: application of an error source typology to cognitive interview data. *Journal of Official Statistics*, 27, 569–599.

Ganzeboom, H. B. G., and Treiman, D. J. (2003). Three internationally standardised measures for comparative research on occupational status. In J. H. P. Hoffmeyer-Zlotnik and C. Wolf (eds), *Advances in Cross-National Comparison: A European Working Book for Demographic and Socio-Economic Variables* (pp. 159–193). New York: Kluwer Academic/Plenum Publishers.

Ganzeboom, H. B. G., and Treiman, D. J. (2012). International Stratification and Mobility File: Conversions for Country-specific Occupation and Education Codes. http://www.harryganzeboom.nl/ISMF/ismf.htm [accessed 2015/12/29].

Grais, B. (1999). Statistical harmonisation and quality. In European Commission (ed.), *The Future of European Social Statistics - Harmonisation of Social Statistics and Quality* (pp. 51–114). Luxembourg: Office of Official Publications of the European Communities.

Granda, P., Wolf, C., and 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, and T. W. Smith (eds), *Survey Methods in Multinational, Multicultural and Multiregional Contexts* (pp. 315–334). Hoboken, NJ: Wiley.

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., and Tourangeau, R. (2009). *Survey Methodology* (2nd ed.). Hoboken, NJ: Wiley.

Harkness, J. A., and Schoua-Glusberg, A. (1998). Questionnaires in translation. In J. Harkness (ed.), *ZUMA-Nachrichten Spezial 3: Cross-Cultural Survey Equivalence* (pp. 87–127). Mannheim: ZUMA.

Harkness, J., Mohler, P. Ph., and van de Vijver, F. J. R. (2003a). Comparative research. In J. Harkness, F. J. R. van de Vijver, and P. Ph. Mohler (eds), *Cross-Cultural Survey Methods* (pp. 3–16). Hoboken, NJ: Wiley.

Harkness, J., van de Vijver, F. J. R., and Johnson, T. P. (2003b). Questionnaire design in comparative research. In J. Harkness, F. J. R. van de Vijver, and P. Ph. Mohler (eds), *Cross-Cultural Survey Methods* (pp. 19–34). Hoboken, NJ: Wiley.

Harkness, J.A., Edwards, B., Hansen, S.E., Miller, D.R., and Villar, A. (2010). Designing questionnaires for multipopulation research. In J. A. Harkness, M. Braun, B. Edwards, T. Johnson, L. E. Lyberg, P. Ph. Mohler, B-E. Pennell, and T. W. Smith (eds), *Survey Methods in Multicultural, Multinational, and Multiregional Contexts* (pp. 33–58). Hoboken, NJ: Wiley.

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

Hoffmeyer-Zlotnik, J. H. P., and Wolf, C. (eds). (2003). *Advances in Cross-National Comparison. A European Working Book for Demographic and Socio-Economic Variables*. New York: Kluwer Academic/Plenum Publishers.

International Labour Organization (2007). *Resolution Concerning Updating the International Standard Classification of Occupations*. Geneva: International Labour Organization. http://www.ilo.org/public/english/bureau/stat/isco/docs/resol08.pdf [accessed 2015/12/29].

ISSP (2012). *International Social Survey Programme (ISSP) Working Principles*. http://www.issp.org/uploads/editor_uploads/files/WP_FINAL_9_2012_.pdf [accessed 2015/12/29].

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

Kerckhoff, A. C., Ezell, E. D., and Brown, J. S. (2002). Toward an improved measure of educational attainment in social stratification research. *Social Science Research*, 31(1), 99–123.

Kolsrud, K., and Kalgraff Skjak, K. (2005). Harmonising background variables in the European Social Survey. In J. H. P. Hoffmeyer-Zlotnik and J. A. Harkness (eds), *ZUMA Nachrichten Spezial 11: Methodological Aspects in Cross-National Research* (Vol. 11, pp. 163–182). Mannheim: ZUMA.

Körner, T., and 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 and J. A. Harkness (eds),

ZUMA Nachrichten Spezial 11: Methodological Aspects in Cross-National Research (Vol. 11, pp. 149–162). Mannheim: ZUMA.

Lee, J. (2014). Conducting cognitive interviews in cross-national settings. *Assessment*, 21(2), 227–240.

Lillard, D. R. (2013). Cross-national harmonization of longitudinal data: the example of national household panels. In B. Kleiner, I. Renschler, B. Wernli, P. Farago, and D. Joye (eds), *Understanding Research Infrastructures in the Social Sciences* (pp. 80–88). Zürich: Seismo.

Lyberg, L. E., and Biemer, P. P. (2008). Quality assurance and quality control in surveys. In E. D. de Leeuw, J. J. Hox, and D. A. Dillman (eds), *International Handbook of Survey Methodology* (pp. 421–441). New York: Lawrence Erlbaum Associates.

Lyberg, L., and Stukel, D. M. (2010). Quality assurance and quality control in cross-national comparative studies. In J. Harkness, M. Braun, B. Edwards, T. P. Johnson, L. Lyberg, Ph. Mohler, B.E. Pennell, and T. W. Smith (eds), *Survey Methods in Multinational, Multiregional, and Multicultural Contexts* (pp. 227–249). Hoboken, NJ: Wiley.

Lynn, P. (2003). Developing quality standards for cross-national survey research: five approaches. *International Journal of Social Research Methodology*, 6, 323–336.

Mair, P., and de Leeuw, J. (2010). A general framework for multivariate analysis with optimal scaling: the R package aspect, *Journal of Statistical Software*, 32(9), 1–23.

Meitinger, K., and Behr, D. (2016). Comparing cognitive interviewing and online probing: do they find similar results? *Field Methods*. doi:10.1177/1525822X15625866

Milfont, T. L., and Fischer, R. (2010). Testing measurement invariance across groups: applications in cross-cultural research. *International Journal of Psychological Research*, 3(1), 111–121.

Miller, K., Fitzgerald, R., Padilla, J.-L., Willson, S., Widdop, S., Caspar, R., Dimov, M., Grey, M., Nunes, C., Prüfer, P., Schöbi, N., and Schoua-Glusberg, A. (2011). Design and analysis of cognitive interviews for comparative multinational testing. *Field Methods*, 23(4), 379–396.

Mohler, Ph., Smith, T., and Harkness, J. (1998). Respondent's ratings of expressions from response scales: a two-country, two-language investigation on equivalence and translation. In J. Harkness (ed.), *ZUMA-Nachrichten Spezial 3: Cross-Cultural Survey Equivalence* (pp. 159–184). Mannheim: ZUMA.

Mueller, J. H., Costner, H. L., and Schuessler, K. F. (1970). *Statistical Reasoning in Sociology* (2nd edn). Boston, MA: Houghton Mifflin.

Ortmanns, V., and Schneider, S. L. (2015). Harmonization still failing? Inconsistency of education variables in cross-national public opinion surveys. *International Journal of Public Opinion Research*, online first, doi: 10.1093/ijpor/edv025.

Pollack, R., Bauer, G., Müller, W., Weiss, F., and Wirth, H. (2009). The comparative measurement of supervisory status. In D. Rose and E. Harrison (eds), *Social Class in Europe: An Introdcution to the European Socio-economic Classification* (pp. 138–157). Abingdon: Routledge.

Schneider, S. L. (ed.) (2008). *The International Standard Classification of Education (ISCED-97). An Evaluation of Content and Criterion Validity for 15 European Countries*. Mannheim: MZES. http://www.mzes.uni-mannheim.de/publications/books/isced97.html [accessed 2016/03/02].

Schneider, S. L. (2009). *Confusing Credentials: The Cross-Nationally Comparable Measurement of Educational Attainment*. Oxford: University of Oxford, Nuffield College. http://ora.ouls.ox.ac.uk/objects/uuid:15c39d54-f896-425b-aaa8-93ba5bf03529 [accessed 2015/12/29].

Schneider, S. L. (2010). Nominal comparability is not enough: (in-)equivalence of construct validity of cross-national measures of educational attainment in the European Social Survey. *Research in Social Stratification and Mobility*, 28(3), 343–357.

Scholz, E. (2005). Harmonisation of survey data in the International Social Survey Programme (ISSP). In J. H. P. Hoffmeyer-Zlotnik and J. A. Harkness (eds), *ZUMA Nachrichten Spezial 11: Methodological Aspects in Cross-National Research* (Vol. 11, pp. 183–200). Mannheim: ZUMA.

Smeeding, T. M., O'Higgins, M., and Rainwater, L. (eds) (1990). *Poverty, Inequality, and*

Income Distribution in Comparative Perspective: The Luxembourg Income Study (LIS). Washington D.C.: The Urban Institute.

Smith, T. W. (2003). Developing Comparable Questions in Cross-National Surveys. In J. Harkness, F. J. R. van de Vijver, and P. Ph. Mohler (eds), *Cross-Cultural Survey Methods* (pp. 69–91). Hoboken, NJ: Wiley.

Smith, T. W. (2004). Developing and evaluating cross-national survey instruments. In S. Presser, J. M. Rothgeb, M. P. Couper, J. T. Lessler, E. Martin, J. Martin, and E. Singer (eds), *Methods for Testing and Evaluating Survey Questionnaires* (pp. 431–452). Hoboken, NJ: Wiley.

Smith, T. W. (2005). The laws of studying social change. *Survey Research,* 36(2), 1–5.

Steenkamp, J.-B. E. M., and Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research*, 25, 78–90.

UNESCO (2006). *International Standard Classification of Education: ISCED 1997* (re-edition). Montreal: UNESCO Institute for Statistics.

UNESCO Institute for Statistics (2012). *International Standard Classification of Education - ISCED 2011*. Montreal: UNESCO Institute for Statistics.

Van Deth, J. W. (1998). Equivalence in comparative political research. In J. W. Van Deth (ed.), *Comparative Politics: The Problem of Equivalence* (pp. 1–19). London: Routledge.

Van Tubergen, F. (2004). *International File of Immigration Surveys: Codebook and Machine Readable Data File*. Utrecht: University of Utrecht, Department of Sociology, Interuniversity Center for Social Science Theory and Methodology.

Wirth, W., and Kolb, S. (2012). Securing equivalence: problems and solutions. In F. Esser and T. Hanitzsch (eds), *The Handbook of Comparative Communication Research* (pp. 469–485). New York: Routledge.