

21

Cross-National Web Probing: An Overview of Its Methodology and Its Use in Cross-National Studies

Dorothee Behr¹, Katharina Meitinger², Michael Braun¹, and Lars Kaczmarek^{3,4,5}

¹Survey Design and Methodology, GESIS – Leibniz Institute for the Social Sciences, Mannheim, Germany

²Department of Methodology and Statistics, Utrecht University, Utrecht, The Netherlands

³Library and Archive Services/AUSSDA, University of Vienna, Vienna, Austria

⁴ANU Centre for Social Research and Methods, College of Arts and Social Sciences, Australian National University, Australia

⁵Monitoring Society and Social Change, GESIS – Leibniz Institute for the Social Sciences, Mannheim, Germany

21.1 Introduction

Cross-national surveys are increasingly being set up or used in analyses, and so are cross-cultural surveys in general, including those conducted in different ethnic groups within a single country (Smith 2010; Van de Vijver 2013). It is therefore more pressing than ever to have a sound methodology at one's disposal that allows a researcher to produce equivalent data that can be meaningfully compared across countries and cultures. After all, equivalence is the prerequisite for any sound conclusions. In the words of Johnson (1998, p. 30): "In addition to the traditional reliability and validity requirements for monocultural survey instruments, researchers conducting cross-cultural survey research have the added concern of equivalence. Indeed, cross-cultural research demands a commitment to the establishment of equivalence that is at least equal to the attention routinely reserved for the problems of reliability and validity." From a quantitative perspective and for multiple-item scales, equivalence is often discussed within a three-level framework that distinguishes between configural, metric, and scalar invariance (Meredith 1993; Steenkamp and Baumgartner 1998; Vandenberg and Lance 2000). The lowest level that is required for cross-cultural comparisons is *configural*

invariance where the same constructs are measured in each cultural group. The next-higher level is *metric invariance*; here, the scales have the same unit of measurement even though their origins differ. The highest level is known as *scalar invariance*. With this type of invariance, both measurement unit and origin of scale are the same across cultural groups. Only when scalar invariance is confirmed can researchers study the latent means of the tested constructs. Different types of *bias*, that is, nuisance factors, can reduce equivalence or compromise it altogether. Construct bias, for instance, results from different culture-specific behaviors associated with a construct or only partial overlap of construct definitions across cultural groups. Method bias originates from the sample, from the administration or from the instrument, with bias from the instruments resulting from differential familiarity with survey material or from differential response styles (e.g. acquiescence or extreme response style) across cultural groups. Item bias can be caused by poor translations, ambiguous items or any cultural-specifics being brought to the response process by the respondents (Van de Vijver and Leung 2011). Several qualitative procedures help to prevent or reduce bias during the questionnaire development process – these procedures apply both to multiple-item scales and to single items, but for the latter they are even more crucial since invariance testing procedures such as confirmatory factor analysis cannot be applied to single items. In order to avoid construct bias, intercultural questionnaire development teams work together to operationalize constructs and draft items in such a way that, ideally, equivalence across countries can be achieved. As a next step in the development process, cross-national cognitive interviewing can be implemented. This relatively recent but already well-established procedure comprises the implementation of face-to-face cognitive interviews in several countries with the aim to identify generic, cultural, or linguistic problems early on in the development process (e.g. Miller et al. 2011; Lee 2014; Thrasher et al. 2011). It notably contributes to identifying construct bias or item bias. Even more recent is cross-national web probing, which is the implementation of probing techniques from cognitive interviewing in web surveys in several countries. Also with this procedure, construct or item bias can be identified. Cross-national web probing has been conceived by Braun et al. (2014) in order to collect qualitative data from a large number of respondents from different countries in a standardized and efficient way – data that can then be used for equivalence testing during questionnaire development but also after the fact to assess equivalence and support subsequent data analysis.

It needs to be acknowledged that at about the same time that Braun et al. (2014) developed cross-national web probing, Murphy et al. (2013) pursued web probing activities, too, even though without cross-national or cross-cultural use in mind. Furthermore, Mockovak and Kaplan (2015) explored

potential online applications of cognitive interviewing. In their approach, they asked respondents to think aloud and respond verbally to scripted probes in a mono-cultural web survey. New cognitive pretesting methods or online extensions of cognitive interviewing have therefore been conceived by several research groups at about the same time (see also Edgar 2013).

This chapter focuses on cross-national web probing. We will present (i) the strengths and weaknesses of the method and additionally draw a brief comparison between web probing and cross-national cognitive interviewing, (ii) possibilities of access to respondents in different countries, (iii) the specifics of the web probing implementation, (iv) particularities of translating and coding cross-national probe answers, (v) a selection of substantive results, and finally (vi) an overview of different application scenarios throughout the survey life cycle.

In the following, the term *cross-national* web probing will be used due to the cross-national focus of the research projects that we are mainly summarizing in this chapter. The research projects (2010–2015) were funded by the German Research Foundation (DFG) and aimed at developing and optimizing web probing procedures for cross-national studies. However, the web probing method can be applied in a variety of contexts, including different ethnic groups within a single country or just one ethnic group or country.

21.2 Cross-National Web Probing – Its Goal, Strengths, and Weaknesses

Web probing in general, as we understand it, is “the implementation of probing techniques from cognitive interviewing in web surveys with the goal to assess the validity of survey questions” (Behr et al. 2017, p. 1). The probing techniques referred to in this definition are “additional, direct questions about the basis for responses” given to closed-ended items (Beatty and Willis 2007, p. 289). With the adoption of probing techniques, web probing shares a core feature with cognitive interviewing, at least when considering the probing paradigm of cognitive interviewing. Thus, it comes as no surprise that a set of studies have been set up to compare cognitive interviewing and web probing, albeit, to date, with an exclusive focus on monocultural, monolingual contexts. Meitinger and Behr (2016) found a large overlap between results in a pretesting study in Germany. While interactivity was found to be a great strength of cognitive interviewing, in particular spontaneous respondent comments, web probing helped to prevent local bias in themes by surveying a (geographically) more dispersed set of respondents. Murphy et al. (2013) compared, for a US study, cognitive interviewing with what they called “crowdsourcing in the cognitive interviewing

process.”¹ Comparing the traditional cognitive interview to web surveys that recruited respondents from three crowdsourcing platforms – TryMyUI, Amazon Mechanical Turk, and Facebook – they concluded that cognitive interviewing had its particular strength in enabling spontaneous probing, the fact of which would make it particularly useful for in-depth exploration of (new) items or constructs. The crowdsourcing platforms excelled in terms of speed, geographic dispersion, and partly also motivation of respondents. Thus, Murphy et al. (2013) came to similar conclusions as Meitinger and Behr (2016). While online recruitment and web probing itself certainly take less time, thorough development of a coding scheme and coding of potentially hundreds of answers are time-consuming.

When transferred to the cross-national context, web probing means that probing questions are implemented in cross-national web surveys with the goal to assess comparability of items. It is here where cross-national web probing resembles cross-national cognitive interviewing, but both methods have their own characteristics as well as strengths and weaknesses. Table 21.1 summarizes these, building on a general comparison of web probing versus cognitive interviewing (Behr et al. 2017; Meitinger and Behr 2016) and expanding it to include cross-national particularities.

On the positive side, cross-national web probing is characterized by potentially large sample sizes that allow assessment and comparison of answer patterns or errors across countries. The samples can often be recruited in such a way that respondents are geographically widely dispersed and cover a diverse set of sociodemographic groups. Web probing does not require cognitive interviewers, and hence no recruitment and training of interviewers is needed. Furthermore, the self-administered mode helps to ensure comparability through standardization. On the negative side, the recourse to the online mode means that certain population groups are excluded from this data-collection procedure right from the start. Without an interviewer present, a certain number of respondents may not be motivated enough to answer open-ended probes. The absence of an interviewer also means that, without further interactive follow-ups, spontaneously emerging issues in response behavior or probe answers cannot be dealt with. Moreover, only a limited number of items can be probed; otherwise researchers run the risk of increased nonresponse or even survey break-offs.

In the remainder of the chapter, we will focus on cross-national web probing and refrain from providing further comparisons between cognitive interviewing and web probing. For cross-national studies, an empirical comparison and further delineation between web probing and cognitive interviewing methods still needs to be done.

1 While the term *web probing* focuses on the technique of asking questions, crowdsourcing focuses on the means of recruiting respondents. These may be surveyed using web probing but also using other techniques such as recorded think-aloud.

Table 21.1 Strengths and weaknesses of cross-national cognitive interviewing and cross-national web probing.

Comparison	±	Cross-national web probing	Cross-national cognitive interviewing
Sample size	+	Large sample sizes and good assessment and comparison of answer patterns or errors possible	
	-		Typically small sample sizes, even though larger in the cross-national context than in the national context (Willis 2015)
Coverage of target groups	+		Special target groups, including illiterate, old, poor, ill, etc. persons can be reached
	-	Only online population can be reached	
Geographical and sociodemographic coverage	+	Broader coverage in a country as long as people have internet access	
	-		Typically limited to certain geographical areas and socio-demographic groups in a country
Cognitive interviewers	+	No interviewers are needed, and hence no recruitment or training is necessary	Interviewer can motivate respondents to provide an answer
	-	Interviewer cannot motivate respondents to provide an answer	Careful recruitment and training of interviewers necessary so that cognitive interviewing is done in a similar fashion across countries (Lee 2014; Willis 2015)
Probing	+	Standardized probes → comparability	Flexible, spontaneous probes possible, reacting toward unforeseen, even country-specific issues (Willis 2015)
	-	Standardized probes → potentially insufficient information	If flexible and spontaneous approach prevails → potential lack of comparability, not only within a country, but also across countries
Number of probes	+		Due to a motivating interviewer, more items can be probed
	-	Due to the lack of a motivating interviewer, fewer items can be probed	

Source: Adapted from Behr et al. (2017, p. 3).

21.3 Access to Respondents Across Countries: The Example of Online Access Panels and Probability-Based Panels

The great strength of web probing consists of easy and cost-efficient access to a large sample size. Online access panels are one source to turn to in the search for respondents. Online access panels provide a pool of respondents that have voluntarily signed up to take web surveys at more or less regular intervals. For a fee, researchers can invite these respondents to participate in their surveys. Online access panels are available in many countries, even though not everywhere and certainly with varying levels of quality and varying degrees of penetration in a society. With online access panels, representativity of the general population cannot be achieved since respondent selection typically follows nonrandom recruitment procedures and mostly does not include the offline population. Nevertheless, respondents for one's study can be selected according to quotas (education, age, region, etc.) and thus a balanced sample can be achieved or a specific group targeted. Especially in the cross-national context, quotas for the various country surveys help to ensure that the samples are set up in equivalent ways. Otherwise method bias through dissimilar quotas may impact and reduce the comparability of results obtained.

Given the nonrandom nature of online access panels and the resulting non-coverage of certain population groups (e.g. the elderly, migrants not speaking the language of the panel surveys), these panels are not a panacea. Furthermore, being in the panel does not automatically mean that every panelist is equally able or willing to answer open-ended probing questions. The education level becomes important in this regard, with more-educated respondents being in general more prone to answering open-ended questions and providing longer answers than lower-educated respondents (Oudejans and Christian 2010; Zuell et al. 2015). However, online panels are certainly a useful way for gaining access to respondents in various countries and for increasing sample size and geographical and sociodemographic scope in pretesting or follow-up studies.

Online access panels can especially be useful for random experiments and for general equivalence checks. We have so far used the panels as a kind of add-on study to representative population studies. However, before drawing inferences to these representative surveys, such as the International Social Survey Programme (ISSP), we compared quantitative results, mainly distributions of suspicious closed-ended items, from the ISSP with the online panel data. Only when similar patterns across countries emerged for the item(s) under investigation did we use the online panel data to clarify issues related to peculiarities we found in the ISSP data. Thus, researchers can build in some consistency checks

before using panel data when attempting to retrospectively explain anomalies in regular survey data.²

In some European countries (the Netherlands, Germany, France), representative online panels have been set up that can be used by the research community (Blom et al. 2016); similarly for the United States (Callegaro et al. 2014) and South Korea (Cho et al. 2017). The cross-national web research endeavor is fostered by projects such as the Open Probability-Based Panel Alliance (<http://openpanelalliance.org>), uniting the mentioned panels in Germany, the Netherlands, South Korea, and the United States in one initiative. Any fielding of items in these probability-based panels requires a research proposal and a subsequent review process, which is why any quick access and turnaround is often impossible.

21.4 Implementation of Standardized Probes

Web surveys need to be carefully designed in order to maintain respondents' motivation and reduce detrimental response behavior such as nonresponse. In web surveys, wording, visual features, and overall design all contribute to the respondents' survey experience and are thus decisive in keeping up the motivation and ensuring response quality (Reja et al. 2003). Open-ended questions where respondents write their answers in their own words without any constraints on length (narrative answers) have seen a revival over the past two decades due to their relative ease of implementation in web surveys, and they have produced promising results, particularly when compared to open-ended questions in paper-and-pencil surveys (Oudejans and Christian 2010). Open-ended questions need to be particularly well designed since they are more cognitively demanding than closed-ended questions. They are associated with a higher response burden for respondents due to the lack of answer categories that could guide respondents in answering the question and due to the necessary typing activities. For these reasons, they are more prone to insufficient response in general and nonresponse in particular. Thus, it comes as no surprise that quite a number of studies have looked into design features of open-ended questions and their role for securing the quality of responses as well as other factors influencing response quality. The focus so far has been on different text box sizes, the use of motivational instructions and of follow-ups to open-ended questions, the use of clarification features or the impact of topic interest, and demographic characteristics on response quality (e.g. Denscombe 2008; Holland and Christian 2009; Metzler et al. 2015; Oudejans and Christian 2010; Smyth et al. 2009; Zuell et al. 2015). Since web probes are essentially open-ended questions, we combined findings from these

2 Behr et al. (2017) list further (nonrandom) sources for respondent recruitment.

studies and research strands to design and implement cognitive probes in web surveys. Our design and implementation decisions are presented next.

21.4.1 Probe Placement, Types, Presentation, and Text Box

21.4.1.1 Probe Placement

In our web probing studies, we implemented the embedded or concurrent approach (Willis 2005), that is, probes that are integrated into the usual questionnaire as a direct follow-up to closed-ended items. This approach is essentially a web-based implementation of Schuman's (1966) "random probes." Moreover, in order to disentangle the answering process for the closed-ended item from the answering process for the probe, we found it useful to present the probes on a separate screen. Thus, when answering the closed-ended item, the respondents were not affected by the probe to come, even though a learning process throughout the survey and anticipation of probes could not be ruled out (see Couper 2013, who shows response effects when systematic commenting is allowed). Fowler and Willis (Chapter 18 in this book) additionally tackled the retrospective method in which probes are asked after the web survey is completed.

21.4.1.2 Probe Types and Presentation

We used the following probe types in our studies to identify potential construct or item bias:

- (1) *Category-selection probes* (Prüfer and Rexroth 2005) serve to gain insights into the reasons for a selected answer. An example of a category-selection probe is: "Please explain why you selected 'agree'." A category-selection probe is similar to what is called a *process-oriented probe* ("How did you arrive at that answer?") by Willis (2015), at least when considering attitude items. In our setup with a separate probing screen, we repeated the closed item and the chosen answer category on the probe screen to help recall the item and the answer. Thus, we managed to reduce respondent burden and increase response (Behr et al. 2012). In the case of a numbered scale, we provided respondents not only with the item referred to and the selected answer but also with the range of the answer scale so that the answer itself, being only a number, was put into the larger context.
- (2) *Comprehension probes* serve to uncover the respondents' general understanding of a term. Examples of such probes are: "What ideas do you associate with the phrase 'civil disobedience'? Please give examples" and "What do you consider to be a 'serious crime'?" Comprehension probes are particularly useful in the case of "fuzzy concepts" (Ziegler et al. 2015, p. 1) that lack "clear cut demarcation lines" and thus can particularly affect cross-national research.

- (3) *Specific probes* serve to gather additional information on a detail of an item. A specific probe may be worded as follows: “What particular civil rights did you have in mind when answering the question?” or “Which type of immigrants were you thinking of when you answered the question?”

Figure 21.1 provides screenshots of these three probe types. The use of web probing is certainly not restricted to these probe types or our chosen formulations. But there is one principle that applies to all probe formulations: Probes should be worded in such a way that respondents know what is expected of them (Züll 2016). A spontaneous rewording during the study if respondents do not provide answers as intended is not possible on the web. For cross-cultural cognitive interviewing, research finds that not all probe types work equally well across all cultural groups. For instance, a lack of focused answers and the avoidance of personal views were found among Chinese and Korean respondents in a study by Pan et al. (2010) (see Willis 2015, for an overview of probe types and challenges in the cross-cultural context). Such findings are important to consider and explore further when deciding on probes and when setting up web probing studies beyond the Western context and/or in countries where survey and opinion research is not widely known or used.

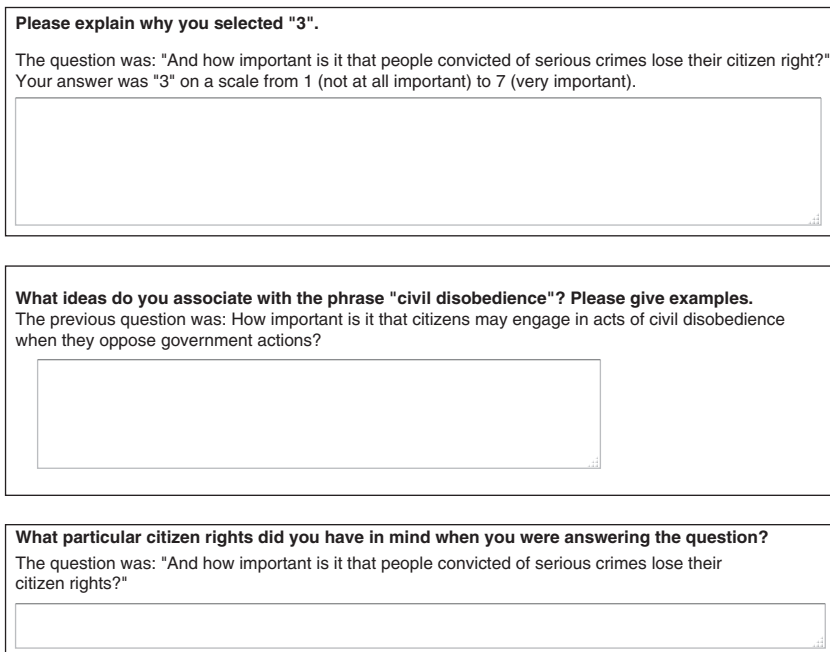


Figure 21.1 Screenshots of examples for category-selection, comprehension, and specific probes.

21.4.1.3 Text Box

An important parameter for open-ended questions is the size of the text box. Numerous studies showed the effect of text box size on answering behavior, with a larger text box producing more text than a smaller box (e.g. Couper et al. 2001; Smyth et al. 2009; Zuell et al. 2015). The same applies to probes as well so that the size of the text box should be adapted to the desired answer type, whether these are examples or more narrative-type answers without length restriction (Behr et al. 2014a). In Figure 21.1, for instance, the category-selection and comprehension probe (first and second) were assigned a larger text box while the specific probe was assigned a smaller text box to cue respondents on the desired length and type of answers.

21.4.2 Sequence of Probes

The aspect of the sequence of probes can be looked at from two different perspectives – sequence of probes for one specific item versus sequences of probes in the entire survey. Neither the cognitive interviewing literature nor research on open-ended questions provided us with answers to these issues so that we had to deal with this without prior input.

21.4.2.1 For One Specific Item

In one experiment, we investigated the best sequence if several probes need to be combined for one specific item. Meitinger et al. (2018) tested two combinations that differed in the sequences of probes: (i) category-selection probe, specific probe, and comprehension probe versus (ii) comprehension probe, specific probe, and category-selection probe. They found that a sequence that had the category-selection probe first increased response rate and motivation and decreased mismatching answers, that is, answers that did not fit the asked probe (e.g. replying with some sort of reasoning answer to a comprehension probe). Interestingly, however, not all effects were equally evident across the countries in our study, which were Germany, Great Britain, Mexico, Spain, and the United States. This at least warrants some caution regarding the uncritical transfer of findings with regard to questionnaire design established in one cultural and linguistic group to other cultural and linguistic groups. More cross-cultural research on probe sequence, nonresponse, and mismatching answer behavior is required.

21.4.2.2 For the Entire Survey

The sequence of probes throughout a survey should not be taken lightly either. Behr et al. (2014a) found that respondents habituated to a specific probe type (e.g. a category-selection probe) in relation to a specific text box size and overall layout when the probe came up repeatedly. The same visual outlook

of a repeatedly occurring probe seemed to have suggested to the respondents the same – known – probe type. Rather than consciously reading the probe question of subsequent probes, respondents answered in terms of their expectations and thus ran the risk of missing any new probe type. In concrete terms: After having been exposed to four or five category-selection probes with identical layout and text box size, when the next probe (a comprehension probe) was reached, the same text box size and overall layout were taken as an indication for a category-selection probe. Thus, we received many responses that in fact matched a category-selection probe but were not a good match for our comprehension probe. We regarded these answers as “mismatching answers” (Meitinger et al. 2018) that were basically useless. In sum, efforts should be directed toward creating a survey where probe type, text box size, and overall layout encourage the respondents to actively read the probe question(s). This could require choosing different text box sizes or layouts for different probe types or conscious decisions on probe sequence. Mismatch conversion does not exist. Particularly with these mismatches, web probing currently reaches its limits. However, with regard to nonresponse we developed first solutions, as described below.

21.4.3 Nonresponse Follow-Ups

Given the high response burden of open-ended questions, probes are particularly prone to nonresponse. In our projects, automated solutions were developed to convert nonrespondents into respondents. The starting point were empirical corpora, first in German, later in English and Spanish, that contained original nonresponse answers to probes. The answers had been coded manually according to different nonresponse categories, as shown in Table 21.2. The various nonresponse answers were used to develop search patterns in the form of regular expressions, that is, generic text strings for automatically identifying patterns in respondents’ answers.³ For instance, the regular expression “`^((be)?cause)? *i? *[a-z]* *do *n.?t *[a-z]* *k* *now*`” finds several variants of English-language “don’t know” answers. When these regular expressions are programmed into the survey software, automated follow-up probes to probe nonresponse can be triggered, possibly with fitting motivating sentences that encourage the respondent to answer the probe despite the first nonresponse. A first set of regular expressions were developed with iterative rounds of testing and validation for German, English, and Spanish. The set of regular expressions as well as the underlying script and technical details are publicly available (Kaczmarek et al. 2017).

³ E.g.: <http://www.regular-expressions.info>.

Table 21.2 Categories of nonresponse.

Category	Type of nonresponse
Category 1	Complete nonresponse: respondent leaves a text box blank
Category 2	No useful answer: response is not a word, e.g. “dfgjh”
Category 3	Don’t know: e.g. “I have no idea,” “DK,” “I can’t make up my mind”
Category 4	Refusal: e.g. “no comment,” “see answer above”
Category 5 ^{a)}	Other nonresponse: responses that are insufficient for substantive coding, e.g. “my personal experience,” “it depends,” “just do,” “just what it is”
Category 6 ^{a)}	One word only: respondent just writes a single word, e.g. “economy”
Category 7	Too-fast response: respondent takes less than two seconds to answer

a) Answers of categories 5 and 6 may for some research questions count as a substantive response.

21.5 Translation and Coding Answers to Cross-Cultural Probes

As we were, in our studies, mainly interested in interpretation patterns across countries, our analyses of web probing almost exclusively relied on a thematic approach. That is, rather than identifying whether certain errors occurred (DeMaio and Landreth 2004; Fitzgerald et al. 2011; Willis and Zahnd 2007), we coded themes that were mentioned in order to detect (non-)equivalent – or biased – patterns across countries: for instance, immigrant groups that respondents thought of in different countries or aspects that made respondents proud of their country. One of the key questions in the analyses was whether team members sufficiently understood all the languages of the study to both develop the coding scheme and code the responses. For certain languages in our first set of studies, this was not the case so that we commissioned professional translators for the task of translating open-ended probing answers into the project language (in our case, German). By specifically instructing the translators for the task (overall goal of research, leeway in translation, examples for required comments) and providing them with space for additional comments, we attempted to narrow the impact of translation on the coding (Behr 2015). We acknowledge, however, that we cannot fully exclude impact on the coding: Whether coding of translated responses leads to different conclusions than coding of original responses, and whether there is a best-practice approach is still an unsolved research question. Being able to understand the responses is needed not only for coding, but also for the development of the coding scheme – if this is question-specific rather than based on generic error types. Especially in an inductive approach where the responses suggest the categories

of a coding scheme, understanding responses from *all* countries is needed in order to develop a balanced coding scheme that takes into account the respective country narratives. Otherwise categories and illustrative examples may favor some countries' themes and perspectives over others and thus introduce some form of bias.

Since coding scheme development and coding (and possibly prior translation) are time-consuming, (semi)-automatic approaches to coding are worthwhile to look at, even though they require a sufficiently large manually coded sample size (about 500 responses) to train a learning algorithm (Schonlau and Couper 2016). Automatic coding has the effect, though, that comments made by the translators, for instance on particularities of translated open-ended responses, cannot be taken into account in code assignments. For a "quick and dirty" problem spotting, one might also try out visualization or text/content analysis tools (see, e.g., wordle.net or tools listed under <http://tapor.ca/tools>) to gain a quick and rough overview of the data, for instance with regard to the frequency of words across countries. If the tools are used on an external server rather than on one's organization's server, data privacy and confidentiality should always be guaranteed, though.

21.6 Substantive Results

The research aim of our group, once the web design challenges were overcome, was to assess equivalence in cross-national surveys by using the collected web probing answers. We have so far tackled items in existing surveys that were identified as problematic during statistical equivalence testing (e.g. inconsistencies or lack of higher levels of invariance) or items that have repeatedly provoked calls in the research community for further research to elucidate their meaning in cross-national contexts.

21.6.1 Issues in Statistical Analyses

In this first line of research, that of shedding light on problematic and suspicious data, we conducted research on the "rights in a democracy scale" of the ISSP 2004, in particular on the item "How important is it that citizens may engage in acts of civil disobedience when they oppose government action?" In the ISSP data, this item showed both high item nonresponse and inconsistent results with regard to the other items in the six-item scale (response scale running from 1 to 7). The inconsistent results – for Canada, Denmark, Germany, Hungary, Spain, and the United States – were as follows: For the index (the response mean across variables) of the first five items in the battery (e.g. importance of all citizens having adequate standard of living or importance of protection of minority rights), the index was quite similar for all countries

(6.2–6.6). However, the last item, the civil disobedience item, markedly divided the countries into two groups, with the mean for Canada, Denmark, and the United States being particularly low (3.8–4.1) and the mean for the other countries ranging between 5.0 and 5.5. To understand what may have driven these results, we implemented the scale in our cross-national web survey in Canada, Denmark, Germany, Hungary, Spain, and the United States (n between 507 and 538) and had the scale followed by a probe split after the civil disobedience item. In each country, respondents were randomly assigned to receive one of two probe versions. Half of the respondents received a category-selection probe inquiring after the reasons for the selected answer and the other half received a comprehension probe asking for the ideas respondents associated with “civil disobedience.” The striking pattern for the scale from the ISSP could be replicated in the web survey data, and the probe answers showed that lower support for civil disobedience in the United States and Canada was partly “real” (due to a higher level of trust in politicians) and partly a methodological artifact due to different associations with the concept of civil disobedience. Respondents from the United States and Canada in particular associated violent actions with civil disobedience, while this answer pattern was much less prevalent in the other countries. We concluded that item bias in the form of different meanings attached to the item’s key term compromised equivalence (Behr et al. 2014b).

Meitinger (2017) took the ISSP 2013 as a starting point. She used multiple group confirmatory factor analysis (MGCFA) to test measurement invariance for five items measuring nationalism and constructive patriotism in five countries: Germany, Great Britain, Mexico, Spain, and the United States. Although the study could confirm metric measurement invariance, (partial) scalar invariance tests failed in MGCFA, which precluded a cross-national comparison of the latent means of the constructs. We then implemented the items in a web probing study and this enabled Meitinger to investigate how three of the five items were understood across countries, namely “And how proud are you of [country] with regard to the way democracy works?” “[...] to its social security system?” and “[...] to its fair and equal treatment of all groups in society?” Based on the probe answers, she found that the lack of scalar invariance could be explained by a major misunderstanding of the term “social security system”⁴ [sistema de seguridad social] by 39% of Mexican respondents (“security on the streets” rather than “state benefits”) and by differences in the perceived scope of the various terms for “social security” in the different languages, pushing respondents’ understanding in one or the other direction.

4 A social security system is a set of measures that fulfill basic needs for citizens. It is put in place by the government and examples include (monetary) benefits for people who are unemployed, require health care, need welfare, are retired, or have children.

21.6.2 Testing Questionable Items

The second line of research, that of focusing on critical or questionable items or terms in general, was followed by Behr and Braun (2015) when looking into respondents' reasons for rating the functioning of democracy in their respective countries in a positive or negative way. In the European Social Survey, the item is worded as follows: "How satisfied are you with the way democracy works in your [country]?" The item, in this or a similar wording, is widely used in cross-national studies but is nevertheless highly controversial due to vagueness, context-dependency, and the fact that it measures a complex concept with just one item (Ariely and Davidov 2011; Canache et al. 2001; Linde and Ekman 2003). We implemented the item in a cross-national web survey in Canada, Denmark, Germany, Hungary, Spain, and the United States and followed up with a category-selection probe asking for the reasons that respondents had in mind when selecting their answers (on a scale from 1 [extremely dissatisfied] to 11 [extremely satisfied]). While a variety of reasoning patterns or dimensions could be found, most notably on the levels of government output, governance, and the political system as such, these dimensions played a role in *all* countries of our study. In particular, probe responses expressing dissatisfaction were strongly linked to output and governance, whereas the political system assembled probe responses that expressed both satisfaction and dissatisfaction. These differences also fitted to the respective country results regarding the closed-ended item. For instance, Denmark stood out as the most satisfied of the five countries, as measured by the closed-ended item. At the same time, satisfaction with the political system and, to a lesser degree, governance was mostly a probe theme put forward by Danish respondent. Given the results, we concluded that some form of comparability did indeed exist for the item even though the item itself could not be nailed down to a single dimension. If some countries in our study had exclusively relied on assessing government output and others on governance (regardless of satisfaction or dissatisfaction with these aspect), we would have been less positive in our conclusion.

In another study, Braun et al. (2013) looked into the meaning of "immigrants" in a cross-national context. Although the term "immigrant" can more or less easily be translated into other languages, this does not mean that the groups associated with "immigrants" are necessarily comparable across countries (see Heath et al. 2005, for a similar concern expressed toward the term "people from poorer countries"). In order to gain insights into respondents' understanding, we took items from the ISSP 2003 module on National Identity, here in particular the item scale on xenophobic attitudes: "immigrants increase crime rates, ... are generally good for country's economy, ... take jobs away from people who were born in [country], ... improve society by bringing in new ideas and cultures." We rotated these items and asked for the first item in each experimental condition what immigrant groups the respondents had in mind when answering the closed-ended item. Braun et al. (2013) found for Canada,

Denmark, Germany, Hungary, Spain, and the United States that respondents thought of the most visible immigrant groups in their respective countries. For instance, in Germany, these were mostly the Turkish immigrants and, in Canada, these were immigrants of Asian origin. As such, immigrant reality was more or less captured in a comparable way, even though the nationalities or ethnic groups of migrants were different across countries.

These examples of substantive analyses show that the different social contexts determine in the end how a translation is understood. Meaning is nested in many contexts, the questionnaire context, the respondents themselves, and in particular in the socio-cultural context in each country: “Meanings and thought patterns do not spontaneously occur within the confines of a respondent’s mind, but rather those meanings and patterns are inextricably linked to the social world [...]” (Miller and Willis 2016, p. 212). This is why cognitive interviewing has become such a useful tool in cross-national questionnaire design and refinement of translations; this is also why web probing with its own strengths and weaknesses is likely to become a promising supplementary method in cross-national survey research.

21.7 Cross-National Web Probing and Its Application Throughout the Survey Life Cycle

In the research presented here, we have mainly implemented web probing as a follow-up study to a main survey to understand what may have caused non-equivalence in items. A use of web probing for pretesting purposes is equally possible. However, any use should be seen in light of and coordinated with well-established pretesting methods, in particular with cognitive interviewing with which web probing shares the probing questions. The integration and the interplay between web probing and cognitive interviewing is currently a matter of debate and testing in general survey methodology, as was described in Section 21.1 of this chapter (see also Behr et al. 2017). The current tendency, if both cognitive interviewing and web probing are possible, is to have cognitive interviewing first since it allows in-depth and interactive exploration of items. More targeted web probing may follow to assess the prevalence of themes or issues in a larger and more geographically dispersed population. For web probing in this scenario, the researcher should already have a thorough understanding of the item – and potential issues and hypotheses in mind – in order to determine and word the probe(s). This sequence, cognitive interviewing first, followed by web probing, can also be implemented with closed-ended probes. Scanlon (Chapter 17 in this book) follow an approach where the cognitive interviewing results help to word closed-ended probes for a subsequent web survey. These closed-ended probes have the advantage that demanding or burdensome open-ended questions,

when seen from a respondent's perspective, and time-consuming coding scheme development and manual coding, when seen from a researcher's perspective, and are not needed. The combination of cognitive interviews and closed-ended probes in field tests (albeit not in web-based field tests) was also followed by Miller and Maitland (2010) (cited by Baena and Padilla 2014) to assess the range of meanings attached to anxiety and the respective prevalence among respondents in Kazakhstan, Cambodia, Sri Lanka, Maldives, Mongolia, and the Philippines.

Web probing as a pretesting method may also become interesting if cross-national cognitive interviewing is not viable (e.g. lack of cognitive interviewers in some countries, time constraints) or should be supplemented with web probing in additional countries to increase the spread of cultural and linguistic groups (Behr et al. 2017). It goes without saying, however, that (besides the practical constraints) the research questions, the desired probes one has in mind (including their complexity and likelihood to trigger follow-up probes) and the target group will eventually decide which method and in which combination to use.

If the main survey is a web survey, one could imagine implementing probes for selected items and selected respondents. Already in the mid-1960s, Schuman brought up the idea of "random probes" whereby randomly selected respondents receive probes for selected questions (e.g. 10 probes per respondents). Schuman argues that

[t]hrough qualitative and quantitative review of random probe responses the survey researcher has an opportunity to increase his own sensitivity to what his questions mean to actual respondents [...] In research in other cultures—and under some conditions in one's own culture—it forms a useful supplement to standard attitude survey methods.
(Schuman 1966, p. 222)

Response burden and potential effects on closed-ended items should be considered in such a design, though.

The use of web probing in cross-national follow-up studies was described earlier, notably in Section 21.5, where substantive analyses were presented. These follow-up studies can help to explain statistical inconsistencies and problems and gather additional qualitative data to aid analysis.

Regardless of the stage of use (pretesting, main study, follow-up study), the web probing data collected can serve to foster systematic mixed-methods approaches in cross-national research. Mixed-methods research refers to a combination of qualitative and quantitative methods and their integration to reach the research objective (Baena and Padilla 2014).

Respondent perceptions and quantitative approaches can be reconciled by this type of research and limitations associated with each research paradigm

offset. Van de Vijver and Chasiotis (2010) have already made a plea for cross-cultural mixed-methods research in general; Baena and Padilla (2014) have echoed this plea, thereby focusing on cognitive interviewing in connection with quantitative methods. In this chapter, we are recommending web probing to be considered in mixed-method research, too.

21.8 Conclusions and Outlook

In this chapter, we have described the methodology and use of cross-national web probing by drawing on the major findings from two research projects conducted by the authors of this chapter. Where available, we supplemented our findings with approaches and applications by other researchers in order to guard against “project bias.” However, cross-national web probing is a new endeavor; therefore, the literature both on theory and practice is still scarce. While the probing technique as such has been adopted from cognitive interviewing, design decisions have mainly been driven by advances in web survey design and here in particular current knowledge on the design of open-ended questions. Findings and innovations in these areas should be duly considered when embarking on web probing research activities.

We have implemented our studies mainly in Western countries (Mexico is an exception), and with a limited number of probe types. In terms of future research, it would be highly useful to test the web probing approach with a greater spread of countries and particularly in non-Western countries. The research questions should inquire whether web probing can elicit meaningful answers from a diverse set of cultural groups, and whether limitations for (certain) probe types in certain cultures exist (similarly to research conducted in cross-cultural cognitive interviewing, see Willis 2015). After all, what is ultimately needed is a method that, without introducing method bias itself, helps to confirm equivalence on the one side or uncover construct or item bias on the other side. Attention should also be paid to conducting comparative studies between cross-national cognitive interviewing and web probing to further delineate these types of methods and provide clearer guidance on when which of these methods can or should be used.

As Van de Vijver and Chasiotis (2010) point out, systematic mixed-methods studies in cross-national and cross-cultural research are still wanting. Due to its relative ease of implementation, web probing can contribute to increasing the number of these studies, whether at the pretesting, main study or follow-up stage. Cross-national web probing is in a unique position, besides cross-national cognitive interviewing, to take into account the socio-cultural contexts of respondents and its influence on understanding and answering (translated) survey questions.

Acknowledgments

This research was funded by the German Research Foundation (DFG) as part of the PPSM Priority Programme on Survey Methodology (SPP 1292) (project BR 908/3-1) and in a follow-up project (BR 908/5-1). Researchers on the projects: Michael Braun, Wolfgang Bandilla, Lars Kaczmirek (grant applicants), Dorothee Behr, and Katharina Meitinger.

References

- Ariely, G. and Davidov, E. (2011). Can we rate public support for democracy in a comparable way? Cross-national equivalence of democratic attitudes in the World Value Survey. *Social Indicators Research* 104: 271–286. <https://doi.org/10.1007/s11205-010-9693-5>.
- Baena, I.B. and Padilla, J.-L. (2014). Cognitive interviewing in mixed research. In: *Cognitive Interviewing Methodology* (eds. K. Miller, V. Chepp, S. Willson and J.-L. Padilla), 133–152. Hoboken, NJ: Wiley.
- Beatty, P. and Willis, G. (2007). Research synthesis: the practice of cognitive interviewing. *Public Opinion Quarterly* 71: 287–311. <https://doi.org/10.1093/poq/nfm006>.
- Behr, D. (2015). Translating answers to open-ended survey questions in cross-cultural research: a case study on the interplay between translation, coding, and analysis. *Field Methods* 27: 284–299. <https://doi.org/10.1177/1525822X14553175>.
- Behr, D. and Braun, M. (2015). Satisfaction with the way democracy works: how respondents across countries understand the question. In: *Hopes and Anxieties. Six Waves of the European Social Survey* (eds. P.B. Sztabinski, H. Domanski and F. Sztabinski), 121–138. Frankfurt am Main: Lang.
- Behr, D., Kaczmirek, L., Bandilla, W., and Braun, M. (2012). Asking probing questions in web surveys: which factors have an impact on the quality of responses? *Social Science Computer Review* 30: 487–498. <https://doi.org/10.1177/0894439311435305>.
- Behr, D., Bandilla, W., Kaczmirek, L., and Braun, M. (2014a). Cognitive probes in web surveys: on the effect of different text box size and probing exposure on response quality. *Social Science Computer Review* 32: 524–533. <https://doi.org/10.1177/0894439313485203>.
- Behr, D., Braun, M., Kaczmirek, L., and Bandilla, W. (2014b). Item comparability in cross-national surveys: results from asking probing questions in cross-national web surveys about attitudes towards civil disobedience. *Quality & Quantity* 48 (1): 127–148. <https://doi.org/10.1007/s11135-012-9754-8>.
- Behr, D., Meitinger, K., Braun, M. et al. (2017). Web probing – implementing probing techniques from cognitive interviewing in web surveys with the goal to

- assess the validity of survey questions. Mannheim, GESIS – Leibniz-Institute for the Social Sciences (GESIS – Survey Guidelines). doi:https://doi.org/10.15465/gesis-sg_en_023.
- Blom, A.G., Bosnjak, M., Cornilleau, A. et al. (2016). A comparison of four probability-based online and mixed-mode panels in Europe. *Social Science Computer Review* 34: 8–25. <https://doi.org/10.1177/0894439315574825>.
- Braun, M., Behr, D., and Kaczmirek, L. (2013). Assessing cross-national equivalence of measures of xenophobia: evidence from probing in web surveys. *International Journal of Public Opinion Research* 25: 383–395. <https://doi.org/10.1093/ijpor/eds034>.
- Braun, M., Behr, D., Kaczmirek, L., and Bandilla, W. (2014). Evaluating cross-national item equivalence with probing questions in web surveys. In: *Improving Survey Methods: Lessons from Recent Research* (eds. U. Engel, B. Jann, P. Lynn, et al.), 184–200. New York: Routledge.
- Callegaro, M., Baker, R., Bethlehem, J. et al. (2014). Online panel research. History, concepts, applications, and a look at the future. In: *Online Panel Research: A Data Quality Perspective* (eds. M. Callegaro, R. Baker, J. Bethlehem, et al.), 1–22. Chichester: Wiley.
- Canache, D., Mondak, J.J., and Seligson, M.A. (2001). Meaning and measurement in cross-national research on satisfaction with democracy. *Public Opinion Quarterly* 65: 506–528.
- Cho, S.K., LoCascio, S., Lee, K.-O. et al. (2017). Testing the representativeness of a multimode survey in South Korea: results from KAMOS. *Asian Journal for Public Opinion Research*. 4 (2): 73–87. <https://doi.org/10.15206/ajpor.2017.4.2.73>.
- Couper, M.P. (2013). Research note: reducing the threat of sensitive questions in online surveys? *Survey Methods: Insights from the Field* <https://doi.org/10.13094/SMIF-2013-00008>.
- Couper, M., Traugott, M., and Lamias, M.J. (2001). Web survey design and administration. *Public Opinion Quarterly* 65: 230–253.
- DeMaio, T.J. and Landreth, A. (2004). Do different cognitive interview techniques produce different results? In: *Methods for Testing and Evaluating Survey Questionnaires* (eds. S. Presser, J.M. Rothgeb, M.P. Couper, et al.), 89–108. Hoboken, NJ: Wiley.
- Denscombe, M. (2008). The length of responses to open-ended questions: a comparison of online and paper questionnaires in terms of a mode effect. *Social Science Computer Review* 26: 359–368. <https://doi.org/10.1177/0894439307309671>.
- Edgar, J. (2013). Self-administered cognitive interviewing. 68th Annual Conference of the American Association for Public Opinion Research, Boston, MA.

- 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.
- Heath, A., Fisher, S., and Smith, S. (2005). The globalization of public opinion research. *Annual Review of Political Science* 8: 297–333. <https://doi.org/10.1146/annurev.polisci.8.090203.103000>.
- Holland, J.L. and Christian, L.M. (2009). The influence of topic interest and interactive probing on responses to open-ended questions in web surveys. *Social Science Computer Review* 27: 196–212. <https://doi.org/10.1177/0894439308327481>.
- Johnson, T.P. (1998). Approaches to equivalence in cross-cultural and cross-national survey research. In: *ZUMA-Nachrichten Spezial 3: Cross-Cultural Survey Equivalence* (ed. J. Harkness), 1–40. Mannheim: ZUMA.
- Kaczmarek, L., Meitinger, K., Behr, D. (2017). Higher data quality in web probing with EvalAnswer: a tool for identifying and reducing nonresponse in openended [sic!] questions. *GESIS Papers* 2017/01. <http://nbn-resolving.de/urn:nbn:de:0168-ssoar-51100-0>.
- Lee, J. (2014). Conducting cognitive interviews in cross-national settings. *Assessment* 21: 227–240. <https://doi.org/10.1177/1073191112436671>.
- Linde, J. and Ekman, J. (2003). Satisfaction with democracy: a note on a frequently used indicator in comparative politics. *European Journal of Political Research* 42: 391–408. <https://doi.org/10.1111/1475-6765.00089>.
- Meitinger, K. (2017). Necessary but insufficient: why measurement invariance tests need online probing as a complementary tool. *Public Opinion Quarterly* 81: 447–472. <https://doi.org/10.1093/poq/nfx009>.
- Meitinger, K. and Behr, D. (2016). Comparing cognitive interviewing and online probing: do they find similar results? *Field Methods* 28: 363–380. <https://doi.org/10.1177/1525822X15625866>.
- Meitinger, K., Braun, M., and Behr, D. (2018). Sequence matters in online probing: the impact of the order of probes on response quality, motivation of respondents, and answer content. *Survey Research Methods* 12: 103–120. <https://doi.org/10.18148/srm/2018.v12i2.7219>.
- Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika* 58: 525–543.
- Metzler, A., Kunz, T., and Fuchs, M. (2015). The use and positioning of clarification features in web surveys. *Psihologija* 48: 379–408. <https://doi.org/10.2298/PSI1504379M>.
- Miller, K. and Willis, G.B. (2016). Cognitive models of answering processes. In: *The SAGE Handbook of Survey Methodology* (eds. C. Wolf, D. Joye, T.W. Smith and Y. Fu), 210–217. London: Sage.
- Miller, K., and Maitland, A. (2010). A mixed-method approach for measurement construction for cross-national studies. *Joint Statistical Meetings*, Vancouver, CA.

- Miller, K., Fitzgerald, R., Padilla, J.-L. et al. (2011). Design and analysis of cognitive interviews for comparative multinational testing. *Field Methods* 23: 379–396. <https://doi.org/10.1177/1525822X11414802>.
- Mockovak, W., and Kaplan, R. (2015). Comparing results from telephone reinterview with unmoderated, online cognitive interviewing. 70th Annual Conference of the American Association for Public Opinion Research (AAPOR), Boston, MA
- Murphy, J., Keating, M., and Edgar, J. (2013). Crowdsourcing in the cognitive interviewing process. 2013 Federal Committee on Statistical Methodology (FCSM) Research Conference.
- Oudejans, M. and Christian, L.M. (2010). Using interactive features to motivate and probe responses to open-ended questions. In: *Social and Behavioral Research and the Internet: Advances in Applied Methods and Research Strategies* (eds. M. Das, P. Ester and L. Kaczmirek), 304–332. London, New York: Routledge.
- Pan, Y., Landreth, A., Park, H. et al. (2010). Cognitive interviewing in non-English languages: a cross-cultural perspective. In: *Survey Methods in Multinational, Multiregional, and Multicultural Contexts* (eds. J.A. Harkness, M. Braun, B. Edwards, et al.), 91–113. Hoboken, NJ: Wiley Blackwell.
- Prüfer, P. and Rexroth, M. (2005). Kognitive Interviews. *ZUMA How-to-Reihe*, 15. https://www.gesis.org/fileadmin/upload/forschung/publikationen/gesis_reihen/howto/How_to15PP_MR.pdf.
- Reja, U., Lozar Manfreda, K., Hlebec, V., and Vehovar, V. (2003). Open-ended vs. close-ended questions in web questionnaires. *Developments in Applied Statistics*, Metodološki zvezki 19: 160–175.
- Schonlau, M. and Couper, M.P. (2016). Semi-automated categorization of open-ended questions. *Survey Research Methods* 10: 143–152. <https://doi.org/10.18148/srm/2016.v10i2.6213>.
- Schuman, H. (1966). The random probe: a technique for evaluating the validity of closed questions. *American Sociological Review* 31: 218–222.
- Smith, T.W. (2010). The globalization of survey research. In: *Survey Methods in Multinational, Multiregional, and Multicultural Contexts* (eds. J.A. Harkness, M. Braun, B. Edwards, et al.), 477–487. Hoboken, NJ: Wiley Blackwell.
- Smyth, J.D., Dillman, D.A., Christian, L.M., and McBride, M. (2009). Open-ended questions in web surveys: can increasing the size of answer boxes and providing extra verbal instructions improve response quality? *Public Opinion Quarterly* 73: 325–337. <https://doi.org/10.1093/poq/nfp029>.
- Steenkamp, J. and Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research* 25: 78–107. <https://doi.org/10.1086/209528>.
- Thrasher, J.F., Quah, A.C., Dominick, G. et al. (2011). Using cognitive interviewing and behavioral coding to determine measurement equivalence across linguistic and cultural groups: an example from the International Tobacco Control Policy

- Evaluation Project. *Field Methods* 23: 439–460. <https://doi.org/10.1177/1525822X11418176>.
- Van de Vijver, F.J.R. (2013 November). Contributions of internationalization to psychology: toward a global and inclusive discipline. *American Psychologist*: 761–770. <https://doi.org/10.1037/a0033762>.
- Van de Vijver, F.J.R. and Chasiotis, A. (2010). Making methods meet: mixed designs in cross-cultural research. In: *Survey Methods in Multinational, Multiregional, and Multicultural Contexts* (eds. J.A. Harkness, M. Braun, B. Edwards, et al.), 455–473. Hoboken, NJ: Wiley.
- Van de Vijver, F.J.R. and Leung, K. (2011). Equivalence and bias: a review of concepts, models, and data analytic procedures. In: *Cross-Cultural Research Methods in Psychology* (eds. F.J.R. van de Vijver and D. Matsumoto), 17–45. Cambridge: Cambridge University Press.
- Vandenberg, R. and Lance, C. (2000). A review and synthesis of the measurement invariance literature: suggestions, practices, and recommendations for organizational research. *Organizational Research Methods* 3: 4–70. <https://doi.org/10.1177/109442810031002>.
- Willis, G.B. (2005). *Cognitive Interviewing: A Tool for Improving Questionnaire Design*. Thousand Oaks, CA: Sage.
- Willis, G.B. (2015). Research synthesis. The practice of cross-cultural cognitive interviewing. *Public Opinion Quarterly* 79: 359–395. <https://doi.org/10.1093/poq/nfu092>.
- Willis, G.B. and Zahnd, E. (2007). Questionnaire design from a cross-cultural perspective: an empirical investigation of Koreans and non-Koreans. *Journal of Health Care for the Poor and Underserved* 18 (6): 197–217. <https://doi.org/10.1353/hpu.2007.0118>.
- Ziegler, M., Kemper, C.J., and Lenzner, T. (2015). The issue of fuzzy concepts in test construction and possible remedies. *European Journal of Psychological Assessment* 31: 1–4. <https://doi.org/10.1027/1015-5759/a000255>.
- Zuell, C., Menold, N., and Körber, S. (2015). The influence of the answer box size on item nonresponse to open-ended questions in a web survey. *Social Science Computer Review* 33: 115–122. <https://doi.org/10.1177/0894439314528091>.
- Züll (Zuell), C. (2016). Open-ended questions. GESIS Survey Guidelines. Mannheim, Germany: GESIS – Leibniz Institute for the Social Sciences. doi: https://doi.org/10.15465/gesis-sg_en_002.