CHAPTER 7

Pretesting Methods in Cross-Cultural Research

Eva Aizpurua

Introduction

In recent years, substantial advances have been made in the field of multinational, multiregional, and multicultural research (commonly referred to as 3MC survey research; Johnson, Pennell, Stoop, & Dorer, 2018). This research magnifies challenges associated with monocultural studies and poses unique ones at both the organizational and methodological levels. Because cross-cultural surveys seek to make comparative estimates across populations, the data must be valid and reliable for each specific group, as well as comparable across them. Even when questionnaires are carefully translated and adapted, groups may systematically differ in the way they interpret certain questions or respond to them, posing a threat to the validity of the comparisons. In this context, pretesting becomes particularly beneficial to identify potential problems in survey questions and to assess comparability (Willis, 2015).

This chapter introduces the concept and importance of pretesting in cross-cultural survey research. The most common methods used to pretest 3MC surveys are described, highlighting recent applications and developments, as well as current challenges. These methods include cross-cultural cognitive interviewing, online probing, vignettes, and behavior coding. Next, reference is made to the combination of multiple pretesting methods to assess and improve cross-cultural surveys. In the last section of this chapter, the main challenges and opportunities of pretesting in comparative contexts are discussed.

Pretesting Methods and Their Role in Cross-Cultural Research

Pretesting refers to a variety of methods designed to assess the adequacy of survey instruments and field procedures (Caspar et al., 2016). The potential of these methods to identify the existence and sources of problems makes

pretesting an indispensable phase of the survey life cycle. In the context of cross-cultural research, pretesting offers valuable information about the role that language and culture play in the question response process, pointing to noncomparability bias, for example, by identifying questions or response options that are interpreted differently across cultural groups, leading to systematic measurement errors that may be attributed to translation issues, cultural connotations, or both. By detecting questions that function differently when translated or administered to different groups, and by providing information about sources of bias, pretesting allows for corrections prior to data collection.

Pretesting methods are often used once the survey materials have been developed and adapted. In these instances, testing all versions of the survey with the target populations is a crucial step to promote equivalence (Goerman & Caspar, 2010). As an iterative process, pretesting involves multiple rounds, in which changes to the instruments are followed by subsequent rounds of testing. Although less frequently observed, pretesting can be used at an earlier stage to inform the design of the questionnaire (e.g., by identifying terms and concepts used by the population of interest). Pretesting can also be used after data collection to facilitate the interpretation of the data. For example, pretesting methods may help interpret unexpected quantitative findings from one or more groups. In the context of repeated cross-sectional and longitudinal surveys, pretesting also informs future design decisions (e.g., modification of survey questions) (Fitzgerald & Zavala-Rojas, 2020).

To promote data quality, several methods have been developed for pretesting and improving questionnaires. These methods have traditionally been used in single-population studies and are gaining popularity in the context of cross-cultural research due to their potential to reduce measurement and comparison errors that restrict the quality of 3MC surveys. Nevertheless, there is a lack of consensus regarding the amount, type, and combination of pretesting that should be conducted (see the forthcoming report of the American Association for Public Opinion Research [AAPOR]/ World Association for Public Opinion Research [WAPOR] Task Force on Comparative Survey Quality). Further, the design and implementation of pretesting in cross-cultural research poses challenges in addition to those encountered in single-population studies. These challenges are the result of an increased number of parties involved, often located in different regions and speaking a variety of languages (Miller, 2018). Recruiting participants from multiple cultural and linguistic groups, designing protocols that are

culturally appropriate and comparable, and adopting consistent methods to report results are some of the aspects resulting in increased logistical complexity of pretesting in 3MC surveys (Sha & Pan, 2013).

Several considerations guide the selection of pretesting methods, including the objectives of this process, the characteristics of the population, and the availability of resources. In the context of 3MC survey research, cultural appropriateness should also be taken into consideration because differences in communication styles and cultural norms may require adaptation of the protocols or implementation of different methods. In the next section, the most frequently used pretesting methods in cross-cultural studies will be discussed, emphasizing recent applications and challenges.

Pretesting Methods: Current Developments and Challenges

Cross-Cultural Cognitive Interviewing

Cross-cultural cognitive interviewing (CCCI) has become the most widely used method for pretesting and evaluating questionnaires in 3MC survey research. Cognitive interviewing refers to a range of techniques that provide information about the way in which respondents process and answer survey questions (Willis & Miller, 2011). To this end, two main strategies are used, alone or in conjunction: thinking aloud and verbal probes. Thinking aloud encourages participants to verbalize their thoughts as they answer survey questions. In contrast, probing requires interviewers to ask follow-up questions to obtain additional information about the response process. These probes can be designed in advance or be spontaneous and nonscripted, triggered by participants' behaviors. Probes administered immediately after tested survey questions are called concurrent probes, whereas probes administered at the end of the survey are referred to as retrospective probes.

Different types of probes serve different purposes (see Table 7-1), and their effectiveness may vary by cultural groups. For example, Martin et al. (2017) found paraphrasing, thinking aloud, and hypothetical probes to be difficult for women in Ethiopia and Kenya with low education levels. Other researchers have identified difficulties with paraphrasing, meaning-oriented probes, and thinking aloud tasks when used with non-English-speaking groups in the United States, regardless of their education levels (e.g., Goerman, 2006; Pan, 2004, 2008). Other multilingual studies have reported significant differences in the effectiveness of various types of probes in eliciting the desired information across linguistic groups, which may reflect

Table 7-1. Frequently used probes

Probe Type	Purpose	Example
Meaning oriented	Assesses respondent interpretation of terms, phrases, or questions	"What does the term 'property' mean to you here?"
Process oriented	Examines the process by which respondents select their answers	"How did you choose that answer?"
Paraphrase	Assesses respondent interpretation of questions	"What is this question asking in your own words?"
Elaborative	Gathers further information about the response process	"Could you explain your answer a little further?"
Hypothetical	Analyzes responses to hypothetical situations	"Please, report babies as age 0 when the child is less than 1 year old. If a person has a 4-month-old baby girl, what age should the respondent write here?"
Evaluative	Investigates the appropriateness of questions and response options	"Was it difficult for you to answer some of these questions here? Which ones?""Does the question here sound natural to you in <language>?"

Note: Examples taken from Park, Sha, and Willis (2016) and Park, Sha, and Pan (2013).

cultural norms and communication styles. The results from a multilingual cognitive project involving five languages indicated that evaluative and hypothetical probes were more effective for English, Russian, and Spanish respondents when compared with Chinese and Korean participants (Pan, Landreth, Park, Hinsdale-Schouse, & Schoua-Glusberg, 2010). Another study reported different outcomes for three types of probes used to assess the sensitivity of a series of translated questions in the Saudi context (Mneimneh et al., 2018). The findings show that proactive indirect probes asking whether "others" would find it uncomfortable to answer the questions resulted in more survey questions being identified as sensitive than direct probes asking about the respondents themselves and general probes asking respondents to elaborate on the questions in general. Further research is needed to better understand how different probes perform across cultural and linguistic groups and to understand the effects of education and culture in probe suitability.

In addition to the probes, the protocols for the interviews require adaptation to ensure that they comply with linguistic conventions and

communication styles. Researchers have encountered difficulties in applying standard protocols developed from the perspective of English speakers to respondents from other cultural and linguistic groups that are less familiar with the interview task (Martin et al., 2017). Park, Goerman, and Sha (2017) compared the performance of different types of practice sessions to help Asian language speakers become more familiar with the cognitive interview process. They found that an action-based enhanced practice worked better than the traditional one translated from English. Interviewers indicated that participants in the enhanced practice felt more comfortable and better understood the purpose of the interview when compared with those presented with the traditional practice. Similarly, in an experimental project testing the American Community Survey (ACS) with Spanish speakers, protocols including additional rapport building and less structured interviews performed better than conventional protocols translated from English (Park & Goerman, 2018). However, more research is needed comparing different approaches to cognitive interview outcomes across languages and cultures.

The selection of participants and interviewers poses unique challenges in CCCI. Given the need to understand what the sources of error are, it is essential for interviewers to be fluent in the language of the pretest, as well as sensitive to cultural and linguistic nuances (Caspar et al., 2016). Although some flexibility in the conduct of the interviews has been advised, a common strategy to compensate for less skilled interviewers in applied settings has been the development of highly structured interviews (Lee, 2014; Miller et al., 2011). Despite the lack of guidelines regarding appropriate sample sizes in cognitive interviews generally (Blair & Conrad, 2011), it has been recommended that the number of interviewees be greater than that normally used in standard cognitive interviewing (Willis, 2015). The rationale behind this recommendation is to increase the likelihood of identifying problems that may arise or be more prevalent only among certain groups (Fitzgerald, Widdop, Gray, & Collins, 2011). Based on 132 interviews conducted in four countries (Bolivia, Fiji, New Zealand, and the United States), Hagaman and Wutich (2017) indicated that sample sizes of 12–16 may be sufficient for studies with homogeneous populations. However, they found that larger sample sizes are required to reach data saturation in heterogeneous and culturally diverse populations. As the literature suggests, several factors should be weighted when determining sample sizes, including participant characteristics, interviewer skills and experience, available economic

resources, pretesting design (e.g., whether CCCI is going to be used alone or in combination with other methods), and anticipated problems (Blair & Conrad, 2011; Lee, 2014).

When testing translated questionnaires, participants may be restricted to monolingual non-English speakers or may include bilingual speakers. Although it was traditionally assumed that only monolingual speakers should be interviewed, recent studies suggest the value of evaluating translated questionnaires with both groups. Results from cognitive interviews of the Chinese and Korean translations of the ACS Language Assistance Guide indicated that the issues reported by monolingual and partially bilingual speakers were similar. When differences were found, they seemed to be driven by demographic differences (age, education, years living in the country) and not as much by language proficiency (Park et al., 2016). Results from cognitive interviews of the 2020 Decennial Census questionnaire with monolingual and bilingual Spanish speakers ratify the added value of including both groups. While bilingual participants identified most of the problems reported by monolinguals, there were a number of issues that were problematic for only one group. For example, the concept of "live or stay somewhere else" was only misunderstood by monolinguals, while the concept of "housemate or roommate" was more frequently misunderstood by bilinguals (Goerman, Meyers, Sha, Park, & Schoua-Glusberg, 2019).

CCCI has been mostly used to assess the cross-cultural equivalence of survey questions and to detect problems associated with translations. For example, a study conducted with participants in the Netherlands and Spain uncovered construct differences in the interpretation of "quality of life" (Benítez, Padilla, van de Vijver, & Cuevas, 2018). Although this term was mainly associated with relationships among Spaniards, it was more generic and linked to happiness for the Dutch. Similarly, findings from another CCCI project in six countries pointed to differences in the interpretation of the scope of "friends and acquaintances." In five of the six countries (Australia, Malaysia, Mexico, United States, and Uruguay), the term encompassed family members, but in Thailand it connoted only non-kin (Thrasher et al., 2011). These examples indicate that equivalent translations do not guarantee functional equivalence because connotations associated with context depend on social, cultural, and linguistic elements. CCCI has also shed light on systematic differences in the interpretation and use of response options. The study conducted by Benítez et al. (2018) showed that, when compared with

the Dutch, Spanish respondents were more influenced by question order effects and showed less consistency across responses.

Despite the wide use of CCCI, there has been a lack of standards for analyzing and reporting on interview data (Ridolfo & Schoua-Glusberg, 2011). Drawing on sociolinguistic approaches, Pan and Fond (2014) developed a coding scheme to classify translation issues leading to measurement error in multilingual surveys. They identified five sources of errors: (1) linguistic rules (e.g., unnatural syntax), (2) cultural norms (e.g., address and naming conventions), (3) social practices (e.g., concepts that do not exist in a target language), (4) production errors (e.g., typographical errors), and (5) respondent errors (e.g., selecting multiple answers for questions when only one response should be selected). Other coding schemes have been developed in recent years, including the Cross-National Error Source Typology (CNEST), which emerged as part of the European Social Survey questionnaire design process (Fitzgerald et al., 2011). The Cross-National Error Source Typology defines three types of errors arising from different sources: source question problems, translation problems, and cultural portability. Source question problems arise when a questionnaire is designed in one language and then translated to another (or others). In these instances, problematic issues in the source questionnaire are likely to be replicated in the translated instruments (e.g., overly complex syntax, use of jargon). Translation problems refer to errors stemming from the translation process, ranging from typographical errors to using terms that are not equivalent in meaning, resulting in a loss of equivalence. Cultural portability problems occur when the concept of interest does not exist in all groups or when it manifests itself in different ways. For example, Pan and Fond (2014) reported difficulties with translations of certain concepts that appeared to be uniquely American, including "mobile homes" and "nursing homes," which were uncommon in the translated languages (Chinese, Korean, Russian, and Vietnamese).

In addition to preexisting tools for the analysis of qualitative data, Q-Notes, a specific software product for data entry and the structured analysis of cognitive interviews, has been developed by the US National Center for Health Statistics. Given its ability to centralize the process of data entry and its analytical flexibility, this software has been used in cross-cultural studies of various scales (Benítez & Padilla, 2014; Miller, 2018; Ridolfo & Schoua-Glusberg, 2011). Among its benefits for CCCI, Q-Notes can be used to analyze entire data sets, as well as examine the performance of

questions across cultural or linguistic groups (Miller, 2018). In terms of reporting, Boeije and Willis (2013) proposed the Cognitive Interviewing Reporting Framework (CIRF) to guide the presentation of findings from this pretesting method in a comprehensive and systematic way. CIRF is a 10-category checklist that includes the following sections, allowing for flexibility in their ordering: (1) research objectives; (2) research design; (3) ethics; (4) participant selection; (5) data collection; (6) data analysis; (7) findings; (8) conclusions, implications, and discussion; (9) strengths and limitations of the study; and (10) report format. CIRF has been used to report cognitive interviewing studies in various countries, as well as mixed-method studies combining cognitive interviews with quantitative methods (Boeije & Willis, 2013; Padilla, Benítez, & Castillo, 2013).

Although CCCI has been mainly used to assess responses to survey questions, it has proven useful in testing multilingual advance materials, such as brochures and advance letters (Chan & Pan, 2011; Pan et al., 2010), and to refine scales measuring latent constructs (Reeve et al., 2011). Research conducted to date provides evidence of the utility of CCCI to identify issues and understand the sources of bias across cultural and linguistic groups (Benítez et al., 2018; Park et al., 2013). However, previous research also emphasizes the need to culturally adapt the protocols because interviewing techniques may not work equally well in all cultural groups.

Online Probing

In recent years, several studies have assessed the potential of online probing to uncover problems with survey questions and identify interpretation differences across countries (see Behr, Meitinger, Braun, & Kaczmirek, 2020, for a review of cross-cultural online probing). In online probing, after answering a survey question, respondents receive one or more probes to explore different aspects of the cognitive process they went through to answer the question. Among the probing techniques, mostly comprehension (e.g., "What does this term mean to you?") and category selection probes (e.g., "Please, explain why you selected this answer.") have been used to explore the country-specific interpretation of questions and assess item comparability (Behr et al., 2014). Despite using probes similar to in-person cognitive interviewing, several aspects vary between the two methods, including the mode, the appropriate sample size, and the level of interactivity (Meitinger & Behr, 2016). Unlike CCCI, online probing does not include interviewers, which removes potential interviewer effects but rigidifies the interview

process. Responses provided by participants cannot be followed up through subsequent probes if the desired information has not been gathered. However, online probing allows for increased standardization and cost-effective recruitment of participants, particularly when they are dispersed across geographical areas (Neuert & Lenzner, 2019).

Meitinger, Braun, Bandilla, Kaczmirek, and Behr (2014) tested a composite scale measuring national pride across five countries in Europe (Germany, Great Britain, and Spain) and North America (United States and Mexico). Their results indicated that online probing was effective in identifying systematic variations across countries. For example, the question about pride in the Social Security system was interpreted differently in the United States and Spain. While respondents in the United States tended to equate the Social Security system with retirement benefits, in Spain most respondents associated "Seguridad Social" with the health care system. Another study exploring the cross-national comparability of a "civil disobedience" item across six countries (Canada, Denmark, Germany, Hungary, Spain, and the United States) pointed to substantial interpretation differences. In particular, respondents in Canada and the United States associated civil disobedience with violence and destruction more often than those in any of the other countries, leading to a lack of cross-national equivalence (Behr et al., 2014).

As part of the same project, Meitinger and Behr (2016) compared the findings from cognitive interviewing and online probing in Germany. They found that online probing resulted in higher nonresponse rates and shorter responses to the probes. Although participants in the standard cognitive interviews uncovered slightly more potential problems, the overlap between the two methods was high. Further research comparing cognitive interviewing and online probing in cross-cultural settings is needed to better understand their performance.

Previous studies suggest that when multiple probes follow a survey question, the sequence in which they are presented may affect the quality of the responses and the motivation of the participants, although the effects seem to vary across countries (Meitinger, Braun, & Behr, 2018). Given the scarcity of studies and the increased popularity of online probing, further research is needed comparing the performance of different combinations of online probes in a wider set of cultural contexts. In addition, more research is needed examining the impact of design features (e.g., probe placement, text box size) and number of probes on the responses to them. Given that most studies have used this pretesting technique with online panelists, who tend to

be experienced survey respondents, future research would benefit from applying online probing to general population samples, furnishing the current evidence with greater validity (Neuert & Lenzner, 2019).

Vignettes

Vignettes are hypothetical situations that can be used to assess survey questions. When applied, participants are provided with one or more scenarios, in textual or visual form, and asked to answer a series of questions regarding the interpretation of terms and the process followed to answer the questions. This method has been often used in the context of cognitive interviews and focus groups; it offers several advantages including the ability to test multiple situations without the challenge of recruiting participants who would correspond to each specific situation. For example, multiple scenarios have been used to assess different categories of the relationship question used on the Census form (e.g., "housemate or roommate," "roomer or boarder," "stepson or stepdaughter," "unmarried partner"), because recruiting participants from each group would become very costly (Sha, 2016). In addition, vignettes can be particularly useful to test sensitive questions, because they shift the focus from participants to hypothetical cases (Goerman & Clifton, 2011). Vignettes have proven to be effective in examining comprehension issues with Spanish and Asian language translations (Goerman & Clifton, 2011; Sha, 2016).

Despite their potential, vignettes have several drawbacks, including that participants' responses to scenarios may differ from their own responses in real-life situations. In the context of cross-cultural research, particular attention should be paid to the cultural appropriateness of the vignettes, as scenarios developed for and tested with a group may not be appropriate in other contexts. For example, Sha (2016) reported some discomfort among Vietnamese participants presented with a scenario describing a couple living together without being married. Similarly, Goerman and Clifton (2011) found that a vignette depicting two women renting a room to an unrelated man was culturally inappropriate for some Spanish speakers.

Vignettes have often been used in combination with other pretesting methods, particularly cognitive interviews. A recent study comparing the performance of vignettes in focus groups and cognitive interviews in seven languages concluded that administering the vignettes in cognitive interviews was more effective for identifying problems with survey questions, particularly for Arabic and Spanish speakers (Meyers, García Trejo, & Lykke, 2017). Because

studies comparing the performance of vignettes across pretesting methods are scarce, more research is needed in this area. In terms of vignette design, although some studies have used textual information only (Sha, 2016), others have combined vignettes with pictures or drawings (Goerman & Clifton, 2011). Considering the cognitive burden posed by vignettes, this latter approach could be particularly useful with participants whose education levels are low.

Behavior Coding

Behavior coding is a method by which behaviors displayed by interviewers and respondents during the question response process are systematically observed, coded, and analyzed (Johnson, Holbrook, et al., 2018). Originally developed to assess interviewer performance, behavior coding is increasingly used to evaluate survey questions and examine difficulties for both respondents and interviewers. The assumption on which this method relies is that deviations from the optimal survey process can help identify problematic questions. These deviations can be reflected in respondents' behavior (e.g., requests for repetition or clarification of questions, answers that do not use the options offered with the questions) or in interviewers' behavior (e.g., not reading the questions exactly as written). Table 7-2 shows examples of codes used in previous research to identify survey problems.

Although behavior coding provides systematic information that can be used to improve survey questions, little is known about the comparability of behavior codes across cultural and linguistic groups. To fill this gap, studies have begun investigating cultural variability in respondents' and interviewers' behaviors during survey interviews. Comparing behavior coding across cultural groups interviewed in English, Holbrook et al. (2006) reported greater comprehension difficulties among the three minority groups participating in their study (African Americans, Mexican Americans, and Puerto Ricans) when compared with non-Hispanic whites. They explained these differences indicating that "questions that are written from the perspective of the dominant cultural group seem to be difficult for members of minority cultural groups" (Holbrook et al., 2006, p. 587). Similarly, findings from a behavior coding study with African American, Latina, and non-Latina white women in the United States suggested cultural variability in comprehension and mapping difficulties. Specifically, Latinas expressed more comprehension difficulties than white respondents, and African Americans were more likely to report mapping difficulties compared to whites (Cho, Fuller, File, Holbrook, & Johnson, 2006).

Table 7-2. Examples of behavior codes

Respondent	
Clarification	Respondent indicates uncertainty about the meaning of a question
	Respondent indicates uncertainty about the time frame of the question
	Respondent indicates uncertainty about the meaning of the response options
	Respondent asks the interviewer to repeat part of or the entire question
Inadequate answer	Respondent provides an answer not using the response options offered with the question
Interviewer	
Incomplete reading	Interviewer does not read the question entirely, omitting parts of it
Poor reading	Interviewer does not read the question as written, by adding or changing one or more words

Note: Examples taken from Holbrook, Cho, and Johnson (2006) and Johnson, Holbrook, et al. (2018).

Differences across languages have also been found in previous research. Using behavioral coding, Pascale (2016) analyzed interviews conducted in English and Spanish to evaluate the ACS Content Test. Nonstandard interviewer behavior was more frequent when interviews were conducted in Spanish. Major changes to the questions, higher rates of skipping, and incorrectly verifying questions occurred more often in interviews conducted in Spanish than in English (54 percent versus 39 percent). More recently, Johnson, Holbrook, et al. (2018) conducted a study in which questions designed to produce difficulties were deliberately introduced (e.g., questions asking about nonexistent policies or objects, double-barreled questions, mismatches between the question stem and the response options). This study included respondents from different cultural backgrounds, who were interviewed in various languages (English, Korean, and Spanish). Their findings suggest that respondents across racial, ethnic, and linguistic groups generally reacted in a consistent way when confronted with questions designed to elicit problems. When compared with nonproblematic questions, they generated more problems, as expressed by behavioral codes. Although most groups reacted to the poorly designed questions in a similar manner, differences were found between Korean Americans and non-Hispanic whites interviewed in English. Specifically, Korean Americans reported fewer mapping difficulties when responding to the questions designed to elicit mapping problems than non-Hispanic whites. In addition to respondents'

behavior, differences were found in interviewers' behavior, with non-English-speaking interviewers misreading questions more often than English-speaking interviewers.

A similar experimental study conducted in Korea raised questions about the effectiveness of behavior coding in identifying problematic survey questions (Park & Lee, 2018). In this experiment, respondents were randomly assigned to an intentionally problematic questionnaire (e.g., omitting response options that were likely to be selected, unusually wide reference periods making recall difficult) or to a control featuring existing questions that have been extensively pretested and fielded. Behaviors indicative of potential problems were found to be very limited. Despite finding a higher number of problematic behaviors among respondents when the flawed questionnaire was used, the differences between the groups were not significant. Moreover, the number of problematic behaviors displayed by interviewers was not higher in the group receiving the flawed questionnaire, with codes suggesting the opposite pattern (a higher number of interviewers' problematic behaviors in the control group).

Another study has pointed to potential differences in the effectiveness of behavior coding across countries, which may be attributed to communication norms and styles. Thrasher et al. (2011) assessed the equivalence of survey questions across six countries (Australia, Malaysia, Mexico, Thailand, Uruguay, and the United States), finding that behavioral coding was more successful identifying problems in the two English-speaking, Western countries (Australia and the United States). In Western countries, where directness and openness are the preferred communication styles, behavior coding may be more effective than in other countries with a preference for indirect styles (Pan et al., 2010; Park & Lee, 2018). Although behavior coding is a promising tool to identify problematic questions in 3MC surveys, further research is needed examining the comparability of behavior codes across cultural and linguistic groups. Because behavior coding is based on overt behaviors, important requirements for comparability include ensuring that members of various groups are equally likely to express problems during survey interviews and that the codes capture cultural variations of these behaviors.

Combining Pretesting Methods

Combining pretesting methods and triangulating their findings provides additional information that helps to make informed decisions. Despite this,

few studies have used multiple methods to assess noncomparability bias across linguistic and cultural groups. Thrasher et al. (2011) combined behavioral coding and cognitive interviewing to identify issues in survey questions for adult smokers across six countries. Their findings suggest that both methods yield similar conclusions, although more potential errors were identified using cognitive interviews. Childs and Goerman (2010) highlighted the benefits of using a mixed-method approach to pretest the US Census Test Nonresponse Followup (NRFU) in Spanish and English. Whereas findings from cognitive interviews were very similar between the languages, behavior coding pointed to significantly more problems with the Spanish instrument. For example, questions in English were administered correctly (i.e., asking questions as worded and correctly verifying information) more often than those in Spanish.

In addition, some studies have combined quantitative and qualitative methods to assess the cross-cultural comparability of constructs. For example, the European Social Survey (Fitzgerald & Zavala-Rojas, 2020) and the European Health and Social Integration Survey (Wilmot, 2020) exemplify two large-scale projects in which a variety of pretesting methods have been used. On a smaller scale, Meitinger (2017) applied multigroup, confirmatory factor analysis and online probing in a mixed methods approach to examine the cross-national equivalence of patriotism and nationalism in five countries (Germany, Great Britain, Mexico, Spain, and the United States). Her findings suggest that online probing can help clarify quantitative results and better understand the reasons for the lack of cross-national equivalence. Similarly, Reeve et al. (2011) combined cognitive interviewing with psychometric methods to evaluate the performance of a scale measuring discrimination in a multiethnic population comprising African Americans, Asian Americans, and Latinos in the United States. Their findings reinforce the notion that qualitative and quantitative techniques complement each other by identifying distinct problems and providing different types of information on the same issues. However, the different focuses of qualitative and quantitative methods may result in situations in which these approaches lead to contradictory solutions. In this study, cognitive interviews suggested that a relatively short, 12-month reference period functioned best, while quantitative findings revealed that few individuals reported experiencing discrimination frequently, which called for a longer recall period to capture both usual and rare acts of discrimination. In these instances, the approach to be taken will depend on the goals of the study and the specific use of the scale.

Because different pretesting methods elicit different problems and may not work equally well across cultural groups, combining them maximizes their benefits, providing information to improve survey instruments in different ways. Of particular note are studies combining qualitative and quantitative techniques because they offer the value of the generalization afforded by quantitative methods with the in-depth information provided by qualitative techniques. Given the singularities of the different groups involved in cross-cultural research, combinations of pretesting methods may also vary across the groups (Caspar et al., 2016). In addition to the specific methods, the sequence in which these methods are used may have major consequences on the results, such that it requires careful consideration.

Concluding Remarks

Recent years have witnessed an increase in the number and scope of cross-cultural surveys. This trend has been accompanied by theoretical developments and innovations in all stages of the survey cycle, including pretesting methods and applications. These methods were originally developed for single-population studies and require adaptation to be used across a range of languages, regions, and cultures. Despite the increased use of pretesting methods in 3MC surveys, there remains no consensus regarding best practices for their design and implementation.

In this chapter, the current state of pretesting in cross-cultural surveys has been reviewed, focusing on recent applications and current challenges. Most of the studies investigating differences across linguistic and cultural groups have used a limited number of pretesting methods, primarily cognitive interviewing. Despite this, best practices for CCCI are underdeveloped, and more empirical evidence is needed to better understand the performance of different interviewing approaches and probe types across groups (Boeije & Willis, 2013; Lee, 2014). This field of study would also benefit from additional research examining appropriate sample sizes and numbers of iteration rounds in cross-cultural research with groups featuring various levels of homogeneity.

In contrast to CCCI, very little is known about the performance of other pretesting methods in the context of cross-cultural research. Of particular note is the scarcity of studies utilizing widely used pretesting methods in single-population studies, such as focus groups, expert reviews, and usability testing. Some exceptions include recent applications of focus groups (Sha, Hsieh, & Goerman, 2018) and expert reviews (Goerman, Meyers, & García

Trejo, 2019) to assess and refine questionnaires and other survey materials in multilingual projects. In addition, a few studies have assessed the usability of translated questionnaires and survey materials with non-English or limited English speakers (Leeman, Fond, & Ashenfelter, 2012; Sha et al., 2018; Wang, Sha, & Yuan, 2017), successfully identifying navigation problems. For example, a usability test of the online version of the Puerto Rico Community Survey found that respondents experienced difficulties entering their names into the single box provided. These difficulties were attributed to differences in naming conventions between the United States, with one family name, and Puerto Rico, where two last names (paternal and maternal) are common, requiring additional boxes to enter the information. The evaluation of these and other pretesting methods across different cultures and linguistic groups is an important area for future research. In addition to expanding the use and combination of pretesting methods, much can be learned by sharing the outcomes of tested questions in cross-cultural projects using repositories that researchers and organizations can consult (e.g., Q-Bank, developed by the US National Center for Health Statistics, SQP software; Saris & Gallhofer, 2014).

Acknowledgments

This chapter stems from the "Prisons: The Rule of Law, Accountability and Rights" project, with funding from the European Research Council under the European Union's Horizon 2020 research and innovation program (grant agreement 679362). The author would like to thank Sophie van der Valk for her insightful comments on a previous draft of this paper and Mandy Sha for her support and encouragement.

References

Behr, D., Braun, M., Kaczmirek, L., & Bandilla, W. (2014). Item comparability in cross-national surveys: Results from asking probing questions in cross-national web surveys about attitudes towards civil disobedience. *Quality & Quantity, 48*, 127–148. https://doi.org/10.1007/s11135-012-9754-8

Behr, D., Meitinger, K., Braun, M., & Kaczmirek, L. (2020). Cross-national web probing: An overview of its methodology and its use in cross-national studies. In P. C. Beatty, D. Collins, L. Kate, J. L. Padilla, G. B. Willis, & A. Wilmot (Eds.), *Advances in questionnaire design, development, evaluation and testing* (pp. 521–544). Hoboken, NJ: Wiley & Sons.

Benítez, I., & Padilla, J. (2014). Analysis of nonequivalent assessments across different linguistic groups using a mixed methods approach. *Journal of Mixed Methods Research, 8*(1), 52–68. https://doi.org/10.1177/1558689813488245

Benítez, I., Padilla, J. L., van de Vijver, F., & Cuevas, A. (2018). What cognitive interviews tell us about bias in cross-cultural research: An illustration using quality-of-life terms. *Field Methods, 30*(4), 277–294. https://doi.org/10.1177/1525822X18783961

Blair, J., & Conrad, F. G. (2011). Sample size for cognitive interview pretesting. *Public Opinion Quarterly, 75*(4), 636–658. https://doi.org/10.1093/poq/nfr035

Boeije, H., & Willis, G. (2013). The Cognitive Interviewing Reporting Framework (CIRF): Towards the harmonization of cognitive testing reports. *Methodology, 9*, 87–95. https://doi.org/10.1027/1614-2241/a000075

Caspar, R., Peytcheva, E., Yang, T., Lee, S., Liu, M., & Hu, M. (2016). Pretesting. *Cross-cultural survey guidelines*. Ann Arbor, MI: Survey Research Center, Institute for Social Research, University of Michigan. Retrieved from https://ccsg.isr.umich.edu/index.php/chapters/pretesting-chapter

Chan, A. Y., & Pan, Y. (2011). The use of cognitive interviewing to explore the effectiveness of advance supplemental materials among five language groups. *Field Methods, 23*(4), 342–361. https://doi.org/10.1177/1525822x11414836

Childs, J., & Goerman, P. (2010). Bilingual questionnaire evaluation and development through mixed pretesting methods: The case of the U.S. Census Nonresponse Followup instrument. *Journal of Official Statistics, 26*(3), 535–557.

Cho, Y. I., Fuller, A., File, T., Holbrook, A. L., & Johnson, T. P. (2006). *Culture and survey question answering: A behavior coding approach.* American Statistical Association 2006 Proceedings of the Section on Survey Research Methods. Washington, DC: American Statistical Association.

Fitzgerald, R., Widdop, S., Gray, M., & 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*(4), 569–599.

Fitzgerald, R., & Zavala-Rojas, D. (2020). A model for cross-national questionnaire design and pretesting. In P. C. Beatty, D. Collins, L. Kate, J. L. Padilla, G. B. Willis, & A. Wilmot (Eds.), *Advances in questionnaire design, development, evaluation and testing* (pp. 493–520). Hoboken, NJ: Wiley & Sons.

Goerman, P. L. (2006). *Adapting cognitive interview techniques for use in pretesting Spanish language survey instruments*. Washington, DC: US Census Bureau, Statistical Research Division.

Goerman, P. L., & Caspar, R. A. (2010). A preferred approach for the cognitive testing of translated materials: Testing the source version as a basis for comparison. *International Journal of Social Research Methodology, 13*(4), 303–316. https://doi.org/10.1080/13645570903251516

Goerman, P. L., & Clifton, M. (2011). The use of vignettes in cross-cultural cognitive testing of survey instruments. *Field Methods, 23*(4), 362–378. https://doi.org/10.1177/1525822X11416188

Goerman, P., Meyers, M., & García Trejo, Y. (2019). *The place of expert review in translation and questionnaire evaluation for hard-to-count populations in national surveys*. (Survey Methodology Working Paper Number 2019-02). Washington, DC: Research and Methodology Directorate, Center for Behavioral Science Methods Research Series, US Census Bureau.

Goerman, P. L., Meyers, M., Sha, M., Park, H., & Schoua-Glusberg, A. (2018). Working toward comparable meaning of different language versions of survey instruments: Do monolingual and bilingual cognitive testing respondents help to uncover the same issues? In T. P. Johnson, B. E. Pennell, I. A. L. Stoop, & B. Dorer (Eds.), *Advances in comparative survey methods: Multinational, multiregional, and multicultural contexts (3MC)* (pp. 251–269). Hoboken, NJ: Wiley & Sons.

Hagaman, A. K., & Wutich, A. (2017). How many interviews are enough to identify metathemes in multisited and cross-cultural research? Another perspective on Guest, Bunce, and Johnson's (2006) landmark study. *Field Methods, 29*(1), 23–41. https://doi.org/10.1177/1525822X16640447

Holbrook, A., Cho, Y. I., & Johnson, T. (2006). The impact of question and respondent characteristics on comprehension and mapping difficulties. *Public Opinion Quarterly, 70*(4), 565–595. https://doi.org/10.1093/poq/nfl027

Johnson, T. P., Holbrook, A., Cho, Y. I., Shavitt, S., Chavez, N., & Weiner, S. (2018). Examining the comparability of behavior coding across cultures. In T. P. Johnson, B.-E. Pennell, I. A. L. Stoop, & B. Dorer (Eds.), *Advances in comparative survey methods: Multinational, multiregional, and multicultural contexts (3MC)* (pp. 271–292). Hoboken, NJ: John Wiley & Sons.

Johnson, T. P., Pennell, B.-E., Stoop, I. A. L., & Dorer, B. (Eds.) (2018). *Advances in comparative survey methods: Multinational, multiregional, and multicultural contexts (3MC)*. Hoboken, NJ: John Wiley & Sons.

Lee, J. (2014). Conducting cognitive interviews in cross-national settings. *Assessment, 21*(2), 227–240. https://doi.org/10.1177/1073191112436671

Leeman, J., Fond, M., & Ashenfelter, K. T. (2012). *Cognitive and usability pretesting of the online version of the Puerto Rico Community Survey in Spanish and English*. (SSM2012-09). Washington, DC: US Census Bureau.

Martin, S. L., Birhanu, Z., Omotayo, M. O., Kebede, Y., Pelto, G. H., Stoltzfus, R. J., & Dickin, K. L. (2017). "I can't answer what you're asking me. Let me go, please.": Cognitive interviewing to assess social support measures in Ethiopia and Kenya. *Field Methods, 29*(4), 317–332. https://doi.org/10.1177/1525822X17703393

Meitinger, K. (2017). Necessary but insufficient: Why measurement invariance tests need online probing as a complementary tool. *Public Opinion Quarterly, 81*(2), 447–472. https://doi.org/10.1093/poq/nfx009

Meitinger, K., & Behr, D. (2016). Comparing cognitive interviewing and online probing: Do they find similar results? *Field Methods, 28*(4), 363–380. https://doi.org/10.1177/1525822x15625866

Meitinger, K., Braun, M., Bandilla, W., Kaczmirek, L., & Behr, D. (2014, July). A*spects of measuring national pride: Insights from online probing*. Paper presented at the XVIII ISA World Congress. Yokohama, Japan.

Meitinger, K., Braun, M., & Behr, D. (2018). Sequence matters in web 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

Meyers, M., García Trejo, Y. A., & Lykke, L. (2017). The performance of vignettes in focus groups and cognitive interviews in a cross-cultural context. *Survey Practice, 10*(3), 1–11.

Miller, K. (2018). Conducting cognitive interviewing studies to examine survey question comparability. In T. P. Johnson, B.-E. Pennell, I. A. L. Stoop, & B. Dorer (Eds.), *Advances in comparative survey methods: Multinational, multiregional, and multicultural contexts (3MC)* (pp. 203–225). Hoboken, NJ: John Wiley & Sons.

Miller, K., Mont, D., Maitland, A., Altman, B., & Madans, J. (2011). Results of a cross-national structured cognitive interviewing protocol to test measures of disability. *Quality & Quantity*, *45*(4), 801–815. https://doi.org/10.1007/s11135-010-9370-4

Mneimneh, Z., Cibelli Hibben, K., Bilal, L., Hyder, S., Shahab, M., Binmuammar, A., & Altwaijri, Y. (2018). Probing for sensitivity in translated survey questions: Differences in respondent feedback across cognitive probe types. *Translation & Interpreting, 10*, 73–88.

Neuert, C., & Lenzner, T. (2019, August 13). Effects of the number of open-ended probing questions on response quality in cognitive online pretests. *Social Science Computer Review*, 1–13. https://doi.org/10.1177/0894439319866397

Padilla, J. L., Benítez, I., & Castillo, M. (2013). Obtaining validity evidence by cognitive interviewing to interpret psychometric results. *Methodology, 9*, 113–122. https://doi.org/10.1027/1614-2241/a000073

Pan, Y. (2004, May). *Cognitive interviews in languages other than English: Methodological and research issues.* Paper presented at the American Association for Public Opinion Research, Phoenix, AZ.

Pan, Y. (2008). Cross-cultural communication norms and survey interviews. In H. Sun & D. Kádár (Eds.), *It's the dragon's turn. Chinese institutional discourses (Linguistic Insights)* (1st ed., pp. 17–76). Bern, Switzerland: Peter Lang.

Pan, Y., & Fond, M. (2014). Evaluating multilingual questionnaires: A sociolinguistic perspective. *Survey Research Methods, 8*(3), 181–194. https://doi.org/10.18148/srm/2014.v8i3.5483

Pan, Y., Landreth, A., Park, H., Hinsdale-Schouse, M., & Schoua-Glusberg, A. (2010). Cognitive interviewing in non-English languages: A cross-cultural perspective. In J. A. Harkness, M. Braun, B. Edwards, T. P. Johnson, L. Lyberg, P. Ph. Mohler, B.-E. Pennell, & T. W. Smith (Eds.), *Survey methods in multinational, multiregional, and multicultural contexts* (pp. 91–113). Hoboken, NJ: Wiley.

Park, H., & Goerman, P. L. (2018). Setting up the cognitive interview task for non-English-speaking participants in the United States. In T. P. Johnson, B.-E. Pennell, I. A. L. Stoop, & B. Dorer (Eds.), *Advances in comparative survey methods: Multinational, multiregional, and multicultural contexts (3MC)* (pp. 227–249). Hoboken, NJ: John Wiley & Sons.

Park, H., Goerman, P., & Sha, M. (2017). Exploring the effects of pre-interview practice in Asian language cognitive interviews. *Survey Practice, 10*, 1–11. https://doi.org/10.29115/sp-2017-0019

Park, H., & Lee, J. (2018). Exploring the validity of behavior coding. *Field Methods, 30*(3), 225–240. https://doi.org/10.1177/1525822x18781881

Park, H., Sha, M. M., & Pan, Y. (2013). Investigating validity and effectiveness of cognitive interviewing as a pretesting method for non-English questionnaires: Findings from Korean cognitive interviews. *International Journal of Social Research Methodology*, *17*(6), 643–658. https://doi.org/10.1080/13645579.2013.823002

Park, H., Sha, M. M., & Willis, G. (2016). Influence of English-language proficiency on the cognitive processing of survey questions. *Field Methods*, *28*(4), 415–430. https://doi.org/10.1177/1525822X16630262

Pascale, J. (2016). Behavior coding using computer assisted audio recording: Findings from a pilot test. *Survey Practice, 9*, 1–11. https://doi.org/10.29115/sp-2016-0012

Reeve, B. B., Willis, G., Shariff-Marco, S. N., Breen, N., Williams, D. R., Gee, G. C., … Levin, K. Y. (2011). Comparing cognitive interviewing and psychometric methods to evaluate a racial/ethnic discrimination scale. *Field Methods*, *23*(4), 397–419. https://doi.org/10.1177/1525822X11416564

Ridolfo, H., & Schoua-Glusberg, A. (2011). Analyzing cognitive interview data using the constant comparative method of analysis to understand cross-cultural patterns in survey data. *Field Methods*, *23*(4), 420–438. https://doi.org/10.1177/1525822X11414835

Saris, W. E., & Gallhofer, I. (2014). *Design, evaluation, and analysis of questionnaires for survey research* (2nd ed.). Hoboken, NJ: Wiley & Sons.

Sha, M. (2016). The use of vignettes in evaluating Asian language questionnaire items. *Survey Practice, 9*, 1–8. https://doi.org/10.29115/sp-2016-0013

Sha, M., Hsieh, Y. P., & Goerman, P. (2018). Translation and visual cues: Towards creating a road map for limited English speakers to access translated Internet surveys in the United States. *The International Journal for Translation & Interpreting Research, 10*(2), 142–158.

Sha, M., & Pan, Y. (2013). Adapting and improving methods to manage cognitive pretesting of multilingual survey instruments. *Survey Practice, 6*(4), 1–8. https://doi.org/10.29115/SP-2013-0024

Sha, M., Son, J., Pan, Y., Park, H., Schoua-Glusberg, A., Tasfaye, C., … Clark, A. (2018). *Multilingual research for interviewer doorstep messages, final report*. (Survey Methodology RSM2018-08). Washington, DC: Research and Methodology Directorate, Center for Behavioral Science Methods Research Series, US Census Bureau.

Thrasher, J. F., Quah, A. C., Dominick, G., Borland, R., Driezen, P., Awang, R., … Boado, M. (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*(4), 439–460. https://doi.org/10.1177/1525822X11418176

Wang, L., Sha, M., & Yuan, M. (2017). Cultural fitness in the usability of U.S. Census internet survey in the Chinese language. *Survey Practice, 10*(3), 1-8. https://doi.org/10.29115/SP-2017-0018

Willis, G. B. (2015). The practice of cross-cultural cognitive interviewing. *Public Opinion Quarterly*, *79*(S1), 359–395. https://doi.org/10.1093/poq/nfu092

Willis, G., & Miller, K. (2011). Cross-cultural cognitive interviewing: Seeking comparability and enhancing understanding. *Field Methods, 23*(4), 331–341. https://doi.org/10.1177/1525822X11416092

Wilmot, A. (2020). Measuring disability equality in Europe: Design and development of the European Health and Social Integration Survey Questionnaire. In P. C. Beatty, D. Collins, L. Kate, J. L. Padilla, G. B. Willis, & A. Wilmot (Eds.), *Advances in questionnaire design, development, evaluation and testing* (pp. 545–570). Hoboken, NJ: Wiley & Sons.