The development and testing of instruments for cross-cultural and multi-cultural surveys

Blair, Johnny; Piccinino, Linda

THE DEVELOPMENT AND TESTING OF INSTRUMENTS FOR CROSS-CULTURAL AND MULTI-CULTURAL SURVEYS

JOHNNY BLAIR & LINDA PICCININO

1 Introduction

There are several motivations for proposing wider and more systematic use of the instrument development and testing methods described in this chapter: the wide range of research contexts in which cross-cultural instrument development occurs suggests a methodical consideration of the design implications of these different contexts for instrument development would be useful. Measurement error research has shown that a number of factors can play an important role in response error and data quality; and the analysis of exactly how those factors function in data collection has grown more sophisticated (Biemer et al., 1991; Presser et al., 2004). The research on measurement error due to the survey instrument has developed primarily through non-cross-cultural studies and experiences, but there is no reason to believe its findings do not have important consequences for cross-cultural surveys as well. Finally, the range of pretesting techniques has grown, as has research on the strengths and weaknesses of those different techniques.

The potential importance of the research context becomes apparent when one considers the range of situations where cross-cultural or multi-cultural surveys are done. A simple listing of some of these research contexts makes it evident that the need to consider cross-cultural factors when designing a survey instrument can occur for a number of reasons. Researchers sometimes design instruments "from scratch" for use in cultures not their own. An instrument already administered in one cultural setting may need to provide comparable measurements in a different culture. Or an instrument may be designed to use across different cultures, either in a single survey or in multiple independent surveys. The crossing of cultural boundaries may or may not involve traversing national or language boundaries as well. Different language groups and quite distinct cultures are, of course, encountered within a single country, as well as internationally.

Until fairly recently, the literature on cross-cultural survey instrument development largely focused on the issue of translation. But there are many other factors that affect measurement error in addition to the accuracy of the translation. Conversational implicature affects people's interpretation of things they hear beyond the literal meaning of the words in a statement or question (Grice, 1989). The impact of question order effects may vary from one culture to another (Schwarz, 2003). Respondents' understanding of the general intent of a survey, planned uses of the data, and assurances of confidentiality are also factors that affect how they understand and respond to survey questions. Finally, some aspects of cognition that affect response behaviors and response effects can differ across cultures (Johnson et al., 1997). While there have been important developments in cross-cultural survey design (see for example Grosh & Glewwe, 2000), much of the literature and technical reports of particular surveys' methodology suggest that frequently they do not take full advantage of the available instrument development and testing techniques. Many problems encountered in designing instruments for cross-cultural and multi-cultural research certainly could be addressed using one or more of the pretesting techniques that are now in common use.

The literature on cross-cultural survey design and the technical reports of such surveys' methodologies suggest that many of the surveys do not take full advantage of the available instrument development and testing techniques. Many problems encountered in designing instruments for cross-cultural and multi-cultural research certainly could be addressed using one or more of the pretesting techniques that are now in common use.

It may well be that these techniques themselves will sometimes need to be adapted to accommodate cultural considerations, which is still another area that would benefit from careful reports of experiences on particular surveys as well as from methodological research. It is important to keep in mind that in developing cross-cultural instruments, the researcher confronts all the usual issues of writing clear questions that capture the construct of interest and that present respondents with tasks that are reasonable; but that added to these is a range of cultural and communicative issues.

Cultural issues also exist for researchers designing surveys for their own cultures and countries. However, many factors, particularly societal norms for behaviors and interactions, might be taken into account almost unconsciously. These types of issues, however, can also be unconsciously overlooked when designing a survey for another culture, transferring a survey between cultures, or designing a single survey meant to be adaptable to multiple cultures. All these diverse factors support the potential value of applying a systematic approach to defining the instrument development issues and potential problems, selecting the appropriate instrument testing methodologies and carefully implementing them.

2 Response Effects

Concern about response effects has a long history in survey research, but seems to have played much less of a role in cross-cultural survey methodology than in research and surveys that do not address cross-cultural issues. Response effects play a crucial role in survey measurement error. Much of the response effects research followed the finding of Sudman and Bradburn's meta-analysis showing that beyond the text of survey questions, the nature of the response tasks can strongly influence respondent answers (Sudman & Bradburn, 1974). In reporting the results of a large series of experiments and re-analyses Schuman & Presser (1981) provided extensive evidence of effects resulting from alternative question wording.

Up until now, little was known about the underlying pragmatic and psychological mechanisms that underlie those effects. The Cognitive Aspects of Survey Methodology (CASM) movement began to change that. A number of works have shed light on how cognitive and communicative processes influence survey response to produce some of the response effects that have been observed (Schwarz & Sudman, 1996; Schwarz, 1996; Tourangeau, Rips & Rasinski, 2000). It would seem a logical step to extend the cognitive testing and analytic approach to cross-cultural studies and certainly some important work in that direction has been done. There is strong evidence that respondents from different cultures can vary in their response behaviors in reaction to the same survey question. In an important paper, Johnson et al. (1997) make a convincing case for the potential of cultural differences to impact each stage of the response process: comprehension, recall, response formation and reporting. Such effects result, in part, from the fact that "Cultural groups are also known to vary along the dimensions of individualism versus collectivism, emotional control versus emotional expressiveness, masculinity versus femininity, and the acceptability of physical contact" (Johnson et al., 1997: 89).

Memory retrieval and judgment formation can likewise be affected by cultural differences. In particular, semantic memory – in which information storage is linked to conceptual categories – may be structured differently across cultures. Whether such differences would affect survey recall tasks is an open question. In forming some types of judgments, respondents will sometimes rely on a frame of reference (for example, what constitutes a reasonable expectation, say, for health care) or an anchor point (such as what are the norms or what constitutes "average" behavior in their culture for, say, time spent on recreation) and decide on their answer in relation to these heuristic devices. There is certainly a possibility that these frames of reference and anchor points may differ across cultures. If so, response scales may not be used in the same way, an important factor if, for example, results in multiple countries are to be compared. Finally, editing of responses may be affected by respondents' understanding of the reporting task – how much informa-

affected by respondents' understanding of the reporting task – how much information is wanted in response to an open response question or, as Johnson et al. (1997) point out, by such factors as self-presentation and some aspects of the respondent's interaction with the interviewer.

More recently, Johnson & van de Vijver (2003) reviewed what we know about culturally-impacted effects in the area of social desirability. This may be seen as a special, but illustrative, case of response editing. Their discussion focuses on whether social desirability might have a differential effect on respondents from different cultural backgrounds, which may be a function of real differences in response propensities between cultures. But on another view, social desirability may be due to characteristics of the survey question, the interaction between the interviewer and respondent, or both. In particular, they point out that "cultural differences between respondents and interviewers sometimes produces varying patterns of responses." Additionally, they note that survey administration mode can affect reporting, particularly on sensitive questions.

In summary, for many factors affecting response behaviors there is some theoretical basis for expecting cultural differences, but for most, only a small amount of survey-based research has been done. But even the little we know at this point suggests caution in assuming how well survey instruments will "travel" across cultures; and to use multiple 'tools' to assess how an instrument will perform.

Although this research has not focused on the implications of these findings for pretesting instruments that have been created or transformed to work in different cultures, clearly it is a significant issue. As far as can be determined from the literature much cross-cultural instrument development and testing has been concerned with translation. Beyond issues of the quality of the original question's translation, as has been noted, comprehension can be different or difficult (not the same issues) due to cultural factors. Moreover, it is often unclear how much pretesting was done on an instrument before using it in the researcher's *home culture*, let alone how its performance might be affected by transplanting it into another cultural context. It would seem reasonable, considering the research noted above and even an informal observation of common practice in cross-cultural surveys, that those surveys might benefit from using a wider array of pretesting methods. How that might best be done and how those methods themselves need to be adapted to the cross-cultural research context is a subject for future methodological research and practical utilization.

3 Review of Some Issues in Instrument Testing

3.1 Comparability of scales

Another important component of the pretesting and testing process is to uncover societal similarities and differences in the measurement of frequencies of unobservable and observable behaviors. One way to accomplish this is to design and test a set of response scales that are thought to be comparable from one cultural context to another. Depending on the type of response scale used, conclusions drawn about differences in behaviors and their frequency of occurrence can be false or misleading. It is only when the items in questionnaires representing multiple languages and contexts have equivalent response scales that comparable measurement and valid inferences can be obtained (Smith & Wolter, 2004).[1]

Ji, Schwarz & Nisbett (2000) used a set of open-ended responses and a set of frequency scales to explore cultural differences in behavior and memory. They found that the when the frequency scales were used, these had differential effects on reports of frequency of observable behaviors and unobservable behaviors in China and the United States. Ji et al. (2000) also reported that when the open-ended response format was used, it yielded results for observable behaviors that were about equal for China and the United States (as was intended, by design). They concluded that, depending on the response format used, researchers could arrive at different conclusions about cultural differences in behavior. This study illustrated the potential risks associated with the use of frequency scales in cross-cultural research and underscored the need for pretesting of response scales.

3.2 Equivalence

The question of equivalent measures is often an issue in cross-cultural surveys. Such equivalence becomes a central concern when a single survey encompasses more than one cultural setting, or when the same survey is conducted for the purpose of comparing different cultures or countries. There are essential multiple dimensions of equivalence of measures: equivalent comprehension of questions, including response scales, and equivalent respondent use of response scales (as discussed above).

These dimensions seem central, though there is by no means agreement on this point.

Johnson (1998) provides a comprehensive review of alternative concepts of equivalence, listing fifty-two types of equivalence identified in a literature review; and goes on to

1 In their paper, Smith & Wolter (2004) offer some preliminary ideas about what kinds of scales might lead to more equivalent cross-cultural comparisons.

consider their impact on the focus of survey design. Many of these types differ only slightly from one another; and the types also range from high-level, conceptual notions such as *complete equivalence* and *cultural equivalence* to what appear to be very literal notions of equivalence such as *text equivalence*. Among these is *measurement equivalence*, sometimes defined as "...instance in which factor loadings and error variances are identical across groups" and *semantic equivalence* which occurs when "...survey items ...exhibit identical meaning across two or more cultures after translation." These last two categories, the former of which would seem to encompass the question's response categories, seem to suggest a reasonable direction toward an operational definition of equivalence. Additionally, this direction would seem also to lend itself to developing practical procedures for attaining equivalence. We make some suggestions for beginning to specify such procedures.

The first step is the identification of points of non-equivalence. Expert review, focus groups and cognitive interviews can all be potentially useful in such identification. But, unless the degree of non-equivalence is large, the amount of effort and sample sizes necessary to accomplish this identification may be considerably greater than that need for typical pretesting to determine comprehension and other response task performance. A sequential process approach may be useful in this stage. For example, the in-country expert or an expert panel may be able to identify areas of possible non-equivalence. The principal investigator or data analyst would then judge the possible implications of non-equivalence, considering the variables affected, the rough magnitude of the problem and the likely proportion of the sample that may be impacted. On the basis of this assessment, a decision would need to be made about the importance of removing or adjusting for the non-equivalence. Just as with other sources of measurement error, it is seldom possible to correct every potential flaw. This process of identification and assessment will become more reliable over time as experience is gained with different populations and survey topics and measures.

It may be possible to revise the questions or other features of the instrument to achieve equivalence. Failing that, it may be possible to determine the relationship between non-equivalence measures and allow for it in data analysis. For example, if in a survey that includes young Hispanic males it is determined both that they overstate, say, their health status in comparison to young men from other ethnic groups, and some estimate of the extent of the overstatement, a statistical adjustment may be possible. Although the process we suggest may be costly and relatively complex, these factors have to be weighed against the potential impact of non-equivalence on key analyses.

3.3 Validity

The importance of construct validity is to ensure that the items in the instrument designed for another cultural context capture what you are attempting to measure, and are valid representations of what you are trying to measure, within that country or culture.

Miller et al. (1981) in their early paper asked what comprised a reliable index of a concept and its meaning for a specific country, and assessed the parts of the index that were valid across countries (in this case, the United States and Poland). They began their analysis by examining the *within-country validity* of the items in the US survey that were to be compared to the Polish replication survey. The analysis looked for internal consistency of the items supposed to measure the key concept (i.e. authoritarian-conservatism): empirical differentiation of the items in the key concept from other related concepts (i.e. personality), any covariation due to measurement error, and qualities of the index that showed it was an adequate representation of the concept.

They next examined the *cross-national validity* of the items in the index of authoritarian-conservatism by looking at the correlation of the within-country index of one country against the index derived from factor analysis of the other country's data. They looked at the statistical properties of indicators that were common to both countries and of those that were country-specific.

The investigators concluded that there also was a set of core items that had an equivalent meaning in both countries, but that there were country-specific items, and that those items should remain specific only to that respective country's index.

Some researchers (Miller et al., 1981) recognized the utility of certain statistical tools (e.g. factor analysis) for assessing construct validity in cross-cultural survey instruments. Others such as Saris, van der Veld & Gallhofer (2004) pointed to researchers like Campbell & Fiske (1959) who maintained that validity could best be evaluated if more than one method was used to measure the same trait; in this way errors could be detected. The methods they refer to could be as simple as using multiple versions of a response scale and comparing their correlations. Validity could then estimated from the strength of the relationship between the trait of interest and the "true score" (defined as the component of the observed variable that represents the trait and method used).

4 Pretest Techniques

The available pretesting techniques are well known. The potential strengths and shortcomings of these techniques have been noted in a number of sources based on experimental studies (Presser & Blair, 1994; Presser et al., 2004) or books that provided practical advice for implementation (Fowler, 1995; Czaja & Blair, 2005), proposed process quality frameworks for pretesting (Blair & Piccinino, 2004), or considered pretesting techniques from a theoretical perspective (Martin, 2001). All of these perspectives are potentially important for the application of the techniques in cross-cultural survey instrument pretesting. Below we list the main techniques and note selected theoretical and practical issues that may be relevant to cross-cultural surveys, as well as some suggestions about how techniques might be adapted to address some of the issues we have raised.

Presser & Blair (1994) compared four pretesting methods using a single questionnaire in repeated trials of each method. The methods tested were: conventional pretests, behavior coding, cognitive interviews, and expert panels. A model-based coding scheme was used that classified problems as respondent-semantic, respondent-task, interviewer-task, or analysis. On average, expert panels were found to be the most productive measured by the total number of problems identified. Expert panels and behavior coding were more reliable than other methods in the number of problems identified across trials, as well as in their distribution of problems.

The findings that are probably most important in their implications for cross-cultural studies concern respondent-semantic and respondent-task problems. Cognitive interviews were consistently better at identifying respondent-semantic problems; while conventional pretests and expert panels were best at identifying respondent-task difficulties.

Below we provide an overview of the main pretest techniques and note additional theoretical and practical issues that may be relevant to cross-cultural surveys, as well as some suggestions about how techniques might be adapted to address some of the issues we have raised.

4.1 Expert review

Expert reviews are a generic term for a number of different activities, involving different kinds of experts whose advice is elicited in different ways. Most relevant here are experts about the particular country or culture where the survey will be conducted; and experts in the language, dialect or patois in which the survey will be conducted. What makes a person an expert? On some level, just being indigenous or fluent in a language may qualify as sufficient expertise. It also is important to be aware of possible within-culture class

differences. Certainly different experts may come to different conclusions and provide conflicting advice or solutions to question problems. Even conflicts of this type may be useful in identifying issues that need additional attention in design or testing.

The problem of selecting experts may be true for any survey, but is particularly so in other cultures, where the principal investigator or survey methodologist might not be competent to identify "experts." One approach for helping with this problem that can have broader value as well is the use of expert panels. In this case, usually three or more experts are brought together for a discussion of pertinent survey issues. The experts need not all have the same specialty to be informative about a particular issue. One expert may have a good overall understanding of, for example, how the health care system functions, but be unfamiliar with access difficulties that certain subpopulations in the country experience. While another expert may understand how people living in cities normally handle certain financial matters, but know little about how the same functions work in poor, rural areas.

For example, in a Willingness to Pay survey in Kenya (McGunnigle et al., 2000), an examination of the data showed that respondents and interviewers had problems with survey questions about bid price (price willing to pay), and often entered numeric amounts to yes/no questions. Also, service fee categories in the instrument often did not match actual services being offered. Some discussion with the staff in Kenya after-the-fact was necessary to understand how the service and fees systems worked, and to realize problems with the data. These problems might have been minimized if discussions with Kenya experts occurred before the final development stage of the questionnaire.

When recruiting an expert panel, it is useful to both cover the range of areas considered important for the survey at hand, but also to have some overlaps in expertise so that alternative judgments or points of view can be identified and assessed.

4.2 Cognitive interviewing

The standard four-stage response model (e.g. Tourangeau & Rasinski, 1988) posits a sequential set of cognitive tasks respondents must perform in answering survey questions. This descriptive model has provided a useful general framework for instrument testing as well as other aspects of data collection.

Cognitive interviewing is a generic term for a set of available techniques that can be used in various combinations in one-on-one pretest interviews (Conrad & Blair, 2004). In large part, cognitive interviews include respondents thinking aloud, reporting everything that comes to mind as they answer the questions. Thinking aloud is combined with two types of probes: concurrent, asked during the interview; and retrospective, asked after the inter-

view (or after self-contained interview sections). The probes may be partly written in advance of the cognitive interview as well as improvised during the interview. Cognitive interviewers probe based either on respondents' indications of difficulty (e.g. requests for clarification, changing answers, inability to produce an answer) or on conjectures about aspects of questions that may cause response problems. There is evidence that the former type of probe is more reliable and less likely to turn up false positives (i.e. "identify" nonexistent problems) than the latter type, though these context-free probes may produce more problems as well (Conrad & Blair, 2004).

There is not agreement, however, on the practical implications of these findings. Willis (2004) in "Cognitive Interviewing: A Tool for Improving Questionnaire Design," the most comprehensive treatment of the method to date, takes a more positive view of using what he calls *proactive* probing. Until some methodological research on the technique has been conducted in a cross-cultural setting, caution is advised in the selection of probes.

The main strength of cognitive interviewing is its potential ability to identify problems at any stage of the response process: comprehension, recall, response formation or response reporting. Cognitive interviews often uncover possible reasons for the occurrence of identified problems; such information can help guide question repair. In addition to basic response tasks problems, cognitive interviewing can probably help identify pragmatic communicative issues that may be especially important in cross-cultural surveys. But there is only a small amount of research to date to support this contention.

It is useful to keep in mind that the tasks required in some variants of cognitive interviewing may be very difficult for some respondents. These are not things that people commonly are requested to do; some detailed explanation or examples may be required. Some cognitive interview tasks, like thinking aloud or paraphrasing, require first that the respondents understand what they are being asked to do.

Even if respondents do understand the task, those who may be less articulate or less socially comfortable talking in front of a stranger may have difficulties with some forms of cognitive interviewing. If there are possible cultural barriers to the type of interaction necessary for the successful conduct of cognitive interviews, the in-country experts may be able to point this out early in the planning process. In such a situation, the researcher may choose a version of cognitive interviewing that takes account of such barriers or decide not to use cognitive interviewing at all. Of course, this requires that the researchers clearly explain the cognitive interview process to the in-country experts.

Careful thought should be given to these issues, since difficulty with the cognitive response task may sometimes be mistaken for difficulty with the survey response task.

Looked at from a different perspective, Groves et al. (1992), in discussing direct questioning of respondents as a method to identify problems with meaning, note that "...the questions on meaning would themselves be subject to large measurement errors."

However, if target population members can, on the whole, express themselves adequately, and are comfortable responding to probes and completing other necessary cognitive interview tasks, cognitive testing can be an efficient method of uncovering a range of problems and possible solutions as well.

In work associated with a project in the Philippines in 2003, both structured and self-administered questionnaires for various types of family planning/health care providers were pretested. Through cognitive testing it was discovered that providers (especially midwives) found the self-administered portion of the questionnaire to be irrelevant and therefore did not pay much attention to answering it (Commercial Market Strategies project, 2003).

4.3 Focus groups

Focus groups can take different forms for different purposes. A focus group can consist of experts (as described above) who provide insights into the target population's country, culture, and relevant aspects of their language, or comment on any aspect of the survey. Typically, however, focus groups are composed of target population members. Just as in U.S. focus groups, decisions need to be made about what group composition will best foster the open exchange necessary to produce useful information – whether about particular subject matter or reactions to actual draft survey questions. The project's in-country experts may judge whether or not focus group interactions can be expected to work as necessary, or if not, what sorts of adaptations may be possible.

For example, in a reproductive health study in Jamaica (Young, 2003), focus groups and role-play testing prior to final questionnaire development revealed differences in attitudes of pharmacists toward youth depending on the gender of the youngster and whether they were in or from an inner-city or suburban area. Mystery client scripts were modified to incorporate the separate pharmacy needs of girls and boys, as well as to accommodate the "uptown" and "downtown" language or slang used by these youth. Even though the original instrument developed in the United States was in English, the cultural and language differences in English usage were substantial enough to warrant testing.

In another study, an evaluation of a survey in Kenya (McGunnigle et al., 2000) suggested that if the project had a longer time frame, focus groups could have provided more systematic information on the fee structure of clinics than relying solely on information from client exit interviews.

Still another type of problem was encountered in a survey in Albania (Partners for Health Reform*plus*, 2004a), and to some extent in Kenya, where little variability was found in the five-category response options to questions about client satisfaction with health providers and services. Clients generally reported satisfaction with providers/services as good or excellent, but when asked informally, some admitted they were giving "polite" answers in the structured interviews and had actually experienced a lesser degree of satisfaction than reported. Focus groups and consultation with in-country experts about customs, politics and etiquette in the region might have revealed that it was customary not to voice negative opinions publicly about health care providers/staff. This might have been remedied by permitting the survey to be modified accordingly, perhaps using a different data collection mode, and using a more elaborate introduction to try to alleviate the tendency for the respondent to use non-negative responses.

4.4 Conventional pretesting

Conventional pretesting is so named because it is the most common form of pretesting and, absent other description, what one would assume if told only that a pretest had been conducted. The technique is based on a kind of emulation of the survey. A small sample of respondents from the target population is sampled and the survey is administered to them just as intended in the actual study. A structured debriefing is held afterward in which the interviewers give their overall and their question-by-question assessment of how the interview went, what problems respondents experienced and, possibly, what changes might improve the instrument.

While the conventional pretest may include a post-interview debriefing to supplement the interviewers' impressions, usually the interviewers simply serve as proxy reporters for problems respondents had. Behavior coding (described below) can also be incorporated into conventional pretesting. Both the post-interview respondent debriefing and behavior coding can serve to validate (or not) some of the interviewers' reports.

In cross-cultural surveys, particularly if an in-country contractor is used, it is important that the interviewers go through a pretest training that not only covers issues planned for the full project interviewer training, but also discusses (with examples) the kinds of information that they are expected to be able to report about in the debriefing. They should be encouraged to take notes either during or immediately after each interview to use in the debriefing.

In a recent project in Rwanda, two days of adequately planned interviewer pretest training was truncated to two hours due to unavoidable administrative and managerial resource cuts. As a result, interviewers performed poorly in the field and had to be subjected to retraining mid-way through the field process (Partners for Health Reform*plus* (2004b).

4.5 Respondent debriefings

As noted, conventional pretests can be supplemented in different ways to obtain more and richer information about how the instrument and specific items performed. Immediately after the interview, respondents can be asked by the interviewer what they thought particular questions meant, what items they thought were difficult and why, among other things. Of course, the list of potential problems developed by the pretest team should inform the debriefing interview questions. If cognitive interviewing precedes conventional pretesting, often issues identified in the cognitive testing can inform the construction of the respondent post-interview debriefing.

The debriefing questions may be in either an open- or closed-response format. But there is some evidence (Groves et al., 1992) that open- and closed-response debriefing items may produce different information and, more importantly, present different types of response issues for the respondents. They note that respondents who are more articulate may mention issues not noted by others. But this might give a false sense of the likelihood of such problems occurring and of their distribution across the population. Two lessons that might be taken from this result: first, a mix of types of questions may be better than relying on just one kind; second, one should bear in mind the essentially qualitative nature of pretesting (and the typically small samples), and not expect to learn too much about problem distributions that might occur if flaws were left uncorrected.

4.6 Behavior coding

Behavior coding is based on a conception of what the question-and-answer interview process should be like in the absence of question flaws (Fowler, 1995). Ideally, the interviewer will read the question verbatim, without error, and the respondent will select one of the offered response options. When the process fails, certain kinds of behaviors are likely to be seen, such as respondent interruptions and requests for re-reading or for clarification, interviewer reading errors and the like. These deviations from the ideal interviewer-respondent interaction are taken as indicators of question problems if these occur relatively frequently; with problems occurring in 15% of question administrations a commonly-used threshold.

This technique does raise some questions. Should we expect the same set of 'indicator' behaviors in cross-cultural surveys? Are there supplemental codes that can capture behaviors specific to testing in other cultures? Will people in other cultures indicate problems in the same way that respondents in the United States do? Will they volunteer comments, ask questions, ask for repeats etc., if given license to do so? This is one sort of thing to check with the in-country experts – will the assumptions of this or other testing methods work as

expected in the other cultures? Ultimately, these kinds of questions need to be addressed through a combination of careful methodological research and ongoing technical reports of actual survey experiences.

5 Overview of Process Approach to Pretesting

A process quality approach to questionnaire testing is a way of ensuring coverage of the range of potential types of problems and relevant issues that can occur in different 'realizations' or contexts of cross-cultural surveys. The idea is that after specifying the type of cross-cultural instrument/survey situation, one considers the particular set of potential problems that might arise. This set of possible problem areas guides the selection of pretest techniques to address these areas.

Blair & Piccinino (2004) offer an approach for systematizing the stages involved in cross-cultural instrument pretesting. One intention of the process approach they propose is to ensure a thorough coverage of the tasks, issues and potential problems associated with instrument development. These tasks, issues and problems can be, for example, testing parameters, cultural concerns, and specific defects in questions and supporting materials. This approach requires the collaborative effort of team experts to ensure that proper coverage occurs. Survey instruments often are flawed for reasons that, in retrospect, appear quite simple and apparent. The emphasis on potential problem coverage is recognition of this. It is important to be sure "all the rocks have been turned over" in the search for problems.

Obviously a translated instrument may perform differently in a new language when exactly equivalent words are not available; or, more seriously, when no comparable concept exists in the target culture. Less obvious is that an instrument might contribute to measurement error due to the tasks required of the respondent being at odds with either respondent capabilities or when survey questions are inadvertently in conflict with some cultural norms or expectations.

As an example, questionnaires used in a survey of family planning providers in the Philippines in 2003 (Commercial Market Strategies, 2003) were implemented in English and also in Tagalog. Although the translation into Tagalog was not thoroughly tested, pretesting of the questionnaire helped reveal that respondents found the dialect of Tagalog used in the translation to be too literal so that they sometimes missed the true meaning or interpretation of the terms used. The pretesting also indicated that a more conversational version of Tagalog was preferred by respondents.

Approaches to translation involving various team design and review procedures can address many of these issues (Harkness, 2003; Harkness et al., 2002; de la Puente, Pan & Rose, 2003). The focus of a process quality approach is broader than translation, and applies even when translation is (strictly speaking) not necessary. The approach should encompass the pragmatics of communication, and even extend to practical data collection implementation issues. The process approach recognizes that both technical design issues and operational issues – such as interviewer behaviors, or obtaining thorough reports of pretest results from an in-country contractor – could affect measurement error. The strength of a process approach is the use of a general framework that can be adapted to specific surveys. As the approach is used in different surveys it is likely that variations on that general framework will develop. The dissemination of the documentation of such variations will be essential to the continued development of this approach.

6 Summary

In the above discussion, we have indicated instrument development and testing issues for which alternative combinations of pretesting techniques may prove useful. In addition, we have suggested potential problems one may encounter and possible adaptations of the techniques to make them more suitable for cross-cultural instrument testing. It is important to note that these recommendations are based on judgment and experience, rather than on methodological research. Clearly, it will be necessary to conduct such research to learn how to use these methods to best effect. Even after such research results begin to become available, it still will be necessary to consider each survey's particular characteristics, subject matter, mode of administration and target population.

References

Biemer, P. P., R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, and S. Sudman. 1991. *Measurement Errors in Surveys.* New York: John Wiley & Sons.

Blair, J., and L. Piccinino. 2004. Paper presented at the RC33 Sixth International Conference on Social Science Methodology in Amsterdam, The Netherlands, August 16-20, 2004. (Also published in proceedings volume, ZUMA, May/June 2005.) *A Process Quality Approach to the Development and Testing of Cross-Cultural Survey Instruments.*

Campbell, D., and D. Fiske. 1959. "Convergent and Discriminant Validation by the Multitrait Multimethod Matrices." *Psychological Bulletin* 56:81-105.

Commercial Market Strategies (CMS) project. 2003. *Personal communication.*

Conrad, F., and J. Blair. 2004. "Data Quality in Cognitive Interviews: The Case of Verbal Reports." Pp. 67-87 in *Questionnaire Development, Evaluation and Testing Methods,*

edited by Presser, S., J. Rothgeb, M. Couper, J. Lessler, E. Martin, J. Martin, and E. Singer. New York: John Wiley & Sons.

Czaja, R., and J. Blair. 2005. *Designing Surveys: A Guide to Decisions and Procedures*. Thousand Oaks: Pine Forge Press.

Fowler, F. J. Jr. 1995. *Improving Survey Questions: Design and Evaluation.* Thousand Oaks: Sage Publications.

Grice, P. 1989. *Studies in the Way of Words.* Cambridge: Harvard University Press.

Grosh, M., and P. Glewwe (Eds.). 2000. *Designing Household Survey Questionnaires for Developing Countries. Lessons from 15 Years of the Living Standards Measurement Study.* Washington: The World Bank.

Groves, R. M., N. H. Fultz, and E. Martin. 1992. "Direct Questioning about Comprehension in a Survey Setting." Pp. 49-61 in *Questions about Questions: Inquiries into the Cognitive Bases of Surveys,* edited by Tanur, J. M. New York: Russell Sage Foundation.

Harkness, J. 2003. "Questionnaire Translation." Pp. 35-56 in *Cross-cultural survey methods*, edited by Harkness, J., F. J. R. van de Vijver, and P. P. Mohler. Hoboken: John Wiley & Sons.

Harkness, J., A. Schoua-Glusberg, and B. Pennell. 2002. *Questionnaire Translation and Questionnaire Design.* Proceedings of the International Conference on Questionnaire Development, Evaluation, and Testing Methods (QDET), Charleston, South Carolina, November 14-17.

Johnson, T. P. 1998. "Approaches to Equivalence in Cross-Cultural and Cross-National Survey Research." In *Cross-Cultural Survey Equivalence,* edited by Harkness, J. Mannheim, Germany: ZUMA News Special, Vol. 3.

Johnson, T. P., D. O'Rourke, N. Chavez, S. Sudman, R. Warneke, and L. Lacey. 1997. "Social Cognition and Responses to Survey Questions among Culturally Diverse Populations." Pp. 87-113 in *Survey Measurement and Process Quality,* edited by Lyberg, C., P. Biemer, M. Collin, C. Dippo, E. deLeeuw, N. Schwartz, and D. Trewin. New York: John Wiley & Sons.

Johnson, T. P. and F. J. R. van de Vijver. 2003. "Social Desirability in Cross-cultural Research." Pp. 195-204 in *Cross-Cultural Survey Methods*, edited by Harkness, J., F. J. R. van de Vijver, and P. P. Mohler. Hoboken: John Wiley & Sons.

Ji, L., N. Schwarz, and R. E. Nisbett. 2000. "Culture, Autobiographical Memory, and Behavioral Frequency Reports: Measurement Issues in Cross-Cultural Studies." *Personality and Social Psychology Bulletin* (May 2000).

Martin, E. 2001. *Theoretical Perspectives on the Question and Answer Process: Implications for Pretesting.* Proceedings Quest 2001, U.S. Census Bureau, Washington, DC, October 24-25, 2001, pp. 6-19.

McGunnigle, M., L. Piccinino, and T. J. Ryan. 2000. *Kenya Survey on Willingness to Pay and Client Satisfaction.* Bethesda, MD: Pathfinder/Kenya, Abt Associates Inc.

Miller, J., K. M. Slomczynski, and R. J. Schoenberg. 1981. „Assessing Comparability of Measurement in Cross-National Research: Authoritarian-Conservatism in Different Sociocultural Settings." *Social Psychology Quarterly,* 44 (3):178-191.

Partners for Health Reform*plus*. 2004a. *Primary Health Care Reform in Albania: Baseline Survey of Basic Health Service Utilization, Expenditures and Quality.* Bethesda, MD: The Partners for Health Reform*plus* Project, Abt Associates Inc.

Partners for Health Reform*plus*. 2004b. *USAID/Government of Rwanda National Health Accounts project.* Personal communication.

Presser, S., M. Couper, J. Lessler, E. Martin, J. Martin, J. Rothgeb, and E. Singer (Eds). 2004. *Methods for Testing and Evaluating Survey Questionnaires.* Hoboken: John Wiley & Sons.

Presser, S., and J. Blair. 1994. "Survey Pretesting: Do Different Methods Produce Different Results?" Pp. 75-104 in *Sociological Methodology, Vol. 24,* edited by Marsden, P. Washington: American Sociological Association.

de la Puente, M., Y. Pan, and D. Rose. 2003. Paper presented to the Federal Committee on Statistical Methodology Research Conference, Arlington VA, November 17-19, 2003. *An Overview of Proposed Census Bureau Guidelines for the Translation of Data Collection Instruments and Supporting Materials.*

Saris, W. E., W. van der Veld, and I. Gallhofer. 2004. "Development and Improvement of Questionnaires Using Predictions of Reliability and Validity." Pp. 275-297 in *Methods for Testing and Evaluating Survey Questionnaires,* edited by Presser, S., M. Couper, J. Lessler, E. Martin, J. Martin, J. Rothgeb, and E. Singer. Hoboken: John Wiley & Sons.

Schuman, H., and S. Presser. 1981. *Questions and Answers in Attitude Surveys. Experiments on Question Form, Wording and Context.* Orlando: Academic Press.

Schwarz, N. 1996. *Cognition and Communication: Judgmental Biases, Research Methods and the Logic of Conversation.* Mahwah : Lawrence Erlbaum & Associates.

Schwarz, N. 2003. "Culture-Sensitive Context Effects: A Challenge for Cross-Cultural Surveys." Pp. 93-100 in *Cross-Cultural Survey Methods,* edited by Harkness, J., F. J. R. van de Vijver, and P. Ph. Mohler. Hoboken: John Wiley & Sons.

Schwarz, N., and S. Sudman. 1996. *Answering Questions: Methodology for Determining Cognitive and Communicative Processes in Survey Research.* San Francisco: Jossey-Bass.

Smith, T. W., and K. M. Wolter. 2004. Paper presented to the American Statistical Association, Toronto, August 2004. *Techniques for Calibrating Response Scales across Countries and Languages.*

Sudman, S. and N. Bradburn. 1974. *Response Effects in Surveys: A Review and Synthesis.* Chicago: Aldine.

Tourangeau, R., and K. A. Rasinski. 1988. "Cognitive Processes underlying Context Effects in Attitude Measurement." *Psychological Bulletin* 103:299-314.

Tourangeau, R., L. Rips, and K. Rasinski. 2000. *The Psychology of Survey Response.* Cambridge: Cambridge University Press.

Willis, G. 2004. *Cognitive Interviewing: A Tool for Improving Questionnaire Design.* Newbury Park: Sage Publications.

Young, B. 2003. *Evaluating the Knowledge and Attitudes of Health Providers in the Provision of Emergency Contraception Pills and Condoms to Jamaican Adolescents.* Commercial Market Strategies (CMS) project. Kingston: JA Young Research Ltd.