Cognitive Interviewing versus Behavior Coding[1]

Martha Stapleton Kudela[2], Barbara H. Forsyth[2], Kerry Levin[2], Deirdre Lawrence[3], Gordon Willis[3]
Westat, 1650 Research Blvd., Rockville, MD, U.S.A., 20850[2]
National Cancer Institute, 6130 Executive Blvd., Rockville, MD, U.S.A., 20852[3]

Introduction

In single-language surveys, cognitive interviewing and behavior coding are pretest methods frequently used in tandem to identify potential problems with questionnaire items and fix them before beginning full scale survey data collection. In single language surveys, cognitive interview pretests are typically used relatively early in survey development, to determine whether respondents will have difficulty understanding the items as intended, whether they will have trouble accurately recalling the information they need and whether they will have problems selecting an appropriate response. In single language surveys, behavior coding is typically used later in survey development to verify that problems identified in earlier testing have been fixed and to ensure no large problems have been overlooked before full scale data collection begins.

The common use of cognitive interview and behavior coding pretests in tandem and in sequence reflects practical considerations, but also the complementary strengths of the two methods. At least in single language surveys, cognitive interview pretests are particularly effective for identifying comprehension problems (Willis et al., 1999). The detailed data they produce are useful for suggesting item revisions likely to reduce problems identified during pretesting (Forsyth et al., 2004). However, cognitive interview pretests typically include small numbers of interviews, and the interview setting often does not represent the usual survey interview context. Therefore, important problems may be undetected by cognitive interviewing. Behavior coding typically involves larger samples that are representative of the survey sample and the interviews themselves mimic standard survey conditions. These characteristics of behavior coding pretests enhance the likelihood that problems undetected in earlier cognitive interview pretests will be detected with later behavior coding. However, it can be difficult to use behavior coding results to identify revisions that are likely to reduce or eliminate problems because the results lack detail (Forsyth et al., 2004).

Relatively recently, survey researchers have begun investigations to identify effective methods for pretesting questionnaires administered in multiple languages. Researchers have documented the value of cognitive interviews, respondent debriefing interviews, and interviewer debriefing interviews as methods for identifying and fixing problems with questionnaires administered in multiple-languages (e.g., McKay et al., 1996; Harkness & Schoua-Glusberg, 1998; Harkness et al., 2003).

Multi-language surveys present a pretesting challenge that is not present in single-language surveys. In addition to cognitive issues related to comprehension, recall and response selection, multi-language surveys must wrestle with a linguistic issue – equivalency. Do questionnaire items administered in multiple languages measure the same constructs? As Forsyth et al. (2006) point out, effective translation and review processes are essential to assessing and ensuring equivalence. However, these processes may be incomplete because they rely on the assessments and judgments of multilingual language users. Pretesting is another important step for assessing and ensuring equivalence because pretesting provides an opportunity to elicit information from monolingual language users (Harkness et al., 2003).

The purpose of this paper is to extend the research of our earlier paper (Forsyth, et al., 2006) to explore the complementary strengths and weaknesses of cognitive interview and behavior coding pretest methods for surveys conducted in multiple languages. We will present additional results from pretesting the TUS/CPS survey in Mandarin and Cantonese Chinese, Korean and Vietnamese. Our goals are

[1] Paper presented to American Association for Public Opinion Research, Montreal, Canada, May, 2006.

- To compare results from cognitive interviews and behavior coding for pretesting questionnaires administered in four Asian languages; and
- To ask what we learn from each pretest method. Do the two methods make unique contributions to refining questions and translations? After we have conducted cognitive interview pretests, it is useful to add a behavior coding pretest step?

We will begin with a brief overview of the pretest methods and design. Then we will move onto reviewing pretest results.

Pretest Methods

The previous paper described the TUS and our cognitive interview pretest design. For that phase, we conducted cognitive interviews in three languages: Chinese (Cantonese and Mandarin), Korean and Vietnamese. For most of these languages, we conducted one round of nine cognitive interviews. For the Vietnamese translation, we conducted two rounds for a total of 14 cognitive interviews. In this paper, we will focus on the more trustworthy results from the second round of 5 cognitive interviews.

As the Forsyth et al. paper already described, we used the cognitive interview results to revise the Asian-language questionnaires in preparation for a larger-scale field pretest. As part of the field test, we conducted several additional questionnaire pretesting activities, including respondent debriefing, interviewer debriefing, behavior coding for selected survey items, and behavior coder debriefing.

For the field test, we used the Asian-language translations to conduct roughly 70 interviews each in Cantonese Chinese, Mandarin Chinese, Korean and Vietnamese. We identified respondents largely through vendor-supplied lists. With respondent permission, we tape recorded all interviews and used the tape recordings to behavior code the interviews. We behavior coded only a subset of questionnaire items. We selected 32 items to code based on problems identified during the cognitive interviewing task described above; problems identified during the interviewer debriefing conducted after the field test; and unusual response distributions.

Behavior coders were fluent in English and in one or more of the target Asian languages. We recruited the behavior coders through one of the study's Survey Language Consultants (SLCs).

Behavior coders coded only the first interviewer-respondent exchange for each item, and they assigned one interviewer and one or more respondent codes to each coded item. We used the set of seven codes in Table 1. The codes in Table 1 are based on similar coding schemes developed by Cannell and still widely used (e.g., Cannell et al., 1975). Five of the codes in Table 1 suggest a question problem (question not read; question read incorrectly; respondent interrupts; respondent requests clarification; problem with answer), and two of the codes suggest no problem (question read correctly; adequate answer). We identified questions with problems as any question where 20% or more of the interviews were coded for one or more of the problem codes. Because of skip patterns, some items were administered to relatively small numbers of respondents. We eliminated items administered to fewer than 20 respondents in a single language from the analyses for this paper.

Table 1. Behavior codes for interviewer and respondent behaviors.

Interviewer codes	
Trouble	Question not read
	Question read incorrectly
No trouble	Question read correctly
Respondent codes	
Trouble with question	Interrupts
	Requests clarification
Trouble with answer	Problem with the answer
No trouble	Adequate answer

After the behavior coding task, we conducted a behavior coder debriefing, separately for the Chinese-language coders (Cantonese and Mandarin), the Korean-language coders and the Vietnamese-language coders. We used the debriefing sessions to gather descriptive details about the behaviors that coders observed.

Before we go on to present results from the cognitive interview and behavior coding pretest activities, please notice an important feature of our general pretest design. We used the cognitive interview results to revise the Asian-language versions of the TUS questionnaire before we conducted the field test. Most of the items selected for behavior coding were revised between the cognitive interview pretest and the field test.

Cognitive Interviewing Results

We will begin by reviewing the main cognitive interview results. Gordon Willis presented preliminary analyses in two previous papers (Willis et al. 2005a; 2005b). This paper expands a bit on those earlier results.

One important finding held up across all three languages. In general, the cognitive interviews went smoothly. Respondents seemed to understand most of the survey items and they had little trouble answering questions about their smoking habits. In addition, cognitive interviewers had few problems administering the Asian-language questionnaire versions.

Nonetheless, cognitive interview pretests did identify some problems, and they seemed to fall into three problem categories: general cognitive problems, translation problems, and culture- or language-specific problems. General cognitive problems are caused by difficulties in comprehension, recall or response selection that are common across cultures and languages. Translation problems are caused when the selected translation wording alters the original question intent. Culture- or language-specific problems are caused when the original question intent is difficult to convey using the constructs from a specific language or culture. (In other words, it is easier to measure the construct of interest in some languages or cultures and more difficult in others.) We will use examples to illustrate these three types of problems.

Most of the problems we found through cognitive interviewing were either general cognitive problems or translation problems. We found few culture- or language-specific problems.

General Cognitive Problems. We will start with a few examples of the types of general cognitive problems identified through cognitive testing.

An item intended to measure nicotine dependence asked respondents to report true or false for the following statement: "Even in a bad rainstorm, if you ran out of cigarettes, you would probably go to the store to get some more." In all Asian-language versions of the questionnaire, respondents said "false" because they would never find themselves in this situation. Some said that they always have enough cigarettes on hand to avoid running out at inconvenient times. Others indicated that they would borrow cigarettes from friends. For these respondents, the item seemed to measure something other than nicotine dependence. The measurement issues for this item seem to transcend language or specific translation wording.

Another item gathering information about household smoking rules asked, "In a usual week, does ANYONE who lives here smoke cigarettes, cigars, or pipes anywhere inside this home?" Some respondents did not include themselves in their reports. In other words, "anyone" was interpreted more narrowly than intended. We revised the Asian-language questionnaires to ask, "In a usual week, does ANYONE who lives here, including yourself, smoke cigarettes, cigars, or pipes anywhere inside this home?"

A third item about shifting smoking habits asked, "Which is the MAIN reason you switched from a stronger to a lighter cigarette – as a way to try to quit smoking, or in order to smoke a less harmful cigarette?" Respondents who answered this item were unable or unwilling to select a single response. They wanted to report that both were important reasons for their shifting smoking habits. Again, difficulty choosing a main reason between two related reasons seems to transcend language or specific translation wording.

Translation Problems. Translation problems were roughly as common as general cognitive problems. Several of the translation problems identified involved mistaken omissions from or insertions into the intended questionnaire text. Other relatively common translation problems involved identifying formal language that would be easier to understand with less formal constructions.

Respondent confusion was a typical indicator for this type of problem. A few translation problems were more subtle. For example, in the Chinese-language questionnaire, the item asking: "Have you ever switched from a stronger cigarette to a lighter cigarette for at least 6 months?" Was incorrectly translated as "Since you have switched from regular to light cigarettes, has it been more than half a year?"

Here is second example from the Korean-language questionnaire. An item asking "How soon after you wake up do you typically smoke your first cigarette of the day?" was difficult to translate because the phrase for "how soon" is unusual in Korean. Based on cognitive interview results, we selected an alternative Korean translation that was likely to be more familiar to respondents. Roughly paraphrased, the revised translation asks, "After waking up on the mornings of days that you smoke, how long a period of time goes by before you smoke?"

In addition to finding specific translation problems based on participants' reactions, the cognitive interview testing provided an additional opportunity for translation reviewers to listen to and reconsider the translations. From this additional, informal review step, reviewers identified several refinements to the tested translation. These refinements were an unexpected bonus from our cognitive interview pretest activities.

Culture-Specific Problems. Cognitive interviewing identified very few problems that seemed culture-specific. Here are two examples.

An item on shifting smoking habits asked, "Have you EVER SWITCHED from a stronger cigarette to a lighter cigarette for at least 6 months?" Korean smokers reported that Korean cigarette packaging does not provide information about tar or nicotine levels. Several did not include shifts from Korean to American cigarettes in their reports, even when they switched to "ultra-light" versions of American cigarettes. Similar issues related to cigarette strength came up at different items in all of the Asian-language versions of the questionnaire. Thus, the difficulty may not be specific to a particular culture. Rather, the difficulty may be a factor for any respondent who has smoked unlabeled cigarettes. We identified this problem as "culture-specific" because it seems that the construct is more difficult to measure in some cultures than in others because of differences across societies. In this case, there are differences across societies in their cigarette packaging requirements.

As another example, an item on access to smoking asked, "In your opinion, how easy is it for minors to buy cigarettes and other tobacco products in your community?" Vietnamese respondents reported interpreting the term "community" to mean, "the Vietnamese community" rather than their physical neighborhood as intended. In addition, the Vietnamese term for "neighborhood" conveys a closeness among neighbors that is not part of the intended "community" construct. Again, we classified this problem as a culture-specific problem rather than a translation problem because the "community" construct seemed to differ in important ways across languages or cultures, making the measurement goal easier to achieve in some languages than in others.

Behavior Coding Results

Table 2 shows the number of items identified as problems for each language. The number of items identified as problems ranged from 13 to 22 of the 32 items coded. The numbers of problem items in Table 2 are relatively high in part because many of the items we chose to behavior code were already identified as potentially problematic based on the cognitive testing or from interviewer debriefing comments.

Table 2. Number of problem items, by language

Language	Number of problem items	Percentage of all coded items identified as problems
Cantonese Chinese	21	66%
Mandarin Chinese	22	69%
Korean	13	41%
Vietnamese	13	41%

It appears that behavior coding identified more problems in the Cantonese and Mandarin interviews (21 and 22 items respectively) than in the Korean or Vietnamese interviews. Differences in the translation quality between the Chinese versions and the Korean and Vietnamese versions could explain these differences. To explore this hypothesis, we used the coder debriefing results to discover the causes of the identified behavior coding problems. Table 3 shows the percentages of different problems mentioned during the coder debriefing session, separately for general cognitive problems, translation problems and culture- or language-specific problems.

Coder debriefing comments suggested that there were more translation problems with the Chinese language questionnaire versions than with the Korean or Vietnamese language questionnaire versions. However, for all four versions, general cognitive problems were considerably more prevalent than translation or culture-specific problems.

One similarity between the cognitive interview and behavior coding results is that both methods revealed more problems due to general cognitive issues than problems due to translation issues. Also both methods revealed few problems due to culture- or language-specific issues.

Table 3. Problem types, by language

	Type of Problem				
Language	General cognitive problems	Translation problems	Culture- or language-specific problems	Unclassifiable problems	Total number of problems*
Cantonese	52%	27%	15%	6%	33
Mandarin	60%	17%	17%	7%	30
Korean	53%	20%	13%	13%	15
Vietnamese	62%	12%	6%	19%	16

* In all four language versions, the number of problems identified is larger than the number of problem items identified because some items were identified as having multiple problems.

How similar are the specific problems identified by the two methodologies? To answer this question, we reviewed the item-by-item results from each method. This review indicated that relatively few of the problems identified through behavior coding were previously identified in the cognitive interview pretest. Most of the problems identified through behavior coding were problems that were not evident from the cognitive interview pretest results.

In part, this is because we were able to fix some of the problems identified through the cognitive interview pretests. However, our review suggests that some of the differences in problems identified reflect the different characteristics of the two pretest methods. We provide two examples that illustrate this point.

Example 1: Smoking habits
Which of these best describes the area where you work MOST of the time?

☐ Mainly work indoors
☐ Mainly work outdoors
☐ Travel to different buildings or sites
☐ Somewhere else

Results from cognitive interviews led us to conclude that respondents had no trouble understanding or answering the question. However, in the field test interviews, Cantonese, Mandarin and Korean-language interviewers regularly read only the first two response options, yielding relatively high rates on this item for the "question read incorrectly" code. Cognitive interviewing is not particularly strong for helping researchers to anticipate interviewer difficulties (Forsyth et al., 2004). This seems to be an example of this weakness of cognitive interviewing.

Example 2: Smoking habits
How soon after you wake up do you typically smoke your first cigarette of the day?

_____ minutes/hours

Results from cognitive interviews indicated that respondents had no trouble understanding or answering this item. Cognitive interview respondents described their morning routines and could tell us the point in their routines when they typically smoked their first cigarette. We interpreted this as a positive finding. Respondents understood the question and used predictable strategies to select a response.

During the field test interviews, many respondents referred to their morning routines to answer the question. For example, "after I brush my teeth" or "after breakfast." As a result, behavior coding yielded relatively large numbers of "problem with the answer" codes in all four Asian languages. (The behavior coding percentages ranged from 40% to 60% across the four languages.)

We also observed a few cases where problems identified based on cognitive interviews were re-identified by behavior coding.

Example 3: Smoking history
What is the total number of years you have smoked EVERYDAY? Do not include any time you stayed off cigarettes for 6 months or longer.

Cognitive testing indicated that respondents had trouble interpreting and using the exclusionary statement appropriately. We did not revise the item because it was important to keep the Asian-language versions parallel to the unchangeable English-language version of the questionnaire. During field pretest interviews, Cantonese and Korean respondents requested clarification relatively often (25% and 20% respectively). During debriefing, behavior coders reported that the exclusionary statement was a major cause for frequent requests for clarification.

Notably, the same item yielded relatively high behavior coding frequencies for "problem with answer" – a problem that was not anticipated by cognitive interviewing results. In all four Asian

languages, relatively large numbers of respondents reported their ages when they started smoking and their current ages, rather than a number of years. (Percentages ranged from 21% to 45% across the four languages.)

Summary and Conclusions

We will close by using the results we've reported to answer three questions.

First, what did we learn from early cognitive interviewing?
Early cognitive interviewing identified general cognitive issues that were relevant to questionnaire design across the four Asian languages. In addition, cognitive interviews identified translation issues that we overlooked during the earlier review and adjudication steps of our translation work. Cognitive interviews revealed relatively few culture-specific issues.

Second, what did we learn from the later behavior coding pretest step?
Subsequent behavior coding also identified general cognitive issues that were relevant to questionnaire design across the four Asian languages. Also, behavior coding identified some additional translation issues that remained after the review and adjudication and the cognitive interview pretest steps.

Cognitive interviewing and behavior coding results were similar in that neither revealed many culture-specific issues. Perhaps the low incidence of culture-specific issues is due to the topics covered by the TUS. We anticipate that other topics would involve more culture-specificity. For example, translation and testing work we are currently conducting for food intake surveys suggests that culture-specific concepts and culture-specific vocabulary are very common when the topic is food.

Third, what would we lose without the behavior coding step?

Behavior coding was particularly effective for identifying long or cumbersome question wordings that needed further editing. Other review procedures might be able to accomplish similar goals, but review, adjudication and cognitive interviewing were not sufficient for this task. We believe that another step – whether behavior coding or some other editorial review – makes an important contribution to the comprehensibility and administration of translated questionnaires.

Finally, we point out one important limitation to the work reported here. Because of the sequential pretesting design we used, we are able to make only relatively general comparisons between cognitive interviewing and behavior coding results. Our design included 11 survey items that were unchanged between the cognitive testing and behavior coding pretests. We are currently analyzing results for those specific items to see whether we can make stronger, item-by-item comparisons of problems detected by each method. We think this comparison will provide a firmer basis for identifying the relative strengths, weaknesses and contributions of the two methods to refined questionnaire design in multi-language surveys.

References

Cannell, C.F., Lawson, S.A., and Hausser, D.L. (1975). "A technique for evaluating interviewer performance." Ann Arbor, Institute for Social Research, The University of Michigan.

Forsyth, B., Rothgeb, J.M., and Willis, G.B. (2004). "Does pretesting make a difference? An experimental test." In S. Presser, J.M. Rothgeb, M.P. Couper, J.T. Lessler, E. Martin, J. Martin, and E. Singer (Eds.) Methods for Testing and Evaluating Survey Questionnaires (pp525-546). NY: Wiley.

Forsyth, B., Kudela, M.S., Levin, K., Lawrence, D., and Willis, G. (2006). "Methods for Translating Survey Questionnaires." Presented at the 61st Annual Meeting of the American Association for Public Opinion Research, Montreal, Quebec, Canada.

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

Harkness, J.A., Van de Vijver, F.J.R., and Mohler, P. (2003). Cross-Cultural Survey Methods. Hoboken, N.J.: Wiley.

McKay, R.B., Breslow, M.J., Sangster, R.L., Gabbard, S.M., Reynolds, R.W., Nakamoto, J.M., and Tarnai, J. (1996). Translating survey questionnaires: Lessons learned. New Directions for Evaluation, 70, 93-105.

Willis, G.B., Lawrence, D., Kudela, M., Levin, K., and Miller, K. (2005a). "The use of cognitive interviewing to study cultural variation in survey

response." Paper presented to the Questionnaire Evaluation Standards (QUEST) Workshop.

Willis, G.B., Lawrence, D., Thompson, F., Kudela, M., Levin, K., and Miller, K. (2005b). "The use of cognitive interviewing to evaluate translated survey questions: Lessons learned." Paper presented to the Federal Committee on Statistical Methodology. Washington, DC.

Willis, G.B., Schechter, S., and Whitaker, K. (1999). "A comparison of cognitive interviewing , expert review, and behavior coding: What do the tell us?" Proceedings of the ASA Section on Survey Research Methods. Alexandria, VA: American Statistical Association.