MAXIMIZING THE QUALITY OF COGNITIVE INTERVIEWING DATA: AN EXPLORATION OF THREE APPROACHES AND THEIR INFORMATIONAL CONTRIBUTIONS

Sarah Daugherty, Lauren Harris-Kojetin, Claudia Squire, and Elizabeth Jaël, *RTI*[1]
Lauren Harris-Kojetin, RTI, 1615 M St, NW, Suite 740, Washington, DC 20036

Key Words: Cognitive interviewing, Pre-testing methods, Response errors

Introduction

Cognitive interviewing refers to a set of 1-on-1 interview methods used in questionnaire development to examine sources of response error (DeMaio and Rothgeb, 1996). The intent of cognitive interviewing is to evaluate the question-answering process by assessing how people interpret and comprehend questions, recall information and events, make judgements about how to respond, and identify answer choices that match their internally generated responses (Tourangeau, 1984). Cognitive interviewing methods are more frequently being used in survey design to improve the accuracy and reliability of survey responses (Forsyth and Lessler, 1991). However, there is little research that has examined whether different cognitive interviewing approaches produce different findings about the questionnaire items being pre-tested (for an exception, see Cosenza and Fowler, 2000).

Probes are often the tools that cognitive interviewers use to prompt respondents to reveal information about their question-answering process (Willis, 1999). Cognitive interviewing employs a variety of probes, including scripted probes (developed prior to the interview) and spontaneous probes (generated by the interviewer during the interview process). Variations in the types of interview probes used and the timing or placement of the probes may affect the amount and type of information revealed by the cognitive interview respondent. This paper explores the extent of differences in findings among 3 cognitive interviewing approaches that vary by the type and timing of probes used.

Three different cognitive interviewing approaches

We examine 3 approaches commonly used in cognitive interviewing practice. Exhibit 1 shows where these 3 approaches lie along 2 dimensions: (1) the types of probes used (tailored versus general) and (2) at what time the probes are administered during the cognitive interview session (concurrent versus retrospective). Since general probes are not usually used retrospectively, we chose not to include a retrospective, general probing approach in the study.

Exhibit 1. Three Cognitive Interview Approaches Defined by Different Types and Timing of Probes

	Types of Probes	
Timing of Probes	Tailored	General
Concurrent	✔	✔
Retrospective	✔	

Types of probes. Tailored probes are designed to address a particular questionnaire item or series of items. Tailored probes can be either scripted or spontaneous and fall into a variety of categories including comprehension, paraphrasing, and recall. General, or non-tailored, probes encourage subjects to reveal as much information as possible about their question-answering process in a way that is not specific to a particular questionnaire. Examples of general probes are, "tell me what you are thinking" and "what is going through your mind right now." Compared to general probes, tailored probes provide greater control for the researcher to tailor the discussion during the interview in order to examine potential areas of response error. However, tailored probing can also be subject to interviewer or researcher bias due to the potential presence of leading probes or comments.

Timing of probes. Both tailored and general probes can be delivered either during or after a subject completes a questionnaire. A cognitive interview where the interviewer asks probes while the subject completes the questionnaire is often called a concurrent interview. A cognitive interview where the interviewer asks probes after the subject completes the questionnaire is often called a retrospective interview.

Cognitive interview approaches examined. Each of the 3 cognitive interviewing approaches we examine are defined by a different combination of the type and timing of cognitive interview probes used. In a ***concurrent, tailored*** interview, an interviewer typically

[1] Research for this paper was supported through the Agency for Healthcare Research and Quality under Cooperative Agreement Number U18HS09218. The opinions expressed here are those of the authors and not necessarily those of the authors' institutions. We thank Charles Darby, Christine Crofton, Floyd J. Fowler, Frederick Conrad, Kelly Sand, and Peyton Williams for their advice and support.

uses tailored probes immediately after the subject has answered a questionnaire item or series of items. A ***concurrent, general*** interview is commonly called a "think aloud" interview. In think-aloud interviewing, subjects are trained to verbalize what they are thinking as they work their way through the questionnaire. Once the cognitive interview begins, cognitive interviewers offer only general probes. This approach has the advantage of eliciting information that is less constrained by interviewer-imposed bias, providing an open-ended format that may help facilitate the discovery of unanticipated findings. However, the think-aloud approach places considerable burden on subjects by providing them with complete discretion on what information to reveal. Because the think-aloud approach is less structured than an approach that uses tailored probes, it may be less effective at uncovering true questionnaire problems. Subjects in think-aloud interviews may also be likely to stray from the question-answering task.

The third type of cognitive interview we tested is a ***retrospective, tailored*** interview. This type of cognitive interview is commonly used to supplement a session where the interviewer first observes a subject's navigation through a self-administered questionnaire, then conducts the cognitive interview. While this approach enables a more realistic, uninterrupted assessment of the flow of the questionnaire completion process than does a concurrent approach, it may increase the recall burden on the subject. This increased burden could result in the subject recalling only some of the issues that arose during the subject's earlier question-answering process.

Two Hypotheses

H1: Tailored versus general probes — The tailored probe approach uncovers more questionnaire problems than the general probe approach.

H2: Concurrent versus retrospective probes — The concurrent approach uncovers more questionnaire problems than the retrospective approach.

These 2 hypotheses favor the concurrent approach that uses tailored probes. The first hypothesis compares the effectiveness of the types of probes used and their relative impact on the frequency of problem identification. Because the tailored approach uses more specific probes that address paraphrasing, comprehension, and recall, the information provided by the subject in response to these specific probes may reveal more response errors than the general probes used in think-aloud interviews. This hypothesis examines the difference between the approaches highlighted in the first row of Exhibit 1.

The second hypothesis investigates the timing of the probes in relation to the subject's question-answering process and their relative impact on the frequency of problem identification. Concurrent cognitive interviewing may have a methodological advantage over retrospective interviewing in identifying more response errors, because of the immediacy of the probe to the subject's question answering experience. This hypothesis examines the difference between the approaches highlighted in the first column of Exhibit 1.

Study Instrument

We used the Consumer Assessment of Health Plans Study (CAHPS®) Version 2.0 core questionnaire as the test instrument in this study. The CAHPS® products are an integrated set of standardized surveys and corresponding report templates that obtain reliable and meaningful information from health plan members about their health care and plan experiences and report the information for consumers to use when making a health plan choice. CAHPS® is a well-tested instrument with both a common core survey that applies to all consumers and supplemental survey item sets that target particular populations.

The CAHPS® team has performed over 100 cognitive interviews and multiple psychometric tests to ensure the validity and reliability of these survey questions (e.g., see Harris-Kojetin et al., 1999 and Hays et al., 1999). However, some outstanding CAHPS core question and response wording issues remained. Therefore, we tested alternative wording and response changes in this study.

Study Design

We incorporated this methodological investigation into a cognitive testing study sponsored by the Agency for Healthcare Research and Quality (AHRQ). We conducted 2 rounds of cognitive interviewing. In each round, we assigned subjects randomly to 1 of 2 cognitive interview approaches being conducted. In Round 1, we tested 2 questionnaire versions. Version A contained the CAHPS® V2.0 core questionnaire items and Version B contained modified versions of many of these core questionnaire items. We tested each of the 2 questionnaire versions using 3 concurrent, tailored interviews and 2 concurrent, general interviews, for a total of 10 interviews in Round 1. In Round 2, we tested only 1 version of the questionnaire that we developed based on the Round 1 findings. We conducted 5 retrospective, tailored and 6 concurrent, tailored interviews in Round 2. We conducted 21 interviews across the 2 cognitive interview rounds, distributed as shown in Exhibit 2.

Exhibit 2. Distribution of Cognitive Interviews (n=21)

	Types of Probes	
Timing of Probes	Tailored	General
Concurrent	12	4
Retrospective	5	0

Six interviewers conducted the 10 interviews in Round 1 and the 11 interviews in Round 2. A team of 2 experienced cognitive interviewing researchers conducted 2 briefing sessions for all interviewers prior to the first interview of each round. In each briefing, interviewers (who were also project researchers) reviewed the study objectives and familiarized themselves with the variations in the cognitive interviewing approaches.

Study Methods

We recruited study subjects from the Raleigh-Durham, North Carolina area and the Washington, DC metropolitan area. Compared to Round 1 subjects, Round 2 subjects were more racially and ethnically diverse (30% to 60% were minority), more highly educated (60% to 73% had at least some college), had more experience with the health care system (70% to 91% had more than 3 doctor visits in the past year), and were older (54% to 83% were age 35 or older).

In each interview, we gave the subject a written copy of the questionnaire and asked them to read each question and give a response verbally (concurrent) or to complete the questionnaire at their own pace (retrospective). In the concurrent interviews, the subject read each questionnaire item out loud. All interviews were audio-taped. After each round of interviews, we transcribed the tapes and cleaned the data. Any responses that corresponded to questionnaire items that a subject should have skipped (based on a previous questionnaire item response) were eliminated from the analysis.

We specified a classification scheme containing four types of response problems. This scheme follows Tourangeau's (1984) general model. The first class of response problems addresses "comprehension" — any misunderstanding of a word, phrase, or response option. We also included problems with comprehension intent in this first class of response problems. Problems with comprehension intent occur when subjects do not follow questionnaire instructions for inclusion or exclusion in answering a questionnaire item. We defined a "retrieval" problem, the second class of response problems, as either a recall error or a miscalculation of the time frame stated in the question (telescoping).

The third class of response problems is a "judgement" problem — a scope misunderstanding in which no explicit inclusion or exclusion criteria are provided in the question. This forces subjects to decide how to respond to the question on their own. If a subject's response does not match the investigators' (i.e., the CAHPS® questionnaire designers) intent for the question item, then we marked this as a judgement problem. The fourth class of response problems occurs with "response mapping," where the subject's desired response is missing from the written survey response choices. The study analysis leader (who also conducted some of the interviews) developed decision rules to determine when the transcribed interview text indicated the identification of a problem and where that problem should be placed in the classification scheme.

We marked 1 count for each unique problem identified in an interview. Problems that were repeatedly identified throughout the same interview were only marked once. For example, a section that contained a series of questions about specialists was only marked once even if the subject incorrectly excluded her OB-GYN as a specialist throughout the question series.

To ensure inter-coder reliability, 2 sets of coding partners coded text individually and then discussed the identification of problem text. Once consensus was reached on the presence and classification of problems in each interview transcript, we coded the transcripts in Nud*ist, a Non-numeric Unstructured Data Indexing Searching and Theorizing database.[2] Using NUD*IST, we developed matrices to review bi-variate relationships (i.e., the interview approach by the type of problem coded). We then performed a final data quality check to ensure that the data were coded consistent with the decision rules.

Using the matrices, we summed the problem counts for all interviews completed using the same interview approach and averaged these to provide the mean number of problems identified per interview using a given approach. This statistic allows for a standardized comparison across interview approaches. If no single interview approach identifies more problems than another, the averages will be the same.

Results

Exhibit 3 provides a comparison of the average number of different types of problems identified per interview by the concurrent, general and concurrent, tailored approaches for Version A of the questionnaire [original CAHPS® core] in Round 1. The concurrent, general approach identified slightly more problems in the comprehension and judgement areas than did the

[2] . Nud*ist helps organize large amounts of text for analytic purposes. Nud*ist is currently 1 of the most widely used qualitative data analysis software products in the world (Quality Solutions and Research, 2001).

concurrent, tailored approach. The concurrent, tailored approach, however, identified twice as many problems in the retrieval area compared to the concurrent, general approach. Neither approach identified any response mapping problems in Version A questionnaire in Round 1.

Exhibit 3. Average Number of Problems by Approach for Round 1, Version A Questionnaire

	Cognitive Interview Approach	
Type of Response Problem	Concurrent, General (n = 2)	Concurrent, Tailored (n = 3)
Comprehension	.25	.17
Retrieval	.25	.50
Judgement	.75	.67
Response Mapping	.00	.00

Exhibit 4 summarizes the average number of types of problems identified per interview by the concurrent, general and the concurrent, tailored approaches for Version B of the questionnaire [alternative question and response wording to some CAHPS® core items] in Round 1. More problems were uncovered by the concurrent, general than by the concurrent, tailored approach in the comprehension and response mapping categories. Concurrent, general and concurrent, tailored interviews identified on average the same number of problems in the judgement category. No retrieval problems were found in any interviews.

Exhibit 4. Average Number of Problems by Approach for Round 1, Version B Questionnaire

	Cognitive Interview Approach	
Type of Response Problem	Concurrent, General (n = 2)	Concurrent, Tailored (n = 3)
Comprehension	.50	.00
Retrieval	.00	.00
Judgement	.50	.50
Response Mapping	.50	.16

Exhibit 5 provides a comparison of the average number of types of problems identified per interview by the retrospective, tailored and the concurrent, tailored approaches in Version C of the questionnaire [modified version of CAHPS® core based on Round 1 results] in Round 2. The concurrent, tailored approach identified more problems across all four problem types compared to the retrospective, tailored approach.

Exhibit 5. Average Number of Problems by Approach for Round 2

	Cognitive Interview Approach	
Type of Response Problem	Retrospective, Tailored (n = 5)	Concurrent, Tailored (n = 6)
Comprehension	1.0	1.6
Retrieval	.33	.60
Judgement	.67	.80
Response Mapping	.67	.80

Discussion

Round 1

Looking across the results for versions A and B of the questionnaire in Round 1, the concurrent, general approach identified more comprehension, judgement, and response mapping problems than the concurrent, tailored approach. These types of problems may be more identifiable using the concurrent, general approach because the subject is encouraged to talk continuously throughout the question-answering process. Therefore, even slight problems may be mentioned as the subject talks through his or her answer. In contrast, the concurrent, tailored approach creates a pre-established dynamic in which each question answered by the subject is followed by a specific probe from the interviewer.

Thus, the concurrent, tailored probing may have caused subjects to narrow the provision of their information to fit specific probes asked by the interviewer. Many slight problems may have been overlooked or not mentioned because the pre-established probing-answering dynamic structured the interaction between the subject and interviewer so that the information offered by the subject was only in response to specific probes.

On the other hand, the concurrent, tailored approach uncovered almost twice as many retrieval problems compared to the concurrent, general approach. Many questions that required a task to be performed (i.e., a summation of doctor's visits over the past 12 months) were followed with specific probes in the concurrent, tailored approach asking for the identification of those visits. Through the process of recalling (i.e., retrieving) the visits over the past 12 months, subjects would often realize they had forgotten a visit. Thus, the specific probing in the concurrent, tailored interviews helped trigger the subject's recognition of retrieval errors.

The concurrent, general and concurrent, tailored approaches produced a different distribution of problems (data not shown in an exhibit). The concurrent, tailored approach uncovered the same

problems repeatedly across interviews. In contrast, the concurrent, general approach rarely identified the same problem across interviews. This difference in the diversity of problems uncovered by the 2 approaches appears to stem from fundamental differences in the structure of each approach. If a particular question is problematic, one would expect problems with that question to be detected repeatedly across the concurrent, tailored interviews because the same probes are asked of all subjects. The concurrent, general approach, on the other hand, uncovered more breadth in the distribution of problems found. The concurrent, general approach seemed to be capturing the unanticipated problems that were not elicited by the probing in the concurrent, tailored approach.

An example that may help illustrate the unanticipated discovery of problems facilitated by the concurrent, general approach occurred with the question, "When you went to a doctor's office or clinic in the last 12 months, how often did you see the doctor or nurse you came to see within 15 minutes of your appointment time?" This question in the concurrent, tailored interviews was always followed up with a probe asking if the subject included the exam room time as well as the waiting room time in their answer. Most responses to this concurrent, tailored probe were typically answered in a yes/no format. However, in the concurrent, general approach, at least 1 subject answered the question after "thinking aloud" about whether a medical technician was included in the "nurse" category. From this information, the researchers were able to identify the part of the question that may have been causing some confusion. As a result of this finding, it was clarified in Round 2 that the waiting time calculated by the subject should be the time from check-in to the time when the subject and the "person they came to see" have their first interaction.

Round 2

In Round 2, the concurrent, tailored approach uncovered more problems than the retrospective, tailored approach across all 4 types of response problems. This may be due to the timing of the probes in the concurrent, tailored approach following immediately after the question-answering process. The retrospective, tailored approach may place more retrieval burden on subjects by requiring them to recall their question-answering process for each question after having completed the entire questionnaire.

More problems were identified in Version C of the questionnaire during Round 2 than in either of the prior questionnaire versions tested in Round 1, even though Version C was intended to be an improvement based on the Round 1 findings. The increase in problems identified in Round 2 is due in part to the fact that there were about 10 more items in the Version C questionnaire than in either Versions A or B in Round 1. The increase in problems identified in Round 2 may also be due, in part, to a change in the screening criteria[3] that resulted in a larger proportion of Round 2 subjects having more health care visits compared to Round 1 subjects. Subjects who were high utilizers of the health care system (almost everyone in Round 2) may have skipped fewer questions throughout the questionnaire and thus would potentially have had the opportunity to encounter more response errors. Finally, interviewers using the concurrent, tailored protocol were encouraged in Round 2 to probe beyond the structured protocol to explore inconsistencies in subjects' responses. This spontaneous probing may have provided opportunities for problem investigation that did not occur in Round 1.

Implications

Field studies may benefit from using both types of concurrent approaches (i.e., with general and with tailored probes) in the same study to capture the variation in respondent problems, both those anticipated and unanticipated by the researcher. The concurrent, tailored approach allows researchers to focus on questionnaire items that they believe may contribute to response error. The concurrent, general approach also makes a valuable and unique contribution, by providing an avenue through which unanticipated problems can be identified.

Combining these 2 approaches for pre-testing questionnaires, for example, by randomly assigning a given sample of subjects to 1 approach or the other rather than conducting only 1 approach can provide an opportunity for greater problem exploration than using either alone. Using both approaches in pre-testing may also be fiscally efficient, because interviewers who perform concurrent, general interviews will require less training than interviewers who conduct concurrent, tailored interviews.

Concurrent interviewing appears preferable to retrospective interviewing in identifying questionnaire problems. This may be because concurrent interviewing creates less of a recall burden on respondents than does retrospective interviewing.

Some survey researchers use an observation with debriefing type of cognitive interview that includes an observational component to monitor subject navigation of a self-administered questionnaire followed by retrospective probes to evaluate potential sources of

[3] Round 2 participants were recruited if they had made 3 or more visits to the doctor in the past year compared to only 1 or more visits for Round 1 participants.

response error. Our findings tentatively suggest that such observation with debriefing interviews should drop the debriefing component. An alternative approach for pre-testing self-administered questionnaires is to have half of the interviews be observational (to focus on navigation only) and the other half use some form of concurrent approach (to examine other potential sources of response error).

Study Limitations

There are several study limitations that may have affected the outcome of our results. CAHPS was a well-tested instrument; therefore, the number of problems identified was minimal. With a small sample size, response bias is an important consideration when interpreting the results. A subject may have had particular difficulty with the questionnaire or may have been unusually talkative so that more problems were identified within that particular interview.

Although extensive efforts were made to ensure inter-coder reliability, misclassification of problems may still have occurred. Problems may have been classified inappropriately or not identified at all. This misclassification error could have modified the results and relationships observed among the cognitive interviewing approaches.

The results of this study are only applicable and relevant to the classification of response problems used. One cognitive interviewing approach may have been favored over another based on the problems chosen for comparison. For instance, observation with debriefing cognitive interviews are known for their use in observing subjects' ability to navigate through a questionnaire. However, because navigation was not included among the response problems examined, the results may have been unfairly biased against the retrospective, tailored approach (which includes an observational component that examines navigation).

Future Directions

While this study serves as an initial investigation attempting to address the gap in knowledge regarding the comparative strengths of various cognitive interviewing approaches, several questions still remain. Further methodological investigation needs to be done to confirm our results. A study with a less complicated design that used the same version of a questionnaire and a larger sample size would provide a clearer understanding of the variations in problem identification among the 3 cognitive interviewing approaches examined in this study. Additional studies could also investigate the reliability and validity of the relationships identified by comparing the approaches across multiple survey instruments or various stages of survey development.

References

Cosenza C. and F. Fowler. 2000. "Flexible or Structured: What's the Best Cognitive Interviewing Protocol." Paper presented at the 55th Annual Conference of the American Association for Public Opinion Research.

DeMaio, T and J. Rothgeb. 1996. "Cognitive Interviewing Techniques: In the Lab and In the Field," pp. 177-195. In N. Schwarz and S. Sudman (eds.) *Answering Questions: Methodology for Determining Cognitive and Communicative Processes in Survey Research.* San Francisco, CA: Jossey-Bass Publishers.

Forsyth, B. and J. Lessler. 1991. "Cognitive Laboratory Methods: A Taxonomy." In P. Biemer et al. (eds.), *Measurement Errors in Surveys.* New York: John Wiley and Sons, Inc.

Harris-Kojetin, L., F.J. Fowler, J. Brown, J..Schnaier, and S. Sweeny. 1999. "The Use of Cognitive Testing to Develop and Evaluate CAHPS 1.0 Core Survey Items." *Medical Care*, Vol. 47 (Supplement), MS10-MS21.

Hays, R., J. Shaul, V. Williams, J. Lubalin, L. Harris-Kojetin, S. Sweeny, and P. Cleary 1999. "Psychometric Properties of the CAHPS 1.0 Survey Measures." *Medical Care*, Vol. 47 (Supplement), MS22-MS31.

Quality Solutions and Research. 2001. *Products: Overview of Version N.4* Quality Solutions and Research International Pty Ltd Melbourne, Australia; QSR International Proprietary Limited. Version 4, 1998. Internet site: http://www.qsr-software.com/products/n4.html. Site accessed August 26, 2001.

Tourangeau, R. 1984. "Cognitive Sciences and Survey Methods." In T. Jabine, M. Straf, M. Tanur, and R. Tourangeau (eds.), *Cognitive Aspects of Survey Methodology: Building a Bridge Between Disciplines* Washington, DC: National Academy of Sciences Press.

Willis, G.B. 1999. "Cognitive Interviewing: A 'How To' Guide." Research Triangle Institute.