DOI: 10.1111/j.1475-6773.2005.00472.x

Role of Cognitive Testing in the Development of the CAHPS® Hospital Survey

Roger E. Levine, Floyd J. Fowler, Jr., and Julie A. Brown

Objective. To describe how cognitive testing results were used to inform the modification and selection of items for the Consumer Assessment of Health Providers and Systems (CAHPS®) Hospital Survey pilot test instrument.

Data Sources. Cognitive interviews were conducted on 31 subjects in two rounds of testing: in December 2002–January 2003 and in February 2003. In both rounds, interviews were conducted in northern California, southern California, Massachusetts, and North Carolina.

Study Design. A common protocol served as the basis for cognitive testing activities in each round. This protocol was modified to enable testing of the items as interviewer-administered and self-administered items and to allow members of each of three research teams to use their preferred cognitive research tools.

Data Collection/Extraction Methods. Each research team independently summarized, documented, and reported their findings. Item-specific and general issues were noted. The results were reviewed and discussed by senior staff from each research team after each round of testing, to inform the acceptance, modification, or elimination of candidate items.

Principal Findings. Many candidate items required modification because respondents lacked the information required to answer them, respondents failed to understand them consistently, the items were not measuring the constructs they were intended to measure, the items were based on erroneous assumptions about what respondents wanted or experienced during their hospitalization, or the items were asking respondents to make distinctions that were too fine for them to make. Cognitive interviewing enabled the detection of these problems; an understanding of the etiology of the problem informed item revisions. However, for some constructs, the revisions proved to be inadequate. Accordingly, items could not be developed to provide acceptable measures of certain constructs such as shared decision making, coordination of care, and delays in the admissions process.

Conclusions. Cognitive testing is the most direct way of finding out whether respondents understand questions consistently, have the information needed to answer the questions, and can use the response alternatives provided to describe their experiences or their opinions accurately. Many of the candidate questions failed to meet these standards. Cognitive testing only evaluates the way in which respondents

understand and answer questions. Although it does not directly assess the validity of the answers, it is a reasonable premise that cognitive problems will seriously compromise validity and reliability.

Key Words. Survey research and questionnaire design, hospitals, cognitive interviewing, CAHPS, patient assessment/satisfaction

---

Cognitively testing survey questions has become the accepted first step in the development of a survey instrument. Prior to 1984, survey questions were evaluated using loosely structured respondent debriefings (i.e., asking the respondents whether they had any difficulties in answering the survey items), interviewer debriefings (i.e., asking the interviewers whether the respondents had any difficulties answering the survey items), or through psychometric approaches, if they were evaluated at all. However, a 1984 conference, bringing survey methodologists and cognitive psychologists together, marked a point when survey methodologists began to accept in principle that questions should be assessed in new ways, which came to be known as cognitive testing (Jabine et al. 1984).

The basic idea behind cognitive testing is straightforward. It is important for researchers to know whether or not respondents understand questions consistently and in the way researchers intended. It is also important that respondents have the information needed to answer the questions, and that they be able to provide meaningful answers using the response tasks provided for the questions. A traditional pretest does not provide information about these issues.

Procedures for cognitive testing can vary greatly. Willis (2005) provides a comprehensive look at the various ways in which such testing is carried out, and the pros and cons of the alternative approaches. For all of the approaches, respondents are asked to answer a test question or a series of questions. Then, some researchers have interviewers ask a series of follow-up probes, asking the respondents to explain how each question was understood and how they arrived at the answers they gave. Other researchers have the respondents think out loud, explaining how they came up with their responses, or have the

---

Address correspondence to Roger E. Levine, Ph.D., American Institutes for Research, 1070 Arastradero Road, Palo Alto, CA 94304. Floyd J. Fowler Jr., is with The Center for Survey Research, University of Massachusetts, Boston, MA. Julie A. Brown, B.A., is with the RAND Corporation, Santa Monica, CA.

respondents think out loud and follow this activity with a series of follow-up probes. Interviewers and researchers use the information they obtain to form judgments about the questions. For example, respondent summaries of what they think is being asked can be compared with one another and with what the researchers intended, to identify questions that are unclear. Explanations of how answers are formed can be studied to find out whether or not respondents have the information needed to answer a question accurately and whether or not the answer they give accurately describes what they have to say.

The use of cognitive testing has evolved and grown since 1984 (e.g., Sirken et al. 1999). Different protocols have been used to test questions (Forsyth and Lessler 1992; DeMaio and Rothgeb 1996; Willis, Rothgeb, and Harris-Kojetin 1999; Willis 2005). For example, some researchers have respondents think out loud as they are answering the questions, while others use only probes. Some researchers carefully script probes for interviewers, while others give interviewers more freedom in what they ask. Some researchers probe each question as soon as it is answered; others let respondents answer several or many questions, and then go back over the questions with their follow-up probes. When questions will be self administered, some researchers have respondents fill out answers on the paper form or computer, while other researchers prefer to have all test questions administered orally by the interviewer. No one approach has been demonstrated to be best (DeMaio and Landreth 2004). However, there is a growing body of literature that demonstrates that only a few cognitive interviews can identify problems with questions that can have major effects on data quality (Fowler 1992, 2004; Forsyth, Rothgeb, and Willis 2004). By identifying those problems early, and making appropriate revisions, the quality of survey data can be greatly improved.

Cognitive testing played an important role in the development of the original Consumer Assessment of Health Providers and Systems (CAHPS®) instrument (Harris-Kojetin et al. 1999). The first step in developing this instrument was to identify the potentially relevant questions from the many survey instruments that had been used to gather consumer experience with health plans. Once a set of these candidate questions had been assembled, they were tested by each of the three organizations that constituted the CAHPS I consortium (Research Triangle Institute, RAND, and Harvard). Alternative wordings were tested in addition to alternative response tasks, the response alternatives that respondents are asked to use to answer the questions. In addition to identifying various ways in which question wording was ambig-

uous or poorly understood, there were three kinds of question problems that were identified through cognitive testing that were particularly important in designing the final instrument:

1. Questions that could not be asked of everyone because some respondents had not had the experience needed to answer the question meaningfully. For example, questions about how medical decisions were made only make sense and can only be meaningfully answered by respondents who have had to make a significant medical decision within the reference period.
2. Questions for which the respondents do not have the information needed to provide meaningful answers. For example, test questions included whether or not respondents' doctors had been told about the results of their visits to specialists. It took only a few cognitive interviews to learn that respondents do not know what their doctors have and have not been told.
3. When the response task given to respondents does not enable them to answer the question meaningfully. For example, CAHPS Health Plan Survey respondents were asked to report their experiences over a 12-month period. Often, their experiences differed from doctor to doctor, or sometimes even across interactions with the same physician. When respondents had both good and bad experiences in a particular domain, it was essential that the answer options enabled them to indicate this. Any question design that assumes things are constant, when in fact they can vary, will be problematic and will not capture what respondents have to say.

Similar procedures were used in the development and evaluation of candidate questions for a survey of patient experiences with the care they received during a recent hospitalization. This survey, called the CAHPS Hospital Survey, is intended "to produce data on patient perspectives on care that allows objective, meaningful comparisons between hospitals that can help consumers make more informed hospital decisions" (Farquhar 2004). Candidate questions were selected from numerous instruments or sets of survey items that had been used to gather the experiences of hospitalized patients. The instruments and survey items that the CAHPS Hospital Survey Instrument Team reviewed were generated by a call for measures published in the Federal Register. Alternative approaches to asking questions covering similar content were grouped together, and a set of test questions was assembled either by adopting wording from those candidate questions directly or by

adapting them to fit a question form that we thought would be appropriate for this instrument. That is to say, these candidate items were not developed de novo. They were, in the expert judgment of the authors and their associates, the best items from those that were submitted in response to the call for measures to provide useful and meaningful information about hospitalized patients' care experiences.

These candidate questions were then subjected to cognitive testing by the three survey organizations that constitute the CAHPS II consortium (the American Institutes for Research [AIR], Harvard, and RAND). This paper describes the procedures that were used, the issues that were identified, and the bases on which candidate questions were dropped, retained, or revised in order to create an instrument for the next phase of instrument evaluation.

METHODS

After preparation of the initial "cognitive testing" version of a CAHPS Hospital Survey instrument, a draft cognitive interviewing protocol was prepared and administered by each of the grantee teams. As it was believed that the instrument would ultimately be administered in at least two modes (by mail, with a telephone follow-up for mail survey nonrespondents), the items were designed for dual-mode administration.

The cognitive interviewing protocol provided scripted probes that could be used to provide insights into each respondent's cognitive processes as he or she answered the pilot items. It concluded with a series of general questions about the items, to allow the respondent to provide additional feedback about the items and to help assess the comprehensiveness of the instrument. A think-aloud training exercise, with practice questions and a scripted response for the interviewers to model appropriate think-aloud behaviors, was also included. After the initial round of cognitive testing, items were revised and new items were developed. A similar protocol was prepared and administered for the second round of cognitive testing.

Slightly different procedures were used for administration of these protocols by each of the three CAHPS II consortium grantee teams. These procedures enabled the testing of the items as both self-administered and interviewer-administered items. The different procedures also reflected the methodological preferences of each of the grantee teams.

To learn how the items functioned as self-administered items, AIR researchers administered the survey as a self-administered instrument. After

training, respondents were asked to think out loud as they completed each item; scripted and unscripted probes were used, as necessary. Harvard researchers used interviewers to administer the survey, conducting some of the interviews as in-person interviews and others as telephone interviews. Prior to the start of testing, respondents were asked several questions about their hospitalization experiences. These introductory items were used to prepare the respondent for answering a series of items about their hospitalization. In addition, the information obtained was used to validate responses, triggering unscripted probing to determine the etiology of perceived inconsistencies (Cosenza and Fowler 2000). RAND researchers administered the survey as an in-person, interviewer-administered instrument for the first round of cognitive testing and, in the second round, as a self-administered instrument, with administration of scripted probes following each item.

All of the teams used the protocol as a guide, administering scripted probes as deemed appropriate and developing and using other probes as needed. The grantee teams conducted a total of 18 interviews in the initial round of testing (December 2002–January 2003) and conducted 13 interviews in the second round of testing (February 2003). Thirteen of these were conducted as self-administered interviews; 18 were conducted as interviewer-administered interviews.

Respondent Characteristics

In both rounds of cognitive testing, respondents were individuals who had been hospitalized for at least 24 hours within the previous 5 months. They were recruited through flyers, word-of-mouth, and postings on electronic bulletin boards. Respondents were compensated from $50 to $75, depending on the location in which they were interviewed. Ten interviews were conducted in Palo Alto, CA; three were conducted in the Raleigh-Durham, NC area; eight were conducted in the Los Angeles, CA area; and 10 were conducted in the Boston, MA area. Twelve of the respondents were males; 19 were females. Respondent ages ranged from 22 to 91 years (mean = 52.0). Twenty-three respondents were white; five were black; and three were Hispanic.

Analysis

Each grantee team used its own analytic procedures to determine which items appeared to be in need of further revision. For example, AIR cognitive interviewers would review their notes and videotapes to prepare a written summary of each interview. This summary would list responses to probes and,

for each item, the interviewer's determination as to whether the interview provided evidence of a "definite problem," a "possible problem," or "no evidence of a problem." If an item was identified as a "possible problem" or a "definite problem," the interviewer would write an explanation of the reason(s) for this judgment. These summaries would be combined, so that every respondent's answers to a probe were listed underneath the probe and, for each item, all judgments as to whether or not the item was problematic for each respondent would be listed together. The interviewers and a senior staff member (R.E.L.) who participated in the development of the item met and reviewed these summaries, identifying problems not previously noted by the interviewer (such as items that measured constructs other than those the items were developed to measure). The results of these reviews were summarized, in terms of general issues and item-specific issues, in a memo report. Each research team prepared a similar report. Senior staff from all three teams then reviewed and discussed each team's memo report as part of the item development process. Items that were felt to be in need of revision, because of compelling evidence provided by one report or because common problems were noted by at least two teams, either underwent revision or were deleted.

These findings were used to inform the development of a pilot test instrument. The pilot test instrument is presently undergoing testing at several dozen different hospitals. This testing will, almost certainly, inform further instrument revisions.

Results and Discussion

Over 70 percent of the items tested were revised or deleted as a result of the cognitive testing. Five broad categories of item problems were identified. These problem categories, with representative items, are presented below.

Items for Which Respondents Lacked Information Needed for Response

As with cognitive testing of the original CAHPS Health Plan Survey, we found that patients lacked the information they needed in order to answer some of the questions. This characterized a series of items designed to measure coordination of care and integration of services. Five candidate items in this domain (Table 1, items 1–5) were drafted for the initial cognitive testing.

All of these items were shown to be seriously flawed. These items tried to measure coordination of care by asking whether hospital staff knew about specific patient characteristics as a result of communication with other staff.

Table 1: Items for Which Respondents Lacked Information Needed for Response

Original Wording	*Revised Wording*	*Reason for Change*
1. Did staff members who cared for you in the hospital know about your condition without having to ask you?	Item deleted	Respondent was not a knowledgeable informant
2. Did staff members who cared for you in the hospital know about any medicines you were taking without having to ask you?	Item deleted	Respondent was not a knowledgeable informant
3. Did staff members who cared for you in the hospital know about any allergies or sensitivities you had without having to ask you?	Item deleted	Respondent was not a knowledgeable informant
4. Did staff members who cared for you in the hospital know about any special needs you had without having to ask you?	Item deleted	Respondent was not a knowledgeable informant
5. How often were staff members who cared for you in the hospital well prepared with information about you? For example, how often did they know about your condition, any special needs you had, the medicines you were taking, or any allergies or sensitivities you had?	Item deleted	Respondent was not a knowledgeable informant; comprehension issues with "condition" and "special needs"
6. During your hospital stay, did you want your family or someone close to you to talk with any of the doctors? a. Did doctors spend enough time talking with your family or someone close to you?	During this hospital stay, how often did doctors, nurses, or other hospital staff involve your family in decisions about your treatment as much as you wanted?	Respondent was not a knowledgeable informant. For revised item, item length compounded comprehension issues; item was deleted

Patients were not knowledgeable informants about these matters. For instance, the items asking about medications and allergies failed because respondents would often respond affirmatively based on their belief that staff must have read their chart rather than on the basis of any experiential knowledge. Similarly, in order to answer questions about whether staff knew about their allergies, medications, or special needs, several respondents used a heuristic:

"The information is in my chart and hospital staff must read the chart." Other respondents felt that staff knew about their allergies because staff looked at their wristband. They used a similar heuristic: "The information is on my wrist band and hospital staff must be looking at my wristband for this information before they give me any medicines." Although this is a reasonable heuristic, affirmative responses to these items were not indications of coordination of care, as no communication between staff was required for a person to gain this information. The item asking about special needs had additional problems: respondents differed in their interpretations of "special needs" and of "condition": some respondents thought "condition" referred to technical knowledge about the condition for which the patient was hospitalized.

Coordination of care is challenging to measure through patient self-report for several reasons. It is a construct that will have its most noticeable impacts only when it is absent and only if this absence is responsible for a problem. As specific behaviors that are reliable, frequently occurring concomitants of well-coordinated care (or are indicators of its absence) could not be identified, coordination of care was a construct for which no acceptable items could be developed.

Other types of items also failed because respondents did not have the information required to answer them. For example, items about family involvement (Table 1, items 6 and 6a) did not function as intended. Patients were not always knowledgeable informants about whether doctors spent enough time talking to family members. As one respondent noted, she "really couldn't say." These items were also problematic because many patients were hesitant to say that they "wanted" a family member to talk with their doctors.

Rather than asking whether doctors spent enough time with family members, a revised item was administered in the second round of cognitive testing. At least three respondents ignored "as much as you wanted" and answered whether they or their family members were involved in treatment decisions. This item did not work for people who either did not want families involved or did not have families. Also, like involvement in decision making, merely communicating information was frequently considered as being involved in decision making. No acceptable item could be drafted for this construct.

Items with Unclear or Ambiguous Terms

Cognitive testing is also an effective way of identifying comprehension problems such as concepts or terms that are understood incorrectly or

inconsistently by respondents. Questions that asked about involvement in decision making and communication about test results provide examples.

Cognitive testing of items 1 and 1a in Table 2, asking about involvement in making treatment decisions, revealed inconsistent interpretations of both "treatment decisions" and involvement. One patient felt that treatment decisions were made prior to hospitalization—the doctors knew what they were going to do; they knew how they were going to manage her pain; and they knew the recovery regimen. Another patient thought that treatment decisions were the decisions that nurses and aides made—not the decisions that doctors make. A revised item was administered in the second round of cognitive testing. Once again, in forming judgments, people included things that were not about their treatments, such as when to get out of bed and when to leave the hospital. Many respondents considered simple explanations by staff about what would be done as involvement in making decisions. No acceptable item about involvement in making treatment decisions (shared decision making) could be drafted.

As noted, all of the existing measures of shared decision making (and coordination of care) were reviewed by survey researchers from three major research organizations in our effort to develop items measuring these constructs. The items that were rejected would almost certainly display similar problems if they were cognitively tested. Accordingly, we urge caution in interpreting indicators of these constructs (in hospital settings) that are based on patient report survey items and encourage further attempts to develop items that will serve as indicators of these constructs.

Cognitive testing also indicated serious comprehension issues with the items about medical tests and the communication of test results (Table 2, items 2 and 2a–c). Respondents failed to generalize the construct "medical tests" and would usually respond to only those examples provided. In response to an explicit probe, most of the respondents indicated they would consider blood pressure checks and having one's temperature taken as medical tests, but they did not think of these behaviors in response to the item. "CT scans" was a term several respondents did not understand. There were also problems with the follow-up questions. The "never-to-always" scale did not work well for the item asking how often staff made sure the patient understood why tests were being performed. For repeated tests, two respondents were told once why they were done—and did not know how to respond. The item asking how often results were provided when the patient was still hospitalized did not work as many tests are not completed during the patient's hospitalization. As there were acceptable items dealing with information and communication, items about communication of medical test results were deleted.

Table 2: Items with Unclear or Ambiguous Terms

Original Wording	*Revised Wording*	*Reason for Change*
1. During your hospital stay, did you want to be involved in decisions about your treatment? a. When decisions about your treatment were made during your hospital stay, how often did doctors involve you as much as you wanted?	During this hospital stay, how often did doctors, nurses, or other hospital staff involve you in decisions about your treatment as much as you wanted?	Lack of consistent interpretation of treatment decisions was confirmed; item was deleted
2. Medical tests include things like drawing blood, X-rays, and CT scans. During your hospital stay, did you have any medical tests? a. How often did doctors or other hospital staff talk with you to make sure you understood why tests were being done? b. How often were you given or told the results of your tests while you were still in the hospital? c. How often did doctors or other hospital staff explain the results of your tests in a way you could understand?	Item deleted	Comprehension issues with "medical tests"; item inappropriate for subgroup (respondents who had the same test performed repeated times)
3. Doctors, nurses and other hospital staff frequently perform medical procedures and tests on patients. For example, they may insert tubes or needles into you, use X-	We want to ask you about medical procedures and tests, for example, drawing blood, taking X-rays, and applying and removing stitches and bandages. During this	Initial item exceeded patient's working memory capacity, resulting in comprehension problems. Double negative in revised item created

continued

Table 2. *Continued*

Original Wording	*Revised Wording*	*Reason for Change*
rays or other machines to take pictures of you, apply and remove stitches and bandages, and move or lift you and the equipment attached to you. How often did these procedures and tests cause you unnecessary pain?	hospital stay did you have any medical procedures or tests? How often were these tests and procedures done without causing you too much pain?	comprehension problems, leading to its deletion

The item about pain associated with tests (Table 2, item 3) began with a two-sentence description of what was meant by medical procedures and tests. Although the description seemed to work adequately when tested as a self-administered item, problems arose when this item was tested as an interviewer-administered item. This introduction led some respondents to think more about testing than about treatment by staff, and was a problem for about half of the respondents to whom the item was interviewer administered. In addition, one respondent answered about pain and did not distinguish between necessary and unnecessary pain, possibly indicating her inability to assess the amount of pain "necessary" or normally associated with a procedure with which she was relatively unfamiliar.

For the second round of cognitive testing, a filter question with a shorter introduction was developed. This filter question allowed a shorter and simpler follow-up item to be developed. Although the second round of testing indicated that the filter question was functioning as intended, cognitive testing revealed serious problems with the revised item. The double negative response (i.e., *never* done *without* causing too much pain) was, not surprisingly, confusing for several respondents. A positive wording might have been developed to overcome this concern. However, the other pain management items appeared to be better candidates for a measure of this domain, leading to the deletion of this item.

Items Not Measuring the Constructs They Were Intended to Measure

Cognitive testing also enabled identification of items that were not serving as good measures of the constructs that they were designed to measure. The following questions about physical environment and comfort provide good examples of this type of problem.

Table 3: Items Not Measuring the Intended Construct

Original Wording	*Revised Wording*	*Reason for Change*
1. How often did the food you were served look and taste good?	Item deleted	Inappropriate for subgroups (those fed via IV; on liquid diets; on special diets)
2. How often was the temperature in your room comfortable?	Item deleted	Inappropriate indicator—many patients were able to control their room's temperature
3. Were you able to be admitted to the hospital as soon as you wanted or needed to be?	Item deleted	Respondents could not consistently determine start of admissions process; item not measuring intended construct

Several problems were associated with the food item (Table 3, item 1), which was intended to serve as an indicator of food quality. One patient was mainly fed intravenously and another was on a liquid diet. Two others had special diets that potentially affected the taste of the foods they were offered; another said he felt so bad that nothing would taste good. For those on no special diet and who felt reasonably well, the question was clear and easily answered. The combination of look and taste did not pose a problem for respondents. But, this item would not serve as an indicator of food quality for many respondents. If it were to be used, an adjustment, such as a screening question, would have to be developed.

The temperature item (Table 3, item 2) was understood and relatively easy to answer. However, we learned that the majority of patients could control their room temperature. For them, the reports of comfort were accurate, but were not a measure of hospital service. So, the item was deleted.

A series of questions designed to measure how well the admissions process was handled by hospitals were also problematic. (These are discussed in greater detail in the next subsection.) With respect to Table 3, question 3, among other problems noted, no evidence was found that responses reflected overfull hospitals or other issues related to access, which is what the question was designed to measure.

Items Measuring Constructs That Are Inapplicable for Many Respondents

A common problem with questions is that they make assumptions about what patients want or need, which are not true for many patients. A series of questions about needs for emotional support and family involvement show how cognitive testing findings were helpful.

Table 4: Items Measuring Constructs That Are Inapplicable for Many Respondents

Original Wording	*Revised Wording*	*Reason for Change*
1. During your hospital stay, were you anxious or concerned about your illness or the effects of your treatment?	During this hospital stay, did you need any emotional support from doctors, nurses, or other hospital staff?	Follow-up item assumes that patients want to talk about anxieties. This proved to be incorrect
2. Did a doctor, nurse, or other hospital staff member talk with you about feeling anxious or concerned?	How often did doctors, nurses, or other hospital staff give you emotional support when you needed it?	Item assumes respondent wants to talk about anxieties
3. Was it easy to find a doctor, nurse, or other hospital staff member to talk to about feeling anxious or concerned?	Item deleted	Item assumes respondent wants to talk about anxieties
4. When you were admitted to the hospital, were there any unreasonable delays before you were taken to your room?	Item deleted	Inappropriate for subgroup (emergency room admits); time period prior to being taken to room is quite heterogeneous, and was not consistently interpreted as intended
5. How often did hospital staff respond quickly when you asked for pain medicine?	During this hospital stay, did you need medicine for pain?	Subgroup of patients did not need medicine for pain

The emotional support items (Table 4, items 1–3) assume that patients want to talk to staff about their anxieties. This was not the case for many respondents, for whom the items did not measure these constructs reliably. For the second round of cognitive testing, new items asking about whether the patient needed any emotional support and how often doctors, nurses, or other hospital staff provided this support were drafted. This testing revealed that emotional support was frequently interpreted as something that only critically ill or severely injured people needed. Eventually, items applicable to nearly all respondents that asked "During this hospital stay, how often did nurses listen carefully to you?" and "During this hospital stay, how often did doctors listen carefully to you?" were drafted and included in the pilot test survey.

As noted previously, questions about the admissions process were often poor questions because they were based on assumptions about the admissions process that did not correspond with the experience of many patients. The

problem with the admissions items (Table 3, item 3 and Table 4, item 4) is that the admissions process is quite varied, and respondents often were not sure what was and was not part of the process. The most complicated were those who were admitted through the emergency room (ER). However, others filled out paper work in a doctor's office before they came to the hospital, or visited the hospital to complete admissions paperwork one or more days in advance of a scheduled procedure. Still others went directly to surgery for some procedures, and then were taken to a room after surgery. When asked about delays, all of these respondents were uncertain which part of their experience to report about. We concluded that we could not draft items that would provide meaningful measures of issues associated with hospital admission that would work for all respondents.

Finally, the item asking about how quickly hospital staff responded to requests for pain medication (Table 4, item 5) did not function as intended because at least four respondents did not ask for pain medicine during their hospital stay. To deal with this situation, and to restrict the pain management items to those respondents who received medicine for pain, a filter question asking whether the respondent needed medicine for pain was developed and incorporated into the pilot test version.

Items Making Discriminations That Are Too Subtle for Many Respondents

Overall ratings of care are found in almost all surveys of hospital care. We used cognitive testing to compare two alternative ways of asking respondents to rate their care. We also used it to see whether ratings of the hospital and the care received in the hospital should be rated separately or whether they were essentially the same ratings in the eyes of respondents.

The second item provided essentially the same information as the 0–10 ratings, except in less detail. Respondents were more willing to give a rating of "10" than a rating of "perfect" as this rating implies that the hospital is providing the best health care it can. A number of patients verbalized their belief that "Nothing is perfect." It was also noticed that some respondents found it easier to produce a numeric rating by combining positive and negative experiences than it was to produce an adjectival rating. The problems that arose were common to both items: several respondents were answering about all of their hospitalization experiences at the focal hospital rather than care associated with the focal visit.

We also found that at least half a dozen respondents thought that the overall health care questions (Table 5, items 1–2) were asking the same thing as

Table 5: Items Making Discriminations That Are Too Subtle for Many Respondents

	Original Wording	*Revised Wording*	*Reason for Change*
1.	We want to know your rating of all the health care you received while you were in the hospital. Use any number from 0 to 10, where 0 is the worst health care possible and 10 is the best health care possible. How would you rate all the health care you received while you were in the hospital?	Item deleted	Many respondents felt that this item was identical with item 3 (below)
2.	In general, how would you rate all of the health care you received while you were in the hospital?" (Poor, Fair, Good, Very Good, Excellent, Perfect)	Item deleted	Many respondents would not say "perfect," decreasing the item's psychometric value
3.	Using any number from 0 to 10, where 0 is the worst hospital possible and 10 is the best hospital possible, what number would you use to rate this hospital overall?	The following was added before the item: Please answer the following questions about the stay at the hospital shown on the cover. Do not include any other hospital stays in your answer	Some respondents would answer about all of their stays at the hospital

the item asking for an overall rating of the hospital (Table 5, item 3). These overall health care items were therefore deleted from the pilot test version. We also proposed that all questions about hospital care should be preceded by an instruction: "Please answer the following questions about the stay at the hospital [shown on the cover]. Do not include any other hospital stays in your answer." This instruction also preceded the first question on the pilot test survey.

General Issues

Several other general issues associated with the development of hospital patient experience survey items were noted as a result of cognitive testing. These include the following:

1. Patients who were in two parts of the hospital during the stay, such as the ER and a regular room, or who were transferred from a regular unit to a rehabilitation unit, had difficulty in answering some items. This was particularly noticeable for those admitted through the ER, who had very different experiences than scheduled admissions. Responses to items about admissions, privacy, and patient safety (being asked about drug allergies, explaining what to eat or drink when taking medications) that are based on ER experiences can be very different from those based on other hospital experiences.
2. The term "hospital staff" has different meanings for different respondents. It does not capture the behavior of doctors reliably. If one wants to capture physician behaviors along with other staff behaviors, a phrase such as "doctors, nurses, or other hospital staff" should be used.
3. If there are complications associated with childbirth, mothers will frequently include the experiences of their child in forming a response to questions about their own experiences.

CONCLUSION

As researchers have been learning consistently as cognitive testing began to be routinely used in the evaluation of survey questions, this process is invaluable in improving the quality of survey measurement. Most of the questions that were tested had been used in some form in other surveys of hospital patient experience. However, they clearly had not been cognitively tested. Over 70 percent of the questions tested were found to have one of five kinds of problems:

1. Respondents did not have the information needed to answer the question.
2. There were unclear or ambiguous terms in the question that caused them to be inconsistently understood (or consistently misunderstood).
3. The answers to the questions were not measuring the constructs they were supposed to measure.
4. The questions included assumptions about what respondents wanted or what they experienced that were not true for many respondents, making the questions unanswerable or the answers lacking in meaning.

5. Questions designed to measure different constructs were basically measuring the same things in the eyes of the respondents, because the distinctions were too fine or subtle for respondents to differentiate them in their answers.

As in the development of the original CAHPS instrument, item problems because of respondents lacking the information needed to answer the question and because of the item's implicit assumptions of the commonality of respondent desires or experiences were detected in cognitive testing. In addition, we detected additional types of item problems and are sharing them in the expectation that this information will facilitate the detection of such problems by others through the use of cognitive testing. Other ways of pretesting questions do not probe the way in which questions are understood and the way respondents actually form their answers so that the meaning of the answers can be assessed. Without cognitive testing, none of these problems would necessarily have been found.

The proof of the importance of cognitive testing is found in the questions that were tested. All of these questions were being used by survey research firms as quality measures for hospitals. The problems that were identified were often very serious. It is almost certain that most of these questions had never been cognitively tested.

There are limitations to cognitive testing alone as a way of evaluating questions. The samples of respondents are typically small and not necessarily representative of the population to be surveyed. Therefore, some issues that affect only a subgroup of the population may be missed. Also, respondents are usually paid volunteers. If they are more motivated than the average survey respondent, they may deal with potential question problems more successfully than unpaid respondents who do not volunteer for a health study. These issues would lead cognitive testing to understate problems with questions.

There is also the possibility that an issue that appears to be problematic in cognitive testing does not actually adversely affect the resulting data—either because it does not occur very often or, for example, because a particular misunderstanding does not have a substantive effect on most answers. Finally, identifying a problem question does not itself ensure that a revised question will be better.

For all these reasons, cognitive testing alone does not guarantee that questions will be good. Additional testing under realistic field conditions, with larger samples that allow psychometric evaluation, has a critical role to play in the presurvey evaluation of survey instruments. However, we, like others (e.g.,

see Presser et al. 2004), have found cognitive testing to be one essential step in identifying problems with questions that would have posed major threats to the validity of the CAHPS Hospital Survey instrument.

ACKNOWLEDGMENTS

Preparation of this manuscript was supported through a cooperative agreement (2U18HS09204-07) from the Agency for Healthcare Research and Quality (AHRQ) and the Centers for Medicare and Medicaid Services (CMS).

REFERENCES

Cosenza, C., and F. J. Fowler. 2000. "Prospective Questions and Other Issues in Cognitive Testing." *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 994–7.

DeMaio, T. J., and A. Landreth. 2004. "Do Different Cognitive Interview Methods Produce Different Results?" In *Questionnaire Development Evaluation and Testing Methods*, edited by S. Presser et al. New York: John Wiley.

DeMaio, T. J., and J. M. Rothgeb. 1996. "Cognitive Interviewing Techniques: In the Lab and in the Field." In *Answering Questions*, edited by N. A. Schwarz and S. Sudman, pp 177–96. San Francisco: Jossey-Bass.

Farquhar, M. 2004. "Update on the CAHPS Hospital Survey: The Development Process." CAHPS 9th National User Group Meeting, Baltimore, December.

Forsyth, B. H., and J. T. Lessler. 1992. "Cognitive Laboratory Methods: A Taxonomy." In *Measurement Errors in Surveys*, edited by P. N. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, and S. Sudman, pp. 393–418. New York: John Wiley.

Forsyth, B. H., J. Rothgeb, and G. Willis. 2004. "Does Pretesting Make a Difference?" In *Questionnaire Development Evaluation and Testing Methods*, edited by S. Presser et al. New York: John Wiley.

Fowler, F. J. 1992. "How Unclear Terms Affect Survey Data." *Public Opinion Quarterly* 56 (2): 218–31.

———. 2004. "Getting Beyond Pretesting and Cognitive Interviews: The Case for More Experimental Pilot Studies." In *Questionnaire Development Evaluation and Testing Methods*, edited by S. Presser et al. New York: John Wiley.

Harris-Kojetin, L. D., F. J. Fowler, J. A. Brown, J. A. Schnaier, and S. F. Sweeny. 1999. "The Use of Cognitive Testing to Develop and Evaluate CAHPS 1.0 Core Survey Items." *Medical Care* 37 (suppl): MS10.

Jabine, T. B., M. L. Straf, J. M. Tanur, and R. Tourangeau, (eds.). 1984. *Cognitive Aspects of Survey Methodology: Building a Bridge between Disciplines.* Washington, DC: National Academy Press.

Presser, S., M. P. Couper, J. T. Lessler, E. Martin, J. Martin, J. M. Rothgeb, and E. Singer. 2004. "Methods for Testing and Evaluating Survey Questions." *Public Opinion Quarterly* 68: 109–30.

Sirken, M. G., D. J. Herrmann, S. Schechter, N. Schwarz, J. M. Tanur, and R. Tourangeau, (eds.). 1999. *Cognition and Survey Method Research.* New York: John Wiley.

Willis, G. B., T. DeMaio, and B. Harris-Kojetin. 1999. "Is the Bandwagon Headed to the Methodological Promised Land? Evaluating the Validity of Cognitive Interviewing Techniques." In *Cognition in Survey Research*, edited by M. G. Sirken, D. Herrmann, S. Schechter, N. Schwarz, J. Tanur, and R. Tourangeau, pp. 133–54. New York: John Wiley.

Willis, G. B. 2005. *Cognitive Interviewing.* Thousand Oaks, CA: Sage Publications.