Using Cognitive Interviews to Develop Surveys in Diverse Populations

Anna M. Nápoles-Springer, PhD,† Jasmine Santoyo-Olsson, MSc,*‡ Helen O'Brien, BA,*†‡ and Anita L. Stewart, PhD*‡*

Background: Conceptual equivalence of measures is essential in research that compares health across diverse racial/ethnic groups. Cognitive interviews are pretest methods to explore the conceptual equivalence of survey items. Systematic approaches for using these methods are emerging.

Objective: We describe an interaction analysis (IA) approach using qualitative data analysis software to analyze transcripts of cognitive interviews in a study to develop a survey instrument of the quality of interpersonal processes of care of diverse patients. Cognitive interviews included standard administration of the survey followed by retrospective probes for selected items.

Subjects: Interviews were completed with 48 Latino, black, and non-Latino white respondents 18 years of age or older with at least one doctor's visit in the past 12 months. Participants averaged 45.8 years in age (standard deviation [SD] = 18.4), 58% were women, and mean education was 14.7 years (SD = 4.0).

Results: Problems were identified in 126 of 159 items (79%). Behavior coding identified 32 problematic items (20%). IA of the transcript of the survey and retrospective probes identified 94 additional problematic items (59%). IA often revealed the nature of the problems, enabling decisions to modify or drop items based on respondents' comments. Behavior coding and IA identified ethnic and language similarities and differences in the use of response sets and the interpretation of items.

Conclusions: IA and behavior coding of cognitive interview transcripts can identify efficiently problems with items and their source to increase the likelihood of the revised items being conceptually equivalent across ethnic groups.

Key Words: measurement, health disparities, survey development, cognitive interviews

(*Med Care* 2006;44: S21–S30)

As the U.S. population becomes more ethnically diverse, research increasingly includes persons from a variety of cultural and socioeconomic backgrounds and examines explanations for extensive racial/ethnic disparities in health and health care. Most self-report measures of health and quality of care were developed for English-speaking respondents with relatively high levels of education, which raises questions about the transferability of these concepts and measures to ethnically diverse groups. Comparing diverse sociocultural groups requires that the measures be conceptually and psychometrically equivalent among groups being compared.[1] Conceptual equivalence is defined as a survey that has equal meaning and content among comparison groups.[2,3] Issues related to psychometric equivalence are discussed elsewhere in this volume.

Cognitive interviews are used widely in questionnaire development to detect items that are not understood by respondents as intended by the survey developers. Generally, cognitive interview methods reflect a theoretical model of the survey response process introduced by Tourangeau[4] that involves 4 stages: comprehension, retrieval, judgment, and response. In other words, the respondent must first understand the question, then recall information, then decide on its relevance, and finally formulate an answer in the format provided by the interviewer.[5] One cognitive interview technique is to ask respondents to verbalize their thoughts while answering survey questions (think aloud). In recent years, cognitive interviewing has relied more heavily upon probes about the interpretation of questions and recall strategies. Such probes may be scripted or spontaneously created by the interviewer; they may be administered immediately after individual survey questions or after completion of the entire questionnaire.[6–8] Results of cognitive interviews identify the types of errors made by respondents, and how they interpret and answer questions.[5,9–11] Cognitive interviews also can be used to revise or develop new items so that they are appropriate to respondents' cultural context and lifestyle. They provide a useful set of tools for examining whether items are being understood similarly across cultures.[3,12–14] Cognitive

From the *Medical Effectiveness Research Center for Diverse Populations and the Center on Aging in Diverse Communities, the †Division of General Internal Medicine, Department of Medicine, and the ‡Institute for Health & Aging, University of California at San Francisco, San Francisco, California.

Supported by grant no. RO1 HS10599 of the Agency for Healthcare Research and Quality, and grant no. P30-AG15272 under the Resource Centers for Minority Aging Research program of the National Institute on Aging, the National Institute of Nursing Research, and the National Center on Minority Health and Health Disparities, National Institutes of Health.

Reprints: Anna M. Nápoles-Springer, PhD, University of California at San Francisco, 3333 California Street, Suite 335, San Francisco, CA 94118-1944. E-mail: anna.napoles-springer@ucsf.edu.

ISSN: 0025-7079/06/4400-0021

interviews usually are used during the pretesting phase of a survey.

Traditional field pretests (administration of a structured survey to smaller representative samples) are limited in that the evaluation of survey items relies heavily on the interviewer's perceptions of respondent comprehension.[15] More recently, field pretests have been augmented with systematic coding of respondent and interviewer behaviors referred to as behavior coding. Behavior coding typically involves reviewing audiotaped interviews and manually assigning to each survey item predetermined codes or categories of interviewer and respondent behaviors, such as the interviewer misreading a question.[16] The frequency with which the problematic behavior occurs is tabulated for each item (eg, 7 respondents requested clarification).[15] Items with a high frequency of problematic codes are reviewed by the research team for potential modification or elimination, and the likely source of the problem is identified to correct the problem. Behavior coding provides useful summary data on how the survey was actually implemented versus how it was intended.[15] Although behavior coding can produce systematic, replicable, and quantitative results,[17] this method may not detect problems when respondents select from available response choices to questions that they have misinterpreted or when respondents answer questions that they did not understand rather than seek clarification.[16] Generally, behavior coding is well suited for identifying some problems, but does not always explain why they exist and may miss other important problems.

A continuum of methods to analyze data from cognitive interviews has been reported. These methods vary in intensity, including review of interviewers' notes and interviewer debriefing sessions,[9,12,14,18,19] subjective review of interview audiotapes or transcripts, including responses to probes,[16] and more objective coding of audiotapes such as behavior coding.[16,20] Although cognitive interviews are used widely, guidelines on optimal methods for conducting or analyzing them are only starting to emerge.[21–23] The use of multiple pretesting techniques has been recommended.[4,8,21]

In this article, we illustrate 2 complementary techniques for analyzing cognitive interview data (transcripts): (1) behavior coding and (2) interaction analyses (IA; content analysis of an interaction between 2 or more people). In the present study, the cognitive interview consisted of a standard administration of the questionnaire followed by cognitive probes on a subset of items. Behavior coding was used to study the initial administration of and response to the survey question. IA was used to study both the administration of the question and subsequent discussions about the meaning of responses obtained from the cognitive probes.[24] We demonstrate how this integrated approach can provide a broad range of feedback on potential issues with survey items through the use of qualitative data analysis software, including the ability to detect when items were not being understood similarly across ethnic or language groups. Qualitative software allowed us to readily access all dialogue (qualitative data) from the administration of an item and/or probe about that item, as well as calculate the frequency with which problem behaviors occurred (quantitative data). We present examples of this methodologic approach from a study to develop a survey instrument of the quality of interpersonal processes of care (IPC) occurring during the medical encounters of black, Latino, and white patients. Ultimately, the utility of a pretest method lies not only in its ability to identify problematic items, but also in the extent to which it suggests ways to improve them.[8]

METHODS

The draft IPC survey was based on a conceptual framework[25] that has 3 major domains, each with several subdomains: Communication (general clarity, elicitation and responsiveness to patient concerns, explanations of condition, processes of care, and self-care), Decision Making (responsiveness to patients' preferences and consideration of patients' ability to comply with treatment), and Interpersonal Style (friendliness, respectfulness, perceived discrimination, cultural sensitivity, emotional support, and empowerment). A fourth domain assesses sensitivity and discrimination among limited English-proficient patients. Based on this conceptual framework, recent literature, and data from 19 focus groups, a pool of over 1000 closed-ended items was developed.

In numerous meetings, the research team carefully reviewed and discussed each item until 159 candidate items were selected for the final survey and cognitive interview pretest based on their potential to be well understood and relevant to the ethnic groups targeted by the study. Pretesting consisted of standard administration of the items followed by scripted and unscripted retrospective probing of selected items. Because respondent burden precluded probing all 159 items, a team of survey researchers experienced in working with diverse populations reviewed the items and selected those items hypothesized as likely to be misinterpreted, culturally inappropriate, or otherwise problematic to the intended audience. Probes were developed for this subset of 41 items. Depending on the potential problem associated with a specific item, 5 types of probes were developed to identify: (1) if respondents understood the intended meaning of specific words or phrases; (2) whether similar questions were perceived as being redundant; (3) the cognitive processes used in responding; (4) if questions were offensive; and (5) if items were culturally appropriate (Table 1). Probes were worded so as to reveal if the hypothesized problem with the item was in fact evident. For example, if a particular term used in an item was thought to be unclear, respondents were asked what that term meant to them.

Initially, all items and probe questions were translated from English into Spanish by 4 bilingual–bicultural researchers with experience in translating surveys (from diverse Latino national origin groups). Terms for some English and Spanish items were culled from transcripts of focus groups conducted in the earlier stages of survey development. English and Spanish versions of items and probes were meticulously evaluated through team meetings of bilingual researchers to discuss discrepancies, which were reconciled by consensus.[26] The aim was to translate the items to be equivalent in meaning and not to perform a literal translation. Thus, if a translated item was not semantically equivalent, the

TABLE 1. Purpose and Example of Probe Questions

Purpose of Probe Question	Example of Probe Question
Explore the meaning of specific words or phrases to respondents	I asked you how often doctors take a genuine interest in you; what does the phrase "genuine interest" mean to you?
Determine whether similar items were perceived as being redundant	How is the phrase "give you advice about your diet and exercise" different from the phrase "talk to you about your diet and exercise"?
Identify cognitive processes involved in answering questions	When I asked you how often doctors tried to understand your culture, you answered *(*read response option selected by respondent such as always, often, and so on); can you tell me what you were thinking when you answered this way?
Determine if questions were viewed as offensive	When I asked you how often you felt discriminated against by doctors because of your race or ethnicity, you answered (read response option selected by respondent such as always, often, and so on); what were you thinking as you picked your answer?
Determine if questions were culturally appropriate	I asked you how often doctors ask you about your health beliefs? What does the term "health beliefs" mean to you?

decentering method was applied, which in this context meant that both the English and the Spanish versions could be modified to maximize their equivalence.[27] Decentering was possible because of the parallel development of a new survey instrument in both languages.

Procedures

Individual cognitive interviews were conducted face-to-face in Spanish or in English (depending on respondent preference) in June and July 2001 at participants' homes or in worksites, cafes, libraries, community-based organizations, or our office. The Institutional Review Board approved all procedures, and written informed consent was obtained from respondents before the interviews. Participants were paid $20.

Four experienced survey interviewers were trained in cognitive interviewing methods. Two of these had prior experience in conducting cognitive interviews. Training included the purpose of cognitive interviews, the importance of preparing respondents on the special nature of these interviews, methods for scripted and spontaneous probing (when the response to the scripted probe was not clearly understood by the interviewer), the critical role of interviewers in assessing the clarity and relevance of items, and the intended meanings of all items. Training included listening to tapes and reviewing transcripts of cognitive interviews. Written protocols for administration of the standard survey and the probes were used. Interviewers were debriefed weekly to identify further training needs. Adherence to the scripted interview protocol was assured by reviewing procedures during these meetings and having interviewers experienced in cognitive methods supervise those who were not until their skills were judged to be adequate. Interviewers were matched by language to interviewees and on ethnicity for most interviews.

Because respondents often do not understand their role in cognitive interviews, we explained that the purpose of the interview was to identify problems with item wording and help us modify the items to improve comprehension. Respondents were given background information about the survey development project. We emphasized their role in helping clarify the questions before administering the final survey to over 1600 patients. To make cognitive interviews work, it is critical to make clear to respondents that their purpose is to provide insight into their interpretations and responses to questions.[28]

Each respondent completed a cognitive interview lasting approximately 60 minutes. We first administered the closed-ended draft IPC survey items in a standard fashion followed by the scripted open-ended probes. Spontaneous probes were used for clarification as needed. We chose to administer the probes after the structured survey because we felt the think aloud or concurrent asking of probes might adversely affect responses or interpretation of subsequent questions.[16,29] In our previous attempts to use think aloud approaches in low socioeconomic and diverse ethnic groups, respondents found it to be awkward and forced. Another reason for using retrospective probing was that we wanted to do behavior coding of the uninterrupted administration of the structured survey as would occur under field conditions. Interviews were audiotaped and transcribed verbatim, including the items and probes. Pauses (but not their length) were indicated in the transcripts as well as nonword verbalizations, eg, "hmmm."

Sample

For the cognitive interviews, Latino, black, and non-Latino white participants 18 years of age or older who had at least one doctor's visit within the past 12 months were recruited from senior centers, community health clinics, unemployment agencies, and colleges throughout the San Francisco Bay area.

Data Analyses

The transcript of each cognitive interview was coded systematically using a qualitative software program, N5: NUD*IST Software for Qualitative Data Analysis.[30] Interviews conducted in Spanish were analyzed in Spanish by 2 bilingual–bicultural investigators experienced in qualitative data coding. Using the qualitative data software permitted simultaneous evaluation of text associated with the item and probe to: (1) systematically and more precisely assign behavior codes using all available information; (2) calculate, by item or code, the frequency of behavior codes; and (3) examine the content of dialogue pertaining to the items and the probes to make decisions as to whether to retain, drop, or modify the items. We describe here the analytic steps involved.

Using the software, the transcript was first reorganized by item so that the dialogue associated with the item and the probe could be reviewed together. An a priori coding scheme for problematic behaviors was developed for all survey questions based on categories in the behavior coding litera-

TABLE 2. Interviewer and Respondent Behavior Coding Categories

Rank	Interviewer Codes	Type of Problem	Definition
0	No change	No problem	Interviewer reads the question as written
1	Accidental skip	Format	Interviewer accidentally skips question
2	Purposeful skip	Format	Interviewer purposely skips question because judges that question does not apply to respondent
3	Opinion of response category	Response scale	Interviewer interprets response provided by respondent that does not correspond to available response options
4	Hard to read	Item	Interviewer experiences difficulty reading the question
5	Slight change	Item	Interviewer slightly changes question but meaning is not affected
6	Repeat question	Item	Interviewer repeats question without being asked to repeat it
7	Opinion of question	Item	Interviewer initially reads question, but adds his/her interpretation of the question, which influences the respondent's answer
8	Major change	Item	Interviewer alters the meaning of the question or response choice at initial reading
Rank	**Respondent Codes**	**Type of Problem**	**Definition**
0	Adequate answer	No problem	Respondent gives satisfactory answer that meets question objective (ie, always, often, sometimes, rarely or never)
1	Own scale	Response scale	Respondent gives an answer but uses a different response scale than that provided (ie. between always and often, all of the time, almost never when available responses were always, often, sometimes, rarely or never)
2	Qualified answer	Instruction	Respondent unsure how to answer since experience varies depending on circumstances
3	Over reports	Instruction	Respondent over reports when answering question (eg, "it only happened once, so always")
4	Inadequate and related	Other	Respondent does not give an adequate answer, but instead tells a story that relates to the question
5	Redundant question	Item	Respondent implies that the question asked is repetitive or similar to a previous question (eg, "you just asked me that")
6	Not applicable	Item	Respondent feels the question is not applicable (eg, when asked "How often did doctors talk to you about smoking?," they answer "I am a nonsmoker so it never comes up")
7	Inadequate and unrelated	Item	Respondent does not give an adequate answer, but instead tells a story that is unrelated to the question
8	Do not know	Item	Respondent indicates he/she does not know the answer to the question
9	Repeat question	Item	Respondent asks to have the question repeated
10	Clarification	Item	Respondent asks for clarification of question or indicates uncertainty regarding meaning
11	Refusal	Item	Respondent refuses to answer the question

ture.[15,31] One scheme was developed for interviewers and one for respondents. These included problems with items, format, instructions, response scales, and "other." Each problem behavior received a code that had been ranked in order of "severity" of the problem, ie, the degree to which it threatened the validity of the answer. For example, interpreting a question for a respondent by offering one's own opinion could influence the respondent's answer and was thus considered a more serious problem than slightly altering the reading of the question without changing its intended meaning. The ranking was used to incorporate information about the severity of the problem, although the primary purpose of the ranking was to streamline the analysis given the time-intensive nature of coding. The rankings provided a system for assignment of only one interviewer and one respondent code to each item. Nine interviewer and 12 respondent behavior codes were independently ranked by 2 members of the research team from the least to the most problematic, with 100% agreement on the rankings. The codes and their rankings are presented in Table 2.

Next, the coding process involved reviewing the text for the item and assigning 2 behavior codes: an interviewer code (eg, slight change) and a respondent code (eg, request for clarification) based on the behavior categories. If 2 behavior codes were possible for either an interviewer or a respondent, the most problematic code was assigned. A code was not assigned to an item when the type of error did not fall within the major categories identified or when there was insufficient information available in the transcript to determine an appropriate code. One bilingual–bicultural investigator (J.S.-O.) first assigned the behavior codes for each item. A second bilingual–bicultural investigator (A.N.S.) independently verified the first investigator's coding by reviewing all verbal comments assigned to each code. This step was facilitated by the software as we were able to resort and print the text by behavior code. The 2 coders discussed discrepancies until consensus was reached.

After assigning and verifying the application of the behavior codes using the software, the data was again resorted by item to facilitate the IA of all items and their respective probes if any. IA was conducted to evaluate the need to modify or drop items and identify the source of problems. In this step, both coders independently reviewed item-by-item all the text for each item. Based on their judgment of the content, the coders independently assigned a code of "understood" or "not understood" depending on whether the text indicated that the respondent did or did not comprehend the item as intended. The 2 coders discussed discrepancies with the full research team until consensus was reached. For all items designated as "not understood," the relevant transcript portions were reviewed by the research team until consensus was reached on how to revise items or whether they needed to be dropped. Again, the software facilitated this process by allowing us to resort the data by the understood and not understood codes.

To analyze the behavior codes, the proportion of interviews for which a problematic respondent or interviewer behavior occurred was calculated by item. This way we were able to estimate the frequency of errors detected using the administration of the survey item only if we had not had the resources to conduct the probes. Typically, if any one of the problem behaviors occurs on any single item in 15% or more of the interviews (eg, item #36 was coded as interviewer repeats question in ≥15% of the interviews) the item would be considered as potentially problematic and would be reviewed by investigators for possible modification.[16,32] To illustrate the added value of IA where we examined the content of the dialogue during the administration of the items and the probes, we have organized our results according to traditional criteria for identifying problematic items. We used this cutoff in organizing our analyses to attempt to understand the additional "yield" of interaction analysis for items not identified through behavior coding, although, in practice, they offer complementary information.

In summary, 159 candidate items for a survey were administered in a standard fashion for a pretest. A subset of items judged by the research team to be potentially problematic were identified for additional probing. For this subset of

TABLE 3. Characteristics of Cognitive Interview Participants by Race/Ethnicity and Language

	Black n (%)	Latino (English) n (%)	Latino (Spanish) n (%)	White n (%)	Fisher Exact Test *P* Value
Total	14 (29)	6 (12)	14 (29)	14 (29)	
Sex					0.5217
Female	8 (57)	4 (67)	6 (43)	10 (71)	
Male	6 (43)	2 (33)	8 (57)	4 (29)	
Age in yrs (mean, SD)	50.1, 22.3	32.7, 10.0	46.8, 15.1	45.9, 19.2	0.0035
Range	23–78	23–47	32–82	25–75	
23–34	5 (36)	4 (67)	2 (14)	7 (50)	
35–50	2 (14)	2 (33)	8 (57)	0 (0)	
51–82	7 (50)	0 (0)	4 (29)	7 (50)	
Education in yrs (mean, SD)	14.9, 2.8	15.5, 2.7	11.4, 4.7	17.4, 2.3	0.0022
Range	10–21	12–19	2–17	14–23	
6 yr or less	0 (0)	0 (0)	2 (14)	0 (0)	
>6 yr to <high school	3 (21)	0 (0)	3 (21)	0 (0)	
High school diploma	0 (0)	1 (17)	4 (29)	0 (0)	
Some college or technical training	4 (29)	2 (33)	0 (0)	1 (7)	
College graduate and higher	7 (50)	3 (50)	5 (36)	13 (93)	
Place of birth					<0.0001
US-born	13 (93)	4 (67)	0 (0)	10 (71)	
Foreign-born	1 (7)	2 (33)	14 (100)	4 (29)	
Insurance					0.3099
Private	8 (57)	4 (67)	7 (50)	5 (36)	
Public*	5 (36)	0 (0)	4 (29)	3 (21)	
Uninsured	1 (7)	2 (33)	3 (21)	6 (43)	

*Public insurance includes Medicaid, MediCare, Medicaid and MediCare, and private insurance and MediCare.
SD indicates standard deviation.

TABLE 4. Item Disposition Using Interaction Analysis and Behavior Coding of Cognitive Interviews

Proportion of Interviews With Problematic Behavior Codes	Probe Status	No. Items	Problems Found Using Interaction Analysis	Item Disposition	
				Dropped	Modified
≥15%	Not probed	24	24 (100%)*	13	7
≥15%	Probed	8	8 (100%)	2	6
Subtotal		32	32 (100%)	15	13
<15%	Not probed	94	73 (78%)	49	24
<15%	Probed	33	21 (64%)	8	13
Subtotal		127	94 (74%)	57	37
Total		159	122 (77%)	72	50

*Four items were not changed based on the interaction analysis.

items, a scripted probe was administered to clarify whether the suspected problem with each of the items in fact existed. Then all items were systematically coded. First, the text associated with the standard administration of the survey items was assigned 2 behavior codes; an interviewer and a respondent code based on a predetermined coding scheme, which focused on specific observed behaviors. The proportion of interviews for which a problematic behavior code occurred was then calculated for each item. Finally, a review was conducted by 2 of the researchers of the content of the text (which we refer to as IA) associated with all 159 of the closed-ended survey items and the open-ended probes. The same coders performed both the behavioral coding and IA and they used the information gleaned from both methods in a complementary fashion. In this way, we were able to use all available data to inform decisions about whether or not to retain, modify, or delete items.

RESULTS

Forty-eight cognitive interviews were completed, including 14 with blacks, 20 with Latinos, and 14 with whites. Six Latinos were interviewed in English and 14 in Spanish. Across groups, the mean age of respondents was 45.8 years (standard deviation [SD] = 18.4 years), 58% were women, and the mean educational level was 14.7 years (SD = 4.0 years). Interviews lasted 43 minutes on average (range, 22–95 minutes) (Table 3).

Table 4 presents results organized by whether or not the items were above the threshold for problematic behavior codes. Using the ≥15% criterion, 32 of the 159 items (20%) were identified as potentially problematic on examining respondent behavior codes. Scripted probes were not asked on 24 of the 32 items because problems were not anticipated in advance, and spontaneous probing did not occur because the interviewer did not perceive a problem with the respondent's interpretation at the time of the interview. After reviewing the content of these items, 15 were dropped and 13 were modified; the other 4 were not changed based on our evaluation of the magnitude of the problem.

For the 127 items not reaching the 15% criterion, review of the content associated with the item and probes (IA) identified 94 additional items (74%) with some problem. IA involved the systematic, detailed review of all dialogue associated with all of the closed-ended and/or probe questions by at least 2 reviewers to assign a code indicating that the item was understood or not understood. When the dialogue for items and/or probes that were not understood was reviewed, the reason for the problem with the survey item could often be identified. Thus, the IA provided valuable information used to rewrite or drop items. Reviewing these problems and the dialogue, we dropped 57 and modified 37 of the 94 items.

Tables 5 and 6 describe the frequency of interviewer and respondent behavior codes by ethnic/language group for all items combined. Interviewer problem codes occurred infrequently (≤3% of the coded text), except for the code "skipping of items," which occurred in 7% to 8% of the text coded for black, English-speaking Latino, and white respondents. The most frequently occurring respondent code for all groups except Spanish-speaking Latinos was respondents' use of a response scale that differed from that provided; this code occurred in 23% of the coded text among blacks.

TABLE 5. Overall Frequency (%) of Interviewer Behavior Codes by Ethnic/Language Group

Group (n)	No Change n (%)	Accidental Skip n (%)	Purposeful Skip n (%)	Opinion of Response n (%)	Hard to Read n (%)	Slight Change n (%)	Repeat Question n (%)	Opinion of Question n (%)	Major Change n (%)	Total* n (%)
Black (14)	916 (82)	86 (8)	17 (2)	27 (2)	0	21 (2)	3 (0.2)	31 (3)	12 (1)	1113 (100)
Latino-Spanish (14)	1035 (93)	5 (0.4)	0 (0)	26 (2)	3 (0.3)	10 (1)	32 (3)	2 (0.3)	0	1113 (100)
Latino-English (6)	438 (90)	2 (0.4)	39 (8)	2 (0.4)	1 (0.2)	0	0	0	1 (0.2)	483 (100)
White (14)	993 (88)	2 (0.2)	75 (7)	15 (1)	0	22 (2)	15 (1)	3 (0.3)	0	1125 (100)

*Codes were not assigned to items when the type of error did not fall within the major categories identified or when there was insufficient information available in the transcript to determine an appropriate code. Total percentages may not equal 100% as a result of rounding error.

TABLE 6. Overall Frequency (%) of Respondent Behavior Codes by Ethnic/Language Group

Group (n)	Adequate Answer n (%)	Own Scale n (%)	Qualified Answer n (%)	Over Reports n (%)	Inadequately Related n (%)	Redundant n (%)	Not Applicable n (%)	Inadequately Unrelated n (%)	Do Not Know n (%)	Repeat Question n (%)	Clarification n (%)	Refusal n (%)	Total* n (%)
Black (14)	608 (60)	235 (23)	9 (1)	2 (0.2)	26 (3)	4 (0.4)	14 (1)	49 (5)	23 (2)	13 (1)	26 (3)	1 (0.1)	1010 (100)
Latino–Spanish (14)	873 (79)	57 (5)	7 (0.6)	1 (0.1)	12 (1)	2 (0.2)	17 (2)	96 (9)	5 (0.5)	22 (2)	14 (1)	2 (0.2)	1108 (100)
Latino– English (6)	374 (85)	24 (6)	0	0	5 (1)	2 (0.5)	5 (1)	10 (2)	7 (2)	6 (1)	9 (2)	0	442 (100)
White (14)	806 (79)	60 (6)	4 (0.4)	5 (0.5)	37 (4)	5 (0.5)	21 (2)	30 (3)	14 (1)	10 (1)	30 (3)	1 (0.1)	1023 (100)

*Codes were not assigned to items when the type of error did not fall within the major categories identified or when there was insufficient information available in the transcript to determine an appropriate code. Total percentages may not equal 100% resulting from rounding error.

Among Spanish-speaking Latinos, the most frequently occurring respondent code was telling an unrelated story that yielded an inadequate response. Otherwise, the pattern of respondent codes tended to be fairly similar across groups.

The major types of problems identified in the survey were participants' lack of familiarity with certain words or phrases, lack of relevance of a question to an individual (ie, not applicable), misinterpretation of questions, and lengthy questions. Other problems identified through IA that occurred less often included problems with response scales, formatting, and instructions. Finally, analyses of the probe questions across ethnic/language groups revealed both cultural differences and similarities in item interpretation as well as questions that may have worked in one language but not the other (English or Spanish).

Specific examples of survey problems that were addressed based on results of the behavior coding and IA are described below.

Clarification of Unclear Terms

When we reviewed the comments associated with specific items where respondents asked for clarification, we found 2 general types of clarification that were needed, the first of which occurred with much greater frequency: (1) of specific terms and (2) of general instructions. For example, 26% of respondents asked for clarification of the item, "Have you had any medical tests or procedures in the past year?" A review of comments on item and probe questions revealed that respondents were unsure of the meaning of medical tests and procedures. Comments on a probe that asked what they thought was meant by "medical tests or procedures" indicated that whites appeared to interpret this phrase more broadly and included cosmetic surgery and dentistry. Latinos (both English- and Spanish-speaking) were more likely to say they did not understand the phrase, but then gave blood tests as examples. The item was thus modified to include some examples: "In the past 12 months, have you had any medical tests or procedures, such as blood tests, x-rays, or cancer screening tests?"

In another example, problematic behavior codes were below the threshold (11%) for the item, "How often did doctors give you advice about your diet?" Review of comments about the item revealed that 3 respondents requested clarification or gave an inadequate answer regarding what constituted "advice" about one's diet (eg, whether referral to a dietitian was sufficient). Additionally, IA of the probe that asked, "How is the phrase 'give you advice about your diet' different from the phrase 'talk to you about your diet'" revealed that respondents perceived a difference between the terms "talk" and "advice." Respondents viewed "talking" as being involved in a discussion and "advice" as being told what to do. We revised the item to read, "How often did doctors talk to you about your diet?" to capture the participatory nature of patients being included in their self-care (under the self-care subdomain of the instrument).

Even for items that were not probed, IA of the closed-ended items often identified problems. For example, the questions, "How often did doctors explain what was causing your health problem?" and "How often did doctors explain your diagnosis?" were not probed nor did the frequency of behavior codes detect a problem. However, IA found that some respondents asked what the difference was between the 2 questions (ie, redundancy), while others did not know what was meant by diagnosis. We dropped the "diagnosis" item and retained "How often did doctors explain what was causing your health problem?" Had we not preformed IA of these closed-ended items, we probably would have retained the diagnosis item because it was shorter and because we expected most respondents to understand the item.

Improving the Equivalence of English and Spanish Versions

Sometimes English-speaking respondents sought clarification of specific terms, while Spanish-speaking respondents did not, or vice versa. For example, for the item that asked, "How often did doctors ask you about your health beliefs?" behavior coding indicated that 33% of the English-speaking and none of the Spanish-speaking respondents sought clarification. Reviewing the comments revealed that English-speaking respondents were unclear as to what we meant by "health beliefs"; they were interpreting health beliefs broadly to mean "personal beliefs," which included alternative medicine and religious beliefs. In contrast, the Spanish-speaking respondents interpreted the Spanish translation "ideas y creencias acerca de la salud" as it was intended. We revised the English item to read more specifically, "How often did doctors ask you if you have any personal beliefs about your health?," and then translated the Spanish item to be equivalent with the English version "¿Con qué frecuencia le preguntaron los doctores si tiene algunas creencias personales sobre su salud?"

Using Closed-Ended and Probe Data to Explore Group Differences

Reviewing data for the items and probes provided information that was critical in determining whether or not questions were culturally appropriate or offensive. Probe questions often detected cultural differences in interpreting items. For example, the item, "How often did doctors take a genuine interest in you?" seemed to have good face validity because respondent codes indicated that 100% of respondents gave adequate answers. However, verbal comments to the probe indicated that whites interpreted "genuine interest" to mean that doctors take an interest in their health problems, which was not the meaning we intended. Blacks, however, usually interpreted this question as intended, that doctors saw them as individuals. Finally, many Spanish-speaking Latinos did not know the meaning of the word "genuino" (the Spanish translation of "genuine"). In addition, a review of the text coded under the "hard-to-read" interviewer code indicated that Spanish-speaking interviewers often had trouble reading the term "genuino." Thus, we revised the item to read, "How often did doctors respect you as a person?," and in Spanish "¿con que frecuencia le trataron los doctores con respeto?"

We experienced problems developing questions to capture complex constructs such as cultural sensitivity. For example, the item, "How often did doctors ask you if you wanted to include your family when making decisions?" appeared to be acceptable because 86% of respondents gave adequate answers to the closed-ended question. Most Spanish-speaking Latinos commented during the survey administration and the probing that family involvement was an important aspect of their medical care and it made them feel their doctors cared about them and their family's well-being. However, IA revealed that although black and white respondents provided an adequate answer to the question, they found the question irrelevant and felt that involving family was only appropriate when a genetic or serious health condition existed. Based on this review, the item was dropped.

The probes revealed that the item, "How often did doctors look at you when you were talking?" was not understood similarly across ethnic/language groups. An older black woman indicated it would make her uncomfortable most of the time to have a doctor look directly at her. Several Latinas were unclear whether it meant doctors looking them in the eyes or at their body. Whites stated that it depended on the type of appointment, eg, if it was for a foot problem, they expected doctors would look at their feet and not at their face. Thus, we dropped this item.

In earlier focus groups, blacks indicated that doctors sometimes "had a negative attitude toward you," so we included an item asking, "How often did doctors have a negative attitude toward you?" We did not know if this phrase was salient for all groups. Comments from the probing indicated that all groups interpreted the phrase as meaning that doctors were rude, brusque, disapproving, and critical. Thus, we retained the item.

DISCUSSION

We have illustrated how systematic coding of cognitive interview transcripts using behavior coding and IA can help identify problems in items and, perhaps more importantly, can provide information on the nature of the problems that can be used to modify items. We also have illustrated how much information can be obtained through IA of the basic interview without adding probe questions, which are often more costly to administer and to analyze. We thus offer alternative methods for evaluating how respondents answer questions depending on the availability of resources.

As in previous studies, the behavior codes provided useful quantitative summaries of interview outcomes at the item level.[8] When problems are suspected, researchers also may review the content of the dialogue relating to the survey items and probes. In most cases, systematic IA of the dialogue that ensues during the closed-ended administration of the item is overlooked and rarely integrated in a systematic way with data on the probe and frequency of behavior codes. Reliance on probes alone can be problematic because only a limited set of items can be probed as a result of respondent burden. As has been recommended, the richness of data from the use of multiple methods helped us not only to detect issues with specific items, but to understand the nature of those issues.[8] As previously found,[15] behavior coding and cognitive interview probing were complementary. Our integrated approach identified potentially problematic items and enabled informed decisions as to what to do about them. However, researchers need to bear in mind that evaluation of questionnaire items entails subtle choices where the best choice may not always be clear. For example, modification of an item that was problematic in a subset of respondents may have actually resulted in making the item more ambiguous for respondents who had originally understood it.

We found that respondent-related errors occurred more often than interviewer errors. Deviating from the response set appeared to be especially problematic among blacks, whereas Spanish-speaking Latinos were more likely to tell an unrelated story. Perhaps training respondents on the nature of structured surveys is especially important for these groups. Further research needs to explore whether these are culturally influenced response styles. Examining the content of the dialogue associated with the survey items and probes revealed differences across ethnic/language groups in the interpretation of the meaning and scope of key phrases and their relevance. Submitting reworded questions to further cognitive interviewing is necessary to increase the likelihood that items are being understood similarly across groups.

We used qualitative data analysis software to organize the dialogue, the behavior codes, and the frequency of codes at the item level. In the early analytic stages, this data structure allowed us to review the adequacy of our behavior-coding scheme, make necessary modifications, and more precisely characterize the definition of the problem code. In this study, IA of interviewers' and respondents' comments often provided additional information to identify the source of the problem otherwise not detectable when examining the frequency of behavior codes alone. Behavior codes, on the other hand, provided excellent summary data on the frequency of codes across items and by ethnic/language group. Even for questions that were not probed due to time limitations, organizing the data by item enabled us to review the dialogue that transpired during the administration of an item to decipher the underlying problem. However, the possibility also exists that the detailed review of the text associated with each item resulted in false-positives, the identification of items as "problematic" when they were actually comprehended as intended or would be in a larger sample. Moreover, the probability that any item would be considered problematic is high given the small sample size. Although coding of each item was independently reviewed by a second coder, problems with the reliability and validity of coding across interviewers have been noted by others and deserve further exploration.[4,33] Future studies might compare across groups the degree of differential item functioning of alternate versions of items that have and have not been developed using cognitive interviewing techniques.

Studies using cognitive interviews to pretest surveys rarely provide detailed descriptions of the data analysis methods. Thus, this article should prove useful to researchers seeking to add cognitive interviews to their repertoire of pretesting techniques for survey development. These pretesting techniques are especially important to assess the conceptual adequacy of new or adapted self-report measures across ethnic groups in studies of health disparities.

REFERENCES

1. Stewart AL, Nápoles-Springer AM. Advancing health disparities research: can we afford to ignore measurement issues? *Med Care*. 2003; 41:1207–1220.
2. Barofsky I. The role of cognitive equivalence in studies of health-related quality-of-life assessments. *Med Care.* 2000;38(suppl):II125–II129.
3. Weech-Maldonado R, Weidmer BO, Morales LS. Cross-cultural adaptation of survey instruments: the CAHPS experience. In: Lynamon ML, Kulke RA, eds. *Seventh Conference on Health Survey Research Methods*. Hyattsville, MD: Department of Health and Human Services; 2001:75–81.
4. Tourangeau R, Rips LJ, Rasinski K. *The Psychology of Survey Response.* Cambridge, UK: Cambridge University Press; 2000.
5. DeMaio TJ, Rothgeb JM. Cognitive interviewing techniques: in the lab and in the field. In: Schwarz N, Sudman S, eds. *Answering Questions. Methodology for Determining Cognitive and Communicative Processes in Survey Research*. San Francsico: Jossey-Bass, Inc; 1996:177–196.
6. Beatty P. The dynamics of cognitive interviewing. In: Presser S, Rothgeb JM, Couper MP, et al, eds. *Methods for Testing and Evaluating Survey Questionnaires*. Hoboken, NJ: John Wiley & Sons, Inc.; 2004: 45–66.
7. Willis GB. Cognitive interviewing revisited: a useful technique, in theory? In: Presser S, Rothgeb JM, Couper MP, et al, eds. *Methods for Testing and Evaluating Survey Questionnaires*. Hoboken, NJ: John Wiley & Sons, Inc.; 2004:23–43.
8. Willis GB, DeMaio TJ, Harris-Kojetin B. Is the bandwagon headed to the methodological promised land? Evaluating the validity of cognitive interviewing techniques. In: Sirken MG, Herrmann DJ, Schechter S, et al, eds. *Cognition and Survey Response*. New York, NY: John Wiley & Sons, Inc; 1999.
9. Harris-Kojetin LD, Fowler FJ Jr, Brown JA, et al. The use of cognitive testing to develop and evaluate CAHPS 1.0 core survey items. Consumer Assessment of Health Plans Study. *Med Care.* 1999;37(suppl):MS10–MS21.
10. Jobe JB, Mingay DJ. Cognitive research improves questionnaires. *Am J Public Health*. 1989;79:1053–1055.
11. Royston P, Bercini D, Sirken M, et al. *Questionnaire Design Research Laboratory*. Proceedings of the Section on Survey Research Methods of the American Statistical Association; 1986.
12. Carbone ET, Campbell MK, Honess-Morreale L. Use of cognitive interview techniques in the development of nutrition surveys and interactive nutrition messages for low-income populations. *J Am Diet Assoc*. 2002;102:690–696.
13. Johnson TP, O'Rourke D, Chavez N. Cultural variations in the interpretation of health survey questions. In: Warnecke R, ed. *Health Services Research Methods Conference Proceedings*. Hyattville, MD: National Center for Health Statistics; 1995.
14. Warnecke RB, Johnson TP, Chavez N, et al. Improving question wording in surveys of culturally diverse populations. *Ann Epidemiol*. 1997; 7:334–342.
15. Fowler FJ Jr, Cannell CF. Using behavioral coding to identify cognitive problems with survey questions. In: Schwarz N, Sudman S, eds. *Answering Questions: Methodology for Determining Cognitive and Communicative Processes in Survey Research*, 1st ed. San Francisco, CA: Jossey-Bass Publishers; 1996:15–36.
16. Oksenberg L, Cannell C, Kalton G. New strategies for pretesting survey questions. *J Off Stat*. 1991;3:349–365.
17. Fowler FJ Jr. *Improving Survey Questions: Design and Evaluation. Applied Social Research Methods Series*. Thousand Oaks, CA: Sage Publications, Inc; 1995:1–191.
18. Jobe JB, Mingay DJ. Cognitive laboratory approach to designing questionnaires for surveys of the elderly. *Public Health Rep*. 1990;105:518–524.
19. Riley A, Rebok G, Forrest C Young children's reports of their health: a cognitive testing study. In: Cynamon ML, Kulka RA, eds. *Seventh Conference on Health Survey Research Methods*. Hyattsville, MD: National Center for Health Statistics; 2001:19–26.
20. Loosveldt G. Interaction characteristics of the difficult-to-interview respondent. *International Journal of Public Opinion Research*. 1997;9: 386–391.
21. Presser S, Couper MP, Lessler JT, et al. Methods for testing and evaluating survey questions. *Public Opinion Quarterly*. 2004;68:109–130.

22. Sirken MG, Herrmann DJ, Schechter S, et al. *Cognition and Survey Research*. New York, NY: John Wiley & Sons, Inc; 1999. (Groves RM, Kalton G, Rao JNK, et al, eds. Wiley Series in Probability and Statistics).
23. Willis GB. *Cognitive Interviewing*. Thousand Oaks, CA: Sage Publications; 2005.
24. Blixt S, Dykema J. *Before the Pretest: Question Development Strategies.* Proceedings on the Section of Survey Research Methods, American Statistics Association; 1993:1142–1147.
25. Stewart AL, Nápoles-Springer A, Perez-Stable EJ, et al. Interpersonal processes of care in diverse populations. *Milbank Q*. 1999;77:274, 305–339.
26. Harkness J, Pennell BE, Schoua-Glusberg A. Survey questionnaire translation and assessment. In: Presser S, Rothgeb JM, Couper MP, et al, eds. *Methods for Testing and Evaluating Survey Questionnaires*. New York: John Wiley and Sons, Inc; 2004:453–473.
27. Marín G, Marín BV. *Research With Hispanic Populations*, vol 23. Newbury Park, CA: Sage Publications, Inc; 1991. (Bickman L, Rog DJ, eds. *Applied Social Research Methods*.)
28. Krause N. A comprehensive strategy for developing closed-ended survey items for use in studies of older adults. *J Gerontol B Psychol Sci Soc Sci*. 2002;57:S263–S274.
29. Sudman S, Bradburn NM, Schwarz N. *Thinking About Answers: The Application of Cognitive Processes to Survey Methodology,* 1st ed. San Francisco: Jossey-Bass, Inc; 1996.
30. *N5 (Non-numerical Unstructured Data Indexing Searching & Theorizing)* [qualitative data analysis program]. Version 5.0. 2000.
31. Catania J. *Overview of Methods for Cognitive Testing of Self-Reported Measures.* Excellence Centers to Eliminate Ethnic/Racial Disparities (EXCEED) Pre-Conference Workshop; San Francisco, CA; 2001.
32. Fowler FJ Jr. How unclear terms affect survey data. *Public Opin Q*. 1992;56:218–231.
33. Conrad FG, Blair J. Data quality in cognitive interviews: the case of verbal reports. In: Presser S, Rothgeb JM, Couper MP, et al, eds. *Methods for Testing and Evaluating Survey Questionnaires.* Hoboken, NJ: John Wiley & Sons, Inc; 2004:67–87.