EVALUATING SUBJECTIVE HEALTH QUESTIONS: COGNITIVE AND METHODOLOGICAL INVESTIGATIONS

Paul Beatty, Susan Schechter, and Karen Whitaker, National Center for Health Statistics
Paul Beatty, 6525 Belcrest Road, Room 915, Hyattsville, MD 20782

Key Words: Cognitive interviewing, probing, interviewer behavior

Introduction and Overview

Measurement of subjective phenomena has long been of interest to survey researchers. Although many health surveys focus more attention on objective measurements such as frequency of health-related behaviors, interest in self-assessed "quality of life" measures has grown considerably in recent years (Schechter, 1993). Questions about self-assessed health (physical, mental, and overall) have become an important component of health surveillance, used to track progress toward "Healthy People 2000" goals of maximizing years of healthy life (Erickson, Wilson, and Shannon, 1995). Furthermore, self-assessed health status has proved to be a more powerful predictor of mortality and morbidity than many objective measures of health (Idler, 1992). This power, along with the simplicity of administering these questions, makes them quite valuable to researchers.

One survey that makes extensive use of such questions is the Behavioral Risk Factor Surveillance System, or BRFSS, sponsored by the Centers for Disease Control and Prevention (CDC). The BRFSS is a telephone survey of non-institutionalized adults, conducted at the state level on a continuous basis. Over the last several years, the National Center for Health Statistics (NCHS) has worked on several projects with BRFSS researchers to identify potential problems with these questions and to establish their methodological and conceptual integrity. In 1995-96, NCHS conducted two rounds of cognitive interviews to illuminate the meaning of responses to subjective health questions, and to identify difficulties with comprehension, retrieval of relevant information from memory, response selection, and so on.

In this paper, we discuss results of those cognitive interviews, as well as results of our investigation of cognitive interviewing methodology issues. Observations from initial cognitive interviews suggested that many subjects were unable to provide codeable responses to the questions. Several BRFSS researchers asked us if cognitive interviewing could have actually created the problems we observed. From their perspective, the questions had been successfully fielded in a prior RDD survey. Since cognitive interviews encourage discussion, and interviewers are permitted to depart from standardized interviewing techniques, they wondered if it possible that this style of interviewing found "problems" that would not exist under actual survey conditions. Furthermore, BRFSS researchers asked for evidence of the magnitude of these problems, beyond anecdotes and qualitative observations.

This project had three major components: (1) the initial round of cognitive interviews; (2) the development and application of a coding system to analyze the cognitive interviewing data; and (3) a second round of laboratory interviews conducted without intensive probing, to explore the relationship between probing and responses in cognitive interviews. We discuss each of these components in turn.

Initial Cognitive Interviews

Methods

NCHS tested BRFSS Quality of Life (QoL) questions through cognitive interviews conducted at the Questionnaire Design Research Laboratory in Hyattsville, MD. Eighteen subjects were recruited, most of them through newspaper advertisements. However, in order to fulfill a BRFSS request for some subjects over the age of 75, additional subjects were recruited through a local senior citizens center. All subjects were paid $30 for participation in a one-hour interview. Subjects were evenly divided among males and females, ranging in age from 21 to 94, with a median age of 62. Their education ranged from 4th grade to college graduates, with a median level of 12th grade.

All interviews were conducted by NCHS staff members, who used scripted and unscripted concurrent probes (by concurrent, we mean that the probes were administered immediately following the response to a given survey question). For example, one QoL question was "Now thinking about your physical health, which includes physical illness and injury, for how many days during the past 30 days was your physical health not good?" The scripted probes were: *How did you decide on that number of days? Describe for me the illnesses and injuries that you included in your answer. Did you have any difficulty deciding whether days were "good" or "not good"?* Interviewers were instructed to rely on their own discretion regarding which of the scripted probes to actually use. Unscripted probes were also allowed based on interviewer discretion.

Results

There were eight "30-day reference period" QoL questions on the instrument we tested, each of which asked subjects to report "How many days during the past 30 days" some subjective health evaluation applied to them. The intent of the question was for respondents to provide a number from zero to 30 as their answer.

Interviewers agreed that many subjects had difficulty providing these "numbers of days" responses. Sometimes answers were simply indirect (e.g. "I have no complaints" or "I feel that way all the time"). Some survey researchers would argue that interviewers could accept such answers as equivalent to numeric codes, but probing revealed that numerical assumptions were not always warranted. One subject claimed that she experienced pain "like every day," but follow-up probing revealed that actually "in the last couple of weeks I haven't noticed it as much." Often, subjects talked in general terms--they would describe how often they did not feel well, but did not provide quantitative responses, even when probed. Below is an example of our analysis regarding one of the questions:

Q2. Now thinking about your physical health, which includes physical illness and injury, for how many days during the past 30 days was your physical health not good?

Probe: How did you decide on that number of days? Describe for me the illnesses and injuries that you included in your answer. Did you have difficulty deciding whether days were "good" or "not good"?

Only two of eighteen subjects provided clear numerical responses (both of whom said "2 to 3 days"). Eight subjects gave responses that implied "zero." Most gave answers such as "my physical health is good, I don't remember in thirty days seeing any doctors." None of these subjects spontaneously offered "zero" or "none," and only two of them did so after probing.
Two subjects answered "all the time" or "every day."

The first subject initially answered by discussing her arthritis in general terms. After a repeat of the question she said "Every day I have pain in my neck and shoulder, but I try to distract my mind." More probing was required before she confirmed that "every day" she considered her physical health to actually be "not good."

Six subjects gave responses between zero and thirty. Two of them provided numeric answers without any probing; the others talked more generally and had to be probed before providing one. One reported that "last night arthritis bothered me an awful lot. Aside from that it's fairly good... [that was] the first time in my whole life." This implied an answer of one day, but repeated probing failed to elicit a numerical answer. Two subjects could not answer at all-- one gave a response too ambiguous to code, and another refused to accept the idea of "number of days" in answering.

Discussion

As mentioned previously, our observation that subjects had difficulties adhering to question format were surprising, because BRFSS tabulations showed that a very small percentage of RDD responses were non-numeric. Although we expect cognitive interview results to reveal more problems than survey results, the magnitude of this discrepancy seemed striking. Every cognitive interviewer for the project agreed that the response problems seemed pervasive and serious, especially as compared to the survey data. One possible explanation for this discrepancy was that survey interviewers accepted imprecise answers. Perhaps some field interviewers "glossed over" these imprecisions, making the data appear cleaner than they actually are (e.g., recording "gosh, it might have been the whole time" as "30 days"). The BRFSS is conducted by individual states-- such decentralization could reduce the consistency of interviewer training for handling such situations (Fowler and Mangione, 1990).

A second possibility was that field interviewers often elicited precise answers, but that doing so required a high degree of probing. That would explain why numeric answers were prevalent in telephone interviews, and would not contradict our claim that the questions would probably be difficult for many respondents. Furthermore, several studies suggest that interviewer errors increase when questions require extensive probing (Mangione, Fowler, and Louis, 1992; Fowler and Mangione, 1990). We recommended that BRFSS staff monitor survey interviews to explore how much probing was taking place during administration of these questions.

A third possibility was that respondents did provide straightforward answers in the field-- and it was only during cognitive interviews that these problems emerged. If true, this raises another question: did the cognitive interviews provide insight into "real" problems that were hidden in the field, or were these problems merely artifacts of the cognitive interviewing method (i.e., created by the less formal, discussion-oriented mode of operation)?

Addressing these possibilities required a more rigorous look at the cognitive interviewing data. The initial analysis relied on interviewers' subjective assessments of whether subjects gave "clear numerical answers." Coding the interview data using more objective criteria might strengthen the evidence of our conclusions. Objective codes might also allow us to examine the relationship between probing and responses in cognitive interviews.

Coding the Cognitive Interview Data

Methods

We coded the cognitive interview data at the question level based on transcriptions from tape-recordings of the interviews. For each question, we recorded (1) the actual response to the survey question; (2) a code for how closely the response matched the expected format of response (which we labeled "precision"); and (3) a code for the amount and type of probing that preceded the response.

1) Coding the response

Identifying a subject's response to the survey question is not always straightforward in a cognitive interview. Our objective was to find the "best" answer given in terms of adherence to the question format. That is, precise figures (such as "3 days") would be preferable to closed ranges ("3-5 days"), closed ranges would be preferable to vague ones ("at least three days"), and so on. In reviewing the transcripts, the first answer that we thought would have been acceptable to a survey interviewer was coded as the response. If subjects did not provided a precise answer, we recorded the response that was closest in adherence to the response format. Sometimes, of course, there was no response that fit the response format-- subjects rejected the premise of the question ("I can't tell you in days") or gave answers so vague that no answer could be reasonably inferred ("I have these minor problems, but it's more of an inconvenience"; "I think it's most of the time"). Those responses were labeled "uncodeable."

2) Coding precision of response

After identifying the "best" answer, we evaluated its *precision*. We use this term to refer to how closely the response conformed to the question format, not the validity of the response. We devised a four-point scale and assigned values as follows:

Precision Codes

Code 0: The answer requires virtually no rounding, judgment, or interpretation from the coder. Examples:

- *Precise quantities*: ("30 days"; "4 days").
- *Certain colloquialisms*: "Every day" is acceptable as 30; "Never" is acceptable as zero.

Code 1: The answer requires minimal interpretation from the coder. Examples:

- *Moderate qualifiers* ("Probably every day"; "I think a day or so").
- *Narrow range of days* ("Six or eight days")-- record the midpoint.
- *Fractions* ("Half of the days" can be recorded as 15 days, etc.)
- *Lazy format where the meaning is obvious* (for example, a subject answers "I have no problems" after a series of "zero" answers).

Code 2: The answer requires considerable interpretation from the coder.

- *Broad ranges of days* ("sixteen to twenty")-- take the midpoint.
- *Anchored but qualified responses* ("more than 15 days")-- Accept the "anchor" point.

Code 3: The answer cannot be coded ("I can't put the answer in days"; "For a while I was in horrible pain").

3) Coding probes

We distinguished between two types of probes. A re-orienting probe asks the subject to re-focus on the task of providing an answer to the question. For example, "So how many days would that be?" or "Which of the answers you gave is closer?" would both qualify as re-orienting probes. They are geared toward generating a response within the designated categories.

In contrast, elaborating probes are typical cognitive interviewing probes, designed to get information beyond the specific answer to the survey question. These would include probes such as "Tell me what you were thinking about while answering," "How would you describe your emotions in the last 30 days?", or "How would you answer the question if I used a different response scale [such as...]?" These probes are geared toward obtaining additional information beyond the codeable response.

Using transcripts, probes administered before the "best" answer were assigned one of the following codes:

Probe Codes

Code 0: No probes before the response.
Code 1: One re-orienting probe was used; no elaborating probes.
Code 2: More than one re-orienting probe was used; no elaborating probes.
Code 3: One elaborating probe was used; no re-orienting probes.
Code 4: More than one elaborating probe was used; no re-orienting probes.
Code 5: At least one of each type of probe (reorienting and elaborating) were used.

We ignored probes that came after the response because we wanted to see specifically how probe type was related to response precision. Probes administered after the response obviously could not have influenced it.

As an illustration, consider the following excerpt from a cognitive interview to be coded:

> Interviewer: Now thinking about your mental health, which includes stress, depression and problems with emotions, for how many days during the past thirty days was your mental health not good?
>
> Subject: Not good? My mental health has always been good.
>
> I: So would you say zero?

S: No, not necessarily, because I do get stressful at times.

I: What kind of stress?

S: Well, the number one thing is I'm very impatient... [subject discusses this for a while]

I: So if you had to come up with a number of days in the past thirty where your mental health was not good which could include stress, depression, emotional problems, or anything like that, what number would you pick?

S: I'd pick overall three to five days.

In this example, "three to five" was the most precise response given, which was coded as "four." It would receive a precision code of "1" because it was a narrow range. Both reorienting and elaborating probes were used, so the probing code would be "5."

Two NCHS staff members coded responses from the 18 cognitive interviews independently. Later, they were able to resolve all discrepancies using the coding rules.

Results

Our first analysis was to examine how "precision" varied across all of the "30 days" questions. Precision of response covered the entire range from very clear (36.3%) to uncodeable (23.0%).

Table 1 (right) shows how response precision varied across questions. For example, responses to Q14 (being healthy and full of energy) were relatively imprecise--41.2% of responses had major precision problems, largely because subjects distinguished between being "healthy" and "full of energy" and consequently had difficulties answering the question. In contrast, there were few precision problems with Q10, because most subjects did not experience such pain and answered "zero."

A somewhat surprising finding was that responses to Q11 (depression) were very imprecise compared to Q3 (mental health) responses-- respectively, 47.1% and 17.6% had major precision problems. Although Q3 and Q11 were closely related conceptually, the precision codes show that subjects were more apt to "discuss" their answers in Q11. The conceptual overlap might have confused subjects about the purpose of the latter question, who took the opportunity to explain their answers in detail-- thus creating response imprecision.

Table 2 (right) examines type of probing performed across all questions. Some elaborating probes were used prior to 31.0% of all responses (adding codes 3+4+5). Thus, this type of probing was common and might have encouraged some digressive behavior from subjects. Note also that reorienting probes were used prior to 30.3% of responses (adding codes 1+2+5). Interviewers may have encouraged elaboration, but a fair percentage of the time they also pushed their subjects to actually answer the question. In fact, interviewers used elaborating probing alone only 14% of the time.

Table 1: Precision of "30 day" subjective health questions
(Questions listed in order from lowest precision to highest)

During the past 30 days, how many days have you...	Percent of responses with **No/minor** precision problems (Codes 0,1)	**Major** precision problems (Codes 2,3)
Felt sad, blue, or depressed (Q11)	52.9	47.1
Felt healthy and full of energy (Q14)	58.8	41.2
Felt that you did not get enough rest or sleep (Q13)	64.7	32.3
Had poor physical or mental health that kept you from usual activities (Q4)	68.7	31.3
Physical health was not good (Q2)	70.6	29.4
Felt worried, tense, or anxious (Q12)	76.5	23.5
Pain made it hard for you for you to do usual activities (Q10)	76.5	23.5
Mental health was not good (Q3)	82.4	17.6

Table 2: Probing: all "30 days" subjective health measures

Probing	Percent of responses	
0 (None)	55.6	
1 (One reorienting)	11.1	
2 (Mult. reorienting)	2.2	
3 (One elaborating)	0.7	
4 (Mult. elaborating)	13.3	
5 (Reorient and elaborating)	17.0	
	100.0	(n=135)

Table 2 also shows that 55.6% of responses were accepted before any probing occurred. Excluding those cases leaves all responses that were preceded by probes (reorienting, elaborating, or both). In that subset, we explored how type of probing related to precision of response. We speculated that precise responses would be more likely to follow re-orienting probes, and imprecise responses would be more likely to follow elaborating probes. We explore this hypothesis in Table 3, below:

Table 3: Precision of response, by type of probes used before response

Precision	Elaborating probes before response	Re-orienting probes before response
0 (Precise)	4.8%	24.4%
1	21.4%	34.1%
2	14.3%	14.6%
3 (Uncodeable)	59.5%	26.8%
	100%	100%
	(n=42)	(n=41)

(Note: columns are not mutually exclusive-- reorienting and elaborating probes were used in 23 cases.)

The prediction was correct: when elaborating probes were used, about 5% of final responses were "precise" (as defined previously). In contrast, when re-orienting probes were used, 24% of final responses were precise. The proportion of uncodeable responses also varied according to probes used. While almost 60% of responses following elaborating probes were uncodeable, only 27% of responses following re-orienting probes had coding problems. It seems likely that re-orienting probes encourage subjects to answer the question in the specified format, while elaborating probes encourage discussion at the expense of answering within format.

Additional cognitive interviews and coding

Methods

To further explore the relationship between probing and responding, and to increase both the size and variety of our dataset, we conducted a second round of interviews, but in a different manner. We trained NCHS cognitive interviewers to use re-orienting probes exclusively and to avoid all elaborating probing. One of our conclusions from the first round of interviews was that many subjects had difficulty responding in the suggested format. We now wanted to know whether restricting the elaborating probes would reduce the response imprecision we had observed.

After administering the questions with the revised procedure, interviewers debriefed subjects, asking their overall impressions of the questions and any difficulties they had answering. The interviews were transcribed and coded using the same guidelines as before.

Results

We began our analysis by confirming that interviewers did probe as instructed, avoiding elaborating probes. Interviewers were successful at the task-- only one elaborating probe appeared in the transcripts of 20 interviews. Re-orienting probes were used prior to 30.4% of responses (almost identical to the incidence of re-orienting probing in the previous round), and no probing was done prior to 69.0% of responses (compared to 55.6% in the previous round).

We had argued previously that the questions were inherently difficult to answer, and hypothesized that many of the difficulties would persist regardless of interviewing style. Although elaborating probes may have influenced some initial subjects to digress while responding, other subjects actually rejected the premise of the question-- thus, their difficulties answering should have persisted even after elaborating probes were removed.

Interestingly, however, removing the elaborating probes also eliminated the great majority of imprecise responses. In Table 4, we compare precision codes for the first and second rounds of interviewing:

Table 4: Precision of responses compared across first and second interview rounds

Precision	Round 1	Round 2
0 (Precise)	36.3%	82.3%
1	32.6%	14.6%
2	8.1%	0.0%
3 (Uncodeable)	23.0%	3.2%
	100.0%	100.0%
	(n=135)	(n=158)

The table shows that when interviewers attempted to get a response through reorienting probes, they were generally successful at doing so. Not only were 82.3% of responses fully precise, but almost all of the remaining response imprecision was very minor (code 1). Only 3.2% of responses were uncodeable, as opposed to 23.0% of responses in the first round.

However, even though subjects gave "precise" answers, debriefings following the interviews revealed that many subjects had the same misgivings about their responses as they did in the first round. Several complained that "days" were inadequate to describe how often they felt unhealthy, depressed, and so on, or expressed other dissatisfaction with their answers. Other subjects admitted that they did not think very carefully about the actual number of days and answered with minimal thought.

Subjects gave numerical responses because they were asked to, and their answers did not reflect the reservations they professed to have about the accuracy and appropriateness of their responses.

Discussion

We set out to determine whether the response problems observed in our first round of cognitive interviews reflected "real" cognitive difficulties; or, whether cognitive interviews, by encouraging discussion and deviation from a standardized questionnaire, actually created response problems that would not appear in a survey interview. We conclude that our qualitative analyses of first-round interviews and second-round debriefings identified legitimate difficulties that some survey respondents face when answering these questions. Some laboratory subjects discussed their difficulty answering the questions in the requested format quite explicitly-- it seems unlikely that the nature of cognitive interviewing would actually create self-assessed difficulty.

Yet, we also observed that probing style is associated with precision of responses. What are we to make of the precision problems following elaborating probes, that did not follow re-orienting probes? We suggest that these problems were not created by cognitive interviewing-- rather, that cognitive interviewing gave us insight into problems that were suppressed in standardized interviews.

Subjects in the second round of interviews were given no indication that interviewers were interested in understanding problems with the questions. When subjects tried to explain their difficulties, interviewers followed their instructions and redirected them toward the requested response format. The subjects therefore responded to the request as best they could, even if the answers had little meaning. Later, when we opened discussion about the meaning of their answers, they were forthcoming in reporting difficulties. They had refrained from doing so earlier because interviewers rejected their attempts at qualifying or elaborating upon their answers.

In other words, subjects were most likely to discuss and explain their responses if they had qualifications and reservations about them. The most straightforward questions would presumably lead to the most straightforward answers. Thus, we argue that precision differences across questions do provide insight into their relative complexity.

Nevertheless, caution is in order when reporting results from cognitive interviews. Although we argue that relative precision problems are meaningful, it is also clear that elaborating probes are associated with increased discussion and digression. In other words, some subjects did not answer the questions as formatted simply because the interviewers did not ask them to. Just as the re-orienting probes may have suggested that digressions were unwanted, elaborating probes may have suggested that responding within the question format was unnecessary. Thus, it is possible to overstate the overall magnitude of cognitive problems if analysts are not careful.

Conclusion

As always, our study bears repeating under more ideal circumstances. Our samples were small, and precision problems may have been confounded with subject age and education levels (which we could not control for). The analyses in our second-round debriefings could have been more systematic as well.

Yet, the coding scheme and analyses here represent an important step toward formalizing methods for analyzing cognitive interviewing results. Comparing "response precision" across questions could be a useful way to evaluate the relative difficulties posed by each of them. There are also methodological applications-- the coding scheme provides opportunities to analyze the effect of probing on response precision, how probing style varies across interviewers and across studies, and so on. Although further research will be necessary, we think this is an important beginning.

References

Erickson, P., Wilson, R., and Shannon, I. (1995). "Years of Healthy Life." Healthy People 2000: Statistical Notes, No. 7. Hyattsville, MD: National Center for Health Statistics.

Fowler F.J., and Mangione, T. (1990). Standardized Survey Interviewing: Minimizing Interviewer-Related Error. Newbury Park, CA: Sage.

Idler, E. (1992). "Self-assessed Health and Mortality: A Review of Studies." In S. Maes, H. Leventhal, M. Johnston, eds., International Review of Health Psychology. New York: John Wiley and Sons.

Mangione, T.W., Fowler, F.J., and Louis, T.A. (1992). "Question Characteristics and Interviewer Effects." Journal of Official Statistics, 8, 293-307.

Schechter, S., ed. (1993). "Proceedings of the 1993 NCHS Conference on the Cognitive Aspects of Self-Reported Health Status." Cognitive Methods Staff Working Paper No. 10. Hyattsville, MD: National Center for Health Statistics.