Session 12

VALIDATION OF COGNITIVE QUESTIONNAIRE PRETESTING METHODS

Evaluating the Generalizability of Cognitive Interview Findings

Paul Beatty, Gordon Willis, and Susan Schechter
National Center for Health Statistics

I. Introduction

One purpose of cognitive interviewing is to confirm that terms are understood consistently; therefore, if we were to conduct a cognitive interview about cognitive interviewing, we might well ask "what does the term *cognitive interview* mean to you?" Since the term may take on different meanings to different people in different organizations, it is probably sensible to begin by deciding upon a working definition of the term. For our purposes, cognitive interviewing entails asking a participant to either "think out loud" while answering survey questions, or respond to probes about question interpretation and thought processes, or both. Survey researchers have generally accepted this technique as a legitimate, valid, effective means to quickly identify and correct questionnaire problems (Jobe and Mingay, 1991; Willis, Royston and Bercini, 1991; DeMaio and Rothgeb, 1996).

Is all the faith we have placed in the method justified? Based on anecdotal evidence, it seems to be. There is no shortage of examples of "bad" survey questions that were identified and improved based on cognitive interview findings. However, there have not been many systematic attempts to demonstrate that cognitive interviewing finds legitimate survey problems. It seems reasonable to ask: does cognitive interviewing find valid results that generalize to field settings?

That question needs to be answered in several steps. First, it is important to ask more basic questions about the goals of cognitive interviews-- what do we expect them to accomplish? After answering that, we can evaluate how well cognitive interviewing meets those goals. We will do that through reviewing some of our recent studies along those lines; in addition, we will outline our plans for continued research.

II. Addressing the problem of small and non-representative samples

One of the most common criticisms regarding cognitive interviewing concerns sample size and representativeness: how can one infer anything from a dozen interviews from a convenience sample? This criticism is closely related to our question about the goals of cognitive interviewing.

First, we should note that this lack of representativeness is very much by design. The idea is to select participants from particular age groups, those with certain health conditions, or whatever characteristics interest us the most. If, for example, we were interested in people over 65 with asthma, we could recruit them specifically. A random sample would be an inefficient approach to finding them-- in fact, several hundred respondents from the general population might tell us less than a dozen selected specifically from a target group of interest. Furthermore, questions nestled into complex skip patterns might rarely or never be administered during the test.

This efficiency is certainly important-- but how can we find a representative sample of cognitive problems that respondents will have when answering these questions, without a representative sample of the people who will be answering them? It is important to note that we are *not* claiming that we do this. Rather, we are proposing that cognitive interviewing does something much more modest-- provide *clues* regarding *potential* sources of survey error. In other words, rather than claiming that we found something that "will be a problem" when a survey is fielded, we would claim that it *might* be a problem, based on interactions with relevant survey participants.

There is also the issue of recognizing the difference between a legitimate questionnaire problem and an "odd case"-- which could be particularly difficult if we only interview a small sample of people. But actually, logic can usually distinguish the odd cases from likely problems. Cognitive interviews usually suggest not only what the problem is, but what aspect of the question creates the problem.

As an example, consider this question that was recently tested in our cognitive laboratory:

> During the past year, on average, on how many days did you drink alcoholic beverages, that is beer, wine, or liquor?
>
> ____ days
>
> a. per week
> b. per month
> c. per year

One laboratory subject expressed confusion. Probing during the cognitive interview revealed the source of this confusion: the question asks for a "number of days in the past year" and also an "average." It would make sense to ask the *average* number of days in a *typical* year-- or, it would make sense to ask about the *number of days* in the last year, dropping "average"-- but as it is, the question asks for both a "one year total" estimate, and an average over an unspecified time period. Apparently for this reason, our subject asked "do you want days last year, or what?"

This example has three vital characteristics: (1) the clue of a potential problem, (2) a reasonable explanation for the source of the problem, and (3) possible solutions. The identification of this potential problem is valuable because it is *logical* that respondents *could* stumble on this problem, and it can be avoided. Whether we discover this with one subject or fifty, the merit of the insight is really determined through a logical judgment. Thus, the interview was not a mechanism for proof, but rather an idea generator about potential problems.

III. Evaluating the value of cognitive interviewing "clues"

Traditionally, the debate about cognitive interview validity has focused on whether or not they uncover the "true" cognitive processes of respondents (Nisbett and Wilson, 1977). That is, are respondents *capable* of telling us how they figure out their answers? When they think out loud or respond to probes, are they telling us what is *really* happening in their minds, or is it actually a re-creation of their thought processes, which is therefore less valid?

It may not be critical to answer this question at this point. Participants provide us with clues that seem to have great value for discovering the sources of survey problems. The challenge for cognitive researchers is to demonstrate that these clues are actually useful to survey research, whether or not they reflect "true" cognitive processes. A more pressing concern is: what if the clues are wrong, or misleading, or otherwise steer us astray?

Until now, we have simply assumed that this is not the case. Researchers who conduct cognitive interviews have made several implicit assumptions about the value of cognitive interviewing clues. At face value, these assumptions seem reasonable, but they have generally not been challenged in a serious manner. The four major assumptions are as follows:

> <u>Assumption #1</u>: Cognitive interviewing finds problems that will carry over to actual surveys.

In other words, the findings of cognitive interviews are not "artifacts" deriving from the method. These interviews, we assume, tell us something that has practical utility.

> <u>Assumption #2</u>: The response process when answering questions in a cognitive laboratory is more or less the same as in a survey interview.

For example, question comprehension processes should be similar enough in a laboratory to a survey setting to be applicable. In other words, using laboratory findings is not comparing cognitive apples and oranges.

> <u>Assumption #3</u>: Cognitive interviewer behavior does not have an undue effect on the content of the interview.

Some interviewer variation is inevitable, of course. We are simply assuming that cognitive interviewer behavior does not <u>radically</u> alter the way subjects answer survey questions, or affect the basic value of our findings.

Assumption #4: The cognitive interviewing process is basically reliable-- if repeated, it would yield similar results.

That is, if one group of cognitive interviewers identified problems with a particular questionnaire, a different group of interviewers should find compatible (though probably not identical) results.

Some may feel that these assumptions have been made too lightly. Our own research has attempted to explore their veracity, focusing in particular on the first and second assumptions. Two studies described below investigate the assumptions through distinctive approaches.

IV. Two studies on the generalizability of cognitive interview findings

STUDY ONE (Willis and Schechter, 1996)

Anecdotally, if one compares survey questions before and after a round of cognitive interviews, it often seems obvious that the new question is "better" than the previous version. But what about actual survey data? Can we show that changes from cognitive interviews have positively impacted actual survey data?

Consider the following survey question, designed to measure time spent performing strenuous physical activity:

On a typical day, how much time do you spend doing strenuous physical activities such as lifting, pushing, or pulling? (hand card)

a. None
b. Less than 1 hour
c. 1-4 hours
d. 5 or more hours

When tested in a cognitive laboratory, many subjects selected the "1-4 hours" response. When they were probed, however, they often admitted that they worked in offices and performed typical office tasks-- not what we would define as strenuous.

The question seemed to produce a bias-- reporting "none" clearly makes one appear sedentary. Given the available response options, it was much more desirable to report some level of activity than absolutely none.

Our clue of a potential problem was the preponderance of "1-4 hours" responses, which disagreed with probe responses. Our explanation of this discrepancy is the undesirability of appearing to be completely sedentary. A possible solution, then, would be to provide respondents with a more socially desirable "out."

A first step was to draft a question that eliminated this problem. An alternative version was written with this additional screener question:

> On a typical day do you spend any time doing strenuous activities such as lifting, pushing, or pulling? (Yes/No)

A "no" response counted as zero; only subjects who answered "yes" received the original frequency question. When we tested this new version, many subjects were perfectly willing to respond "no," sometimes adding comments such as "I work at a computer all day." The screener question may be an improvement because it presents a balanced choice of equally legitimate responses: some people do strenuous activities and others do not. The previous question implied a continuum ranging from sedentary to vigorously active. Respondents' desire not to appear at the low end of this continuum might have influenced their responses.

The next logical question is: does this new version actually make a difference in the field, improving the accuracy of statistics? To test that, both versions were administered in a split ballot-- one with the screener and one without. The following results were observed in a relatively small field pretest, and repeated in a study on the health of women of child-bearing age:

Table 1: Field Pretest Results: Versions Before and After Cognitive Interview Modifications

	Test 1: NHIS Field Pretest		Test 2: Women's Health Study	
Hours	Ver 1	Ver 2	Ver 1	Ver 2
0	32%	72%	4%	49%
<1	32%	18%	42%	16%
1-4	35%	10%	50%	27%
5+	0%	0%	4%	8%
	----	----	----	----
	n=37	n=39	n=93	n=94

As predicted, the distributions of answers are quite different, with many more respondents falling into the "zero" category when a yes/no screener is used (Version 2, in both tests). We presume that the Version 2 responses are more accurate. We do not know that for certain, but given the apparent tendency to overestimate time spent performing strenuous activity, a good case can be made for this conclusion.

This process was repeated using other survey questions over several different split ballot experiments. The results generally matched these findings: hypotheses from cognitive interviews were borne out by field data. This suggests that cognitive interview findings were relevant and applicable to a field setting.

STUDY TWO (Beatty, Schechter, and Whitaker, 1996)

This study was a follow up to cognitive interviews about subjective health assessments. Questions were based on feelings in the last 30 days-- for example, "During the past 30 days, how many days has your physical health been not good?" The questions called for numeric responses between 0 and 30 days, but many subjects had difficulty providing them. Some provided general answers, such as "I feel that way a lot"; others objected to the premise of the question, arguing that "I can't put it in days."

It seemed clear that the questions had problems, since a large proportion of responses were not given in the expected format. However, the survey sponsors had administered these questions in the field with no reports of trouble from interviewers, and very low item nonresponse. Their alternative theory was that the conversational tone and frequent probing in cognitive interviews actually *created* the *appearance* of problems.

The purpose of our study was to examine the relationship between probing style and subjects' answers. Using transcripts of cognitive interviews, we first coded each subject's response to each survey question, or the statement that most clearly resembled a legitimate response.

Second, we developed a code for how closely this response conformed to the expected response format-- that is, a number between 0 and 30. We labeled this "precision," recording it on a scale from 0 to 3 as follows:

> Code 0: The response was clear, requiring virtually no rounding, judgment, or interpretation from a coder. Example: "Four days."
>
> Code 1: The response required minimal interpretation from a coder, such as a moderately qualified answer, or answers given in a narrow range. Examples: "Probably every day," "Six or eight days."
>
> Code 2: The response required considerable interpretation from a coder, such as broad ranges. Examples: "Six to ten days," "More than 15 days."
>
> Code 3: The response could not be coded in the expected format. Examples: "I can't put it in days," "For a while I was in horrible pain," etc.

Third, we coded the type of probes that preceded each response. We distinguished between "re-orienting" and "elaborating" probes. *Reorienting probes* encourage subjects to re-focus on answering the survey question, such as "So how many days out of 30 is that?" *Elaborating probes* are more typical of cognitive interviews, designed to get information beyond the answer to the survey question-- for example, "Tell me what you were thinking about while answering" (which encourages the subject to discuss the answer).

Our analytic goal was to determine how probing style was related to response precision. We found that probing style had considerable influence. When re-orienting probes preceded responses, 24% of responses were "precise"; when elaborating probes preceded responses, only 5% of responses were precise. Similarly, the percentage of "uncodeable" responses changed considerably depending on probing style: 60% were uncodeable following elaborating probes, whereas only 27% were uncodeable following reorienting probes. These results appear in Table 2, below:

Table 2: Response precision, by types of probes preceding 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%
	-------	-------
	n=42	n=41

(Table excludes cases in which no probing preceded response. Because re-orienting probes and elaborating probes were used in 23 cases, columns are not mutually exclusive.)

Next, we conducted additional interviews, this time training interviewers to use only re-orienting probes. This was done to evaluate whether response imprecision could be reduced by curtailing interviewer behavior that led to increased discussion. Interviewers discussed the meaning of subjects' answers only at the end of the interview session, during a debriefing. A comparison of results from the first and second round of interviews appears below:

Table 3: Precision of responses, compared across 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%
	-------	-------
	n=135	n=158

(Table includes all responses, whether preceded by probes or not.)

In the second round of interviews, 82% of responses were precise, and only 3% were uncodeable. At first, it might seem that the charge against cognitive interview findings was correct-- if one removes conversational probes, subjects' responses are much more straightforward. However, post-interview debriefings revealed that subjects still had many of the same misgivings about answering the questions that they had in the earlier cognitive interviews-- they were reluctant to answer in terms of days, or felt their answers were inaccurate. In the later round, however, interviewers denied subjects the opportunity to *express* uncertainty about their answers. If subjects tried to explain or qualify their responses, the interviewer asked them to respond numerically. Thus, we suggest that cognitive interviewing does not *create* the *appearance* of problems, but rather that conventional interviewing *suppresses* the *expression* of response difficulties.

The fact that some subjects deviate from question format in cognitive interviews, in and of itself, is not particularly illuminating-- interviewers *ask* them to do this. However, the *amount* of deviation from format, which varies across questions, may provide a useful measure of relative difficulty answering the questions. A greater desire to discuss the nuances of answers is probably informative. Nevertheless, analyses needs to be performed with sensitivity to the fact that a cognitive interview is quite different from a survey interview.

V. Future directions for empirical work

Several of the assumptions mentioned earlier-- regarding cognitive interviewer effects, and reliability of conclusions-- have not yet been addressed. We have initiated several studies that explore those assumptions, however, and expect to present data in the near future.

Staff at NCHS recently constructed a "methodological questionnaire" to serve as the basis for additional research. The questionnaire was constructed from drafts of questions from various surveys, but the methodological questionnaire will not actually be fielded. It will therefore be possible to explore hypotheses by maintaining complete control of questionnaire content, question wordings, and so on. ("Real survey" pressures often make it difficult to implement this type of methodological work).

NCHS staff conducted 40 cognitive interviews using this questionnaire, which will serve several purposes. First, cognitive interviewer behavior will be coded: we will explore how much interviewer behavior varies, and in what manner. As of this writing, it is too early to tell exactly how much individual interviewers' styles differ, but it is clear that there is a wide variety of activity during cognitive interviews. A preliminary taxonomy of cognitive interviewer behavior distinguishes between numerous types of probes (probes about thought-processes, question interpretation, question difficulty, and probes for information beyond the scope of the survey question); types of feedback (feedback on subject performance, and feedback on content of responses); and other remarks (transitional statements, confirmation of subject responses, and so on).

In addition to coding what interviewers do, we will investigate what interviewers *conclude* about the nature and extent of questionnaire problems. An important component of reliability assessment is determining whether interviewers reach the same conclusions about problems in a particular questionnaire. Also, a contractor will conduct 60 cognitive interviews using the same questionnaire. That will enable comparison of how two independent groups go about evaluating a questionnaire, and comparison of the conclusions they reach.

Finally, the analysis will extend to other pretesting methods. Twenty questionnaire designers have provided "expert reviews," of the methodological questionnaire; also, field pretest interviews were behavior-coded (see Fowler and Cannell, 1996). Comparing the results of these appraisals should provide a sense for how the methods complement each other, rather than demonstrating which techniques are "best."

In summary, much work remains in investigating the generalizability of cognitive interview findings. However, we also have good preliminary indications that cognitive interviews are effective clue-finders that greatly help questionnaire designers perform their jobs. We look forward to sharing more results of our evaluations in the future.

REFERENCES

Beatty, P., Schechter, S., and Whitaker, K. (1996). "Evaluating Subjective Health Questions: Cognitive and Methodological Investigations." Proceedings of the Section on Survey Research Methods, American Statistical Association, in press.

DeMaio, T.J. and Rothgeb, J.M. (1996). "Cognitive Interviewing Techniques: In the Lab and in the Field." In Schwarz, N., and Sudman, S., eds., Answering Questions: Methodology for Determining Cognitive and Communicative Processes in Survey Research. New York: Jossey-Bass.

Fowler, F.J., and Cannell, C.F. (1996). "Using Behavior Coding to Identify Cognitive Problems with Survey Questions." In Schwarz, N., and Sudman, S., eds., Answering Questions: Methodology for Determining Cognitive and Communicative Processes in Survey Research. New York: Jossey-Bass.

Jobe, J.B., and Mingay, D.J. (1991). "Cognition and Survey Measurement: History and Overview." Applied Cognitive Psychology, 5, 175-192.

Nisbett, R.E., and Wilson, T.E. (1977). "Telling More Than We Can Know: Verbal Reports on Mental Processes." Psychological Review, 84, 231-259.

Willis, G.B., Royston, P.N., and Bercini, D. (1991). "The Use of Verbal Report Methods in the Applied Cognitive Laboratory." Applied Cognitive Psychology, 5, 251-267.

Willis, G., and Schechter, S. (1996, July). "Evaluation of Cognitive Interviewing Techniques: Do the Results Generalise to the Field? Paper presented at the Fourth International Sociological Association Conference on Social Science Methodology, Essex, England.

PREDICTING TEST-RETEST RELIABILITY FROM BEHAVIOR CODING[1]

Jennifer C. Hess, Center for Survey Methods Research, Bureau of the Census,
Washington, DC 20233

Eleanor Singer, Survey Research Center, University of Michigan

[1] The authors wish to acknowledge the contribution of John Bushery, Richard McGuinness and their colleagues in the Demographic Statistical Methods Division of the Census Bureau who calculated the reliability data used in this paper.

Introduction

In attempting to move questionnaire design from art to science, researchers use different evaluation techniques to help determine how well questions are working. Techniques such as behavior coding, respondent debriefing, interviewer debriefing, cognitive interviewing, and nonresponse analysis all provide information to help the questionnaire designer assess whether respondents understand questions as intended and whether they are able to provide adequate answers to them. In 1994, Presser and Blair evaluated some of these methods, concluding that behavior coding provided more reliable diagnoses of question difficulties than conventional pretests involving a small number of interviewers followed by an interviewer debriefing.

However, with the possible exception of some types of respondent debriefing questions, these techniques do not actually measure question reliability. Reliability data, such as those that could be obtained in a test-retest experiment (reinterview), are rarely collected as part of pretest activities because they are time-consuming, labor intensive and very costly to collect. Of course, the goal of good questionnaire design is to produce reliable and valid information, not simply questions that are easy for respondents to answer. But it is assumed that questions that pass the screen of the questionnaire evaluation techniques described above are also more likely to produce data that are reliable and valid.

How well do question evaluation techniques in fact predict reliability and validity? Data reported by Belli and Lepkowski (1995) suggest that interviewer behaviors have little predictive value for response accuracy, though respondent behaviors are somewhat more predictive of response accuracy. Recently, the U.S. Department of Agriculture's Food and Consumer Service fielded a new survey, designed to measure the subjective experience of hunger in the United States. This survey provided an opportunity to examine how well some traditional question evaluation techniques predict test-retest reliability. The Census Bureau was asked to help develop the questionnaire, using some of the evaluation methods listed above. In addition, a reinterview was conducted with a sample of households following the survey. In this paper, we use behavior coding data to predict how reliably questions are answered, as measured by an index of inconsistency developed by the Census Bureau.

Methods

Sample

The Food Security Supplement to the Current Population Survey (CPS) was conducted from April 16-25, 1995 on a nationally representative sample of approximately 54,000 interviewed households. Respondents were asked both the CPS labor force questions and the Food Security Supplement questions. The response rate for the CPS was 92.9 percent and for the supplement was 85.4 percent. Approximately 90 percent of the cases were conducted in the field using computer assisted personal interviewing (includes both personal visit interviews and telephone interviews from field representatives' homes) and 10 percent were conducted at the Census Bureau's centralized telephone facilities using computer assisted telephone interviewing.

Approximately 34 percent of the households in the sample were "low income," which, for the purposes of this study, is defined as at or below 185 percent of the poverty level.[2] Three-quarters of the sample households were urban and one-quarter rural. Approximately 85 percent of the households were White, 10 percent were Black, and 6 percent were Hispanic (could be of any race).[3]

The questionnaire included five different sections: food expenditures, program participation, food sufficiency, coping mechanisms and food scarcity, and concern about food sufficiency.[4] Food expenditures were asked of all households. These questions collect information on the actual amount the household spent for food last week and the usual amount the household spends on food per week. The program participation section asks about food stamp recipiency and participation in other government and private programs that provide food, such as the school lunch program and WIC. The food sufficiency section contains questions used to assess whether respondents clearly have enough to eat or whether there are times when their resources are strained and they have difficulty providing themselves or their families with a nutritionally adequate diet. These questions are used to screen respondents either into or out of the remainder of the questionnaire. The coping mechanism and food scarcity section measures the extent of food insecurity in the household as do the questions in the section on concern about food sufficiency.

Behavior Coding

Behavior coding is the systematic coding of the interactions between an interviewer and a respondent (Cannell, Lawson, and Hausser, 1975; Cannell et al., 1989). Interviewers at the Census Bureau's Hagerstown and Tucson Telephone Centers tape recorded a total of 147 cases of which 136 were subsequently behavior coded. (Eleven cases were not used because permission to record the interview was not on the tape.) We used a quota sample for behavior coding, not a random sample. The telephone centers were instructed to tape record interviews with the first 75 low income households.

We coded the first exchange between the interviewer and the respondent for each question. Coders assigned one interviewer code and up to two respondent codes per question. (Two respondent codes were most often assigned when the respondent interrupts the question reading to provide an answer. Thus, one of the codes is a "break-in" and the other may be any of

[2] Our measure of "185 percent of poverty" in this survey is based on family size and family income. The measure, however, is rather imprecise, because the only measure of family income in the CPS is based on a single question about family income in the previous calendar year and is a categorical variable composed of income ranges.

[3] Race of the household is measured by the unweighted race of the reference person. The reference person is the first person listed on the household roster and is the name of the person or one of the persons who owns or rents the house/apartment.

[4] Contact the authors for a copy of the questionnaire.

the remaining respondent codes.) Four experienced coders from the Hagerstown Telephone Center behavior coded the tapes. (See Appendix A for a description of interviewer and respondent behavior codes.)

To assess coder reliability, each coder was asked to complete the same five cases (in addition to the regular workload). The coders averaged 87 percent agreement on interviewer codes, 92 percent agreement on at least one of the two respondent codes, and 83 percent agreement on both respondent codes. The kappa statistics, which take into account the probability that two coders will agree on a code by chance, ranged from .68 to .80 for between coder agreement on interviewer codes, .74 to .93 on at least one of the two respondent codes, and .55 to .84 on both respondent codes. Kappa values above .75 represent excellent agreement and values from .40 to .75 represent fair to good agreement beyond chance (Fleiss, 1981). Thus, our statistics indicate fair to excellent agreement between coders.

An evaluation of the supplement questionnaire based on behavior coding data indicated that the food expenditures section caused the most problems of any section (see Table 1). Eighty-three percent (N=18 questions) of the questions in this section were flagged as problematic by behavior coding. Approximately 60 percent of the questions in the food sufficiency section (N=10 questions) and the concern about food sufficiency section (N=6 questions) were problematic. The remaining two sections, the program participation section and coping mechanisms and food scarcity section, caused fewer problems. Twenty percent of the questions in the program participation section (N=10 questions) and 28 percent of the questions in the coping mechanisms and food scarcity section (N=36 questions) were problematic. However, 15 of the 36 questions in the latter had less than 7 responses. When these cases are excluded, the percentage of problematic cases in this section drops to 10 percent. (Results are for both categorical and continuous variables.)

Table 1. **Percentage of Problematic Supplement Questions By Section**

Section	Question numbers	Total number of questions in section	Percent problematic questions
Food expenditures	1-8	18	83 percent
Program participation	9-9G	10	20 percent
Food sufficiency	11A-16	10	60 percent
Coping mechanisms and food scarcity	17-52	36 21	28 percent 10 percent (excluding questions with less than 7 cases)
Concern about food sufficiency	53-58	6	67 percent

Reinterview

The Food Security Supplement reinterview was conducted from April 17-29, 1995 by CPS supervisors, senior field representatives, and interviewers. Approximately 90 percent of the reinterviews were conducted within 7 days of the original interview, but in some cases, there was up to a 10 day lag.[5] The reinterview was conducted on a nationally representative sample of 1,827 with a response rate of 63.6 percent (1,162 completed interviews). The reinterview was conducted with the same respondent who had answered the original survey. The sample was split between households with family incomes at or below 185 percent of the poverty level and those with family incomes above 185 percent of the poverty level; 929 reinterviews were conducted with the former group and 233 with the latter. This sample was drawn in order to test two important features of the questionnaire: 1) the reliability of the screening questions that determined whether a respondent was asked the remaining questions that measure degree of food

[5] The number of days between the original interview and the reinterview may account for some of the unreliability measured in the index of inconsistency.

insecurity, and 2) the reliability of the questions on food insecurity. Because of cost constraints, most reinterviews were conducted by telephone.[6]

The ma or ob ective of the reinterview was to measure response variance, that is, to determine the degree of inconsistency between the original survey answer and the reinterview answer. The reinterview data contain several measures of response variance. We will use the index of inconsistency in this paper. This is a relative measure of response variance that estimates the ratio of response variance to total variance for each question. In general, an index of less than 20 indicates that response variance is low; an index between 20 and 50 indicates that response variance is moderate; and one over 50 indicates that response variance is high (McGuinness, forthcoming).[7]

Table 2 shows the mean and median index of inconsistency by section of the questionnaire for categorical variables.

[6] Approximately 35 percent of the cases in the original interview were conducted by personal visit and 65 percent were conducted by telephone either from the field representatives' homes or from a centralized telephone facility. Personal visit interviews are primarily month-in-sample one and five cases, thatis, those cases that are in sample for the first time or those cases that are returning to the sample after a four-month hiatus. Thus, as much as 35 percent of the sample may be sub ect to a mode effect and some of the variation in the index may be due to a mode effect. Based on differences in survey data resulting from personal visit vs. telephone mode effects, the consensus at the Census Bureau is that these differences are quite small and would contribute little to the variation in the index.

[7] The index of inconsistency is the simple response variance divided by the total variance. Computationally it is the proportion who change answers between the original interview and the reinterview divided by $(P1*Q2) + (P2*Q1)$
where $P1$= the proportion in category from the original interview
where $Q1$= the proportion not in category from the original interview
where $P2$= the proportion in category from the reinterview
where $Q2$= the proportion not in category from the reinterview

Table 2. Mean and Median Index of Inconsistency for Each Section of the Questionnaire

Section	Mean	Median
Food expenditures	52	52
Program participation	25	19
Food sufficiency	46	47
Coping mechanisms and food scarcity	44	44
Concern about food sufficiency	53	52

In general, these data indicate that four of the five sections of the questionnaire are producing moderately to highly unreliable data, with the notable exception of the program participation section.

Results

Behavior coding guidelines generally state that a question is considered problematic if less than 85 percent of the time interviewers read questions exactly as written or with only slight changes that do not affect question meaning, or if less than 85 percent of respondents give adequate or qualified answers to the question (Oksenberg, et al., 1991). Our analysis is limited to questions with a minimum of 7 cases in the behavior coding data.

We compare the results of behavior coding to those of the reinterview data at the question level. That is, we compare the diagnostic utility of behavior coding in predicting which questions will yield reliable data on reinterview. We do not have matching datasets at the level of the individual respondent, since the samples for behavior coding and for reinterview were drawn independently.

The questionnaire contained 75 questions, plus one split ballot item. There were 55 categorical questions of the "mark one answer" type, 20 continuous questions, and one question that was a "mark all that apply" type. This question had 5 possible responses and is treated as five separate questions in this analysis.

We were unable to use all questions in our analysis for two reasons. First, 3 questions were excluded because they had less than seven cases in the behavior coding data, 16 were excluded because of an unreliable index of inconsistency, and 15 were excluded because of both reasons. In most cases, the index was unreliable because the characteristic of interest is rare in the population and too few respondents were reinterviewed to provide reliable estimates. Thus, 46 questions were available for analysis. Second, because the index of inconsistency is calculated differently for categorical and continuous variables and the small number (N=9) of continuous variables made it impossible to carry out separate analyses for them, we decided to restrict the analysis to categorical variables.[8] The analysis in this paper is, therefore, restricted to the 37 categorical variables for which we have reliable behavior coding and reinterview data.

Table 3 shows the three models we used to test the predictive utility of the behavior coding data. The dependent variable is the index of inconsistency, a continuous variable that, in theory, ranges from 0 to 100.[9] All three models include the two independent variables for the behavior coding data. These variables are percentages ranging from 0 to 100. The respondent behavior code is the percentage of times respondents provided an adequate or qualified answer to the question. The interviewer behavior code is the percentage of times interviewers read the question exactly as worded or with only slight changes that didn't affect question meaning. In addition to the two behavior coding variables, Model 2 includes three dummy variables representing the sections of the questionnaire. Although the questionnaire contains five sections, two of them--food sufficiency and coping mechanisms and food scarcity--are similar in content and are differentiated in the questionnaire only because the former is used to screen respondents either into or out of the remainder of the questions. Accordingly, these two sections were collapsed for the present analysis. The omitted category is the concern about food sufficiency section. The sections of the questionnaire were included in the model since we knew from both the behavior coding data and the reinterview data that not all of the sections performed equally well. Model 3 includes interactions between the respondent behavior code and the sections of the questionnaire.

[8] We did, in fact, run a general linear model separately for the numeric data. Because of sample size only the behavior coding variables could be used to predict the index of inconsistency. Neither the respondent nor the interviewer behavior coding variable was significant.

[9] It is possible for the index of inconsistency to be greater than 100 if the number of observed agreements is less than chance. See Perkins, 1971 for details.

Table 3. General Linear Models for Predicting the Index of Inconsistency (Standard errors in parentheses)

Variable	Model 1 Parameter Estimate	Model 2 Parameter Estimate	Model 3 Parameter Estimate
Intercept	155.7 (57.1)	76.7 (48.0)	-4.9 (69.0)
Respondent behavior code (RBC)	-0.6* (0.2)	-0.5* (0.2)	0.3 (0.8)
Interviewer behavior code	-0.6 (0.6)	0.2 (0.5)	0.4 (0.4)
Food expenditure (Food)		15.3* (6.8)	268.7** (75.5)
Program participation (Program)		-26.5** (7.7)	201.1* (91.0)
Food sufficiency, coping mechanisms and food scarcity (Coping)		-7.5 (6.5)	34.5 (67.4)
RBC*Food			-3.1** (0.9)
RBC*Program			-2.7* (1.1)
RBC*Coping			-0.5 (0.8)
Model r-square	0.20*	0.61**	0.83**
Degrees of freedom	2	5	8
N	37	37	37

**: p<.01 *: p<.05

Model 1 indicates that the respondent behavior code significantly predicts the index of inconsistency. The sign of the parameter estimate is in the expected direction; that is, as the percentage of respondents who provide adequate or qualified answers increases, the index of inconsistency decreases, indicating lower response variance (higher reliability). Interviewer behavior, however, is not significantly related to the index of inconsistency. These results are similar to those found by Belli and Lepkowski (1995).

The lack of association between interviewer behaviors and question reliability is not surprising. Very few questions were identified as problematic based on interviewer reading errors. Interviewer and respondent behavior coding data for the 37 questions of interest is included in Appendix B. Using the 85 percent threshold for determining whether a question was problematic indicates that only 2 of the 37 questions would be considered problematic based on interviewer reading errors. These same two questions plus an additional 12 were determined to be problematic based on respondent codes.

Model 2 includes the dummy variables for the sections of the questionnaire. (The omitted category is the concern about food sufficiency section.) The two behavior coding variables perform similarly in Model 2 as in Model 1. The parameter estimate for the respondent behavior code remains significant and inversely correlated with the dependent variable, and the interviewer behavior codes are not significant. Addition of the three dummy variables contributed significantly to the model R^2. The results indicate that questions in the food expenditures section were associated with higher levels of response variance (more unreliable) and questions in the program participation section were associated with lower levels of response variance (more reliable) than questions in the omitted section. These findings are consistent with the behavior coding data. Using the 85 percent threshold, five of the seven questions from the food expenditures section of the questionnaire that are included in this analysis were identified as problematic based on respondent codes, whereas only one of the five questions in the program participation section of the questionnaire was identified as problematic based on respondent behavior codes.

Model 3 includes interaction terms between the respondent behavior coding data and the section of the questionnaire. The increase in the R^2 value between Model 2 and Model 3 is significant, indicating that the interaction terms contribute significantly to the amount of variation explained in the dependent variable. The interaction terms indicate that the ability of the respondent code to predict the dependent variable is contingent on the section of the questionnaire. The respondent code is significantly associated with the index of inconsistency only in the food expenditures and program participation sections. The respondent code was not significantly associated with the index in the combined food sufficiency/coping mechanisms sections. Appendix B shows that questions in this section performed well according to respondent behavior coding data, but produced relatively unreliable data according to the index. And respondent behavior coding data for the concern about food sufficiency section were mixed, whereas the index indicated the questions were uniformly unreliable.

Discussion

Why does behavior coding predict reliability of response in some sections of the questionnaire but not in others? On a purely statistical level, the lack of variation in the independent variable (respondent behavior code) in the combined food sufficiency/coping mechanisms and food scarcity section or the dependent variable in the concern about food sufficiency section is probably sufficient to preclude a significant effect of the behavior coding variable in those sections. The more interesting question, however, has to do with how these sections of the questionnaire differ from the others either in terms of the content of the questions, or in terms of their structure.

One way in which these sections differ from the others is that questions in the food expenditures and program participation sections are of a more clearly factual nature than those in other sections. The food expenditure section includes questions on whether the respondent shopped at various locations (supermarkets and grocery stores, other stores, and restaurants), whether they included all purchases regardless of how they paid for them, how often they shop at supermarkets and grocery stores, and whether the amount they spent last week is the usual amount they spend per week. The program participation questions ask about food stamp recipiency, and participation in other food-related programs such as the school lunch and breakfast program and WIC. The remainder of the questionnaire measures the extent of food insecurity in the household. Questions in the concern about food sufficiency section are intended to measure a more subjective dimension of food insecurity than questions in the food sufficiency/coping mechanisms section. However, one could argue that several of the questions in the latter section are subjective as well (see particularly questions 32, 33, 35, 38 in the questionnaire).

A second difference is the reference period used in the questions. The food expenditure questions ask about shopping "last week," and the program participation questions ask about the "last 30 days." Questions in the other sections of the questionnaire have either long or nonexistent reference periods. Out of 25 questions, 19 ask about the "past 12 months," 3 ask about the "past 30 days," and 3 mention no reference period. Perhaps the long reference period results in respondents using recall strategies that produce unreliable data. Unfortunately, the data collected in this study do not allow us to investigate these hypotheses further.

Conclusions

For a long time, researchers have used behavior coding as a guide in questionnaire development, on the assumption that when respondents and interviewers are able to ask and answer questions without difficulty, the quality of the information obtained will be better. This assumption has been based largely on faith rather than empirical evidence. The findings in the present paper provide empirical support for the assumption, but they also appear to qualify it in some important respects. First, interviewer behavior coding has no predictive value for reliability, at least in a study such as this one, where interviewers perform at a uniformly high level. These findings might well differ in studies with greater variability among interviewers. Second, respondent behavior coding data do not appear to predict all types of reliability equally well. Prediction appears to be better for factual questions, and/or for questions with a relatively short recall period. When these conditions are not met, people may be able to answer the questions--and, therefore, behavior coding data may give no indication of difficulty--but the reliability of answers (and, hence, their validity) may nevertheless be low. Clearly, more research is needed into the characteristics of questions for which behavior coding is a valid predictor of test-retest reliability.

In concluding, we would also like to draw attention to some limitations of our data that make us offer these conclusions with a great deal of caution. First, our results are not generalizable. The behavior coding data were not drawn from a random sample of households. They are primarily low income households from the first 75 low income cases interviewed at two of the Census Bureau's centralized telephone facilities. Moreover, the samples for behavior coding and reinterview are different. The reinterview sample is nationally representative, but was oversampled for low income households and suffers from a low response rate (64 percent). Second, because of differences in sample design and sample size, our analysis is at the question level, not the individual level. This analysis would be more precise if we had matched individual level data. Third, the number and type of questions contained in this analysis are very small and the questions are not constructed to deliberately vary either content or structure. Although there were 80 questions in the original survey, we were only able to include 37 questions in our model. Questions were excluded primarily because the characteristic of interest is so rare in the population that the reinterview sample was too small to produce a reliable index of inconsistency. Moreover, we had to exclude continuous variables from the model because the index is calculated differently for categorical and continuous variables and there were too few continuous variables to produce a separate model. Fourth, although approximately 90 percent of the reinterviews were done within seven days of the original interview, the elapsed time between the original interview and the reinterview may account for some of the unreliability measured in the index of inconsistency, and the impact of the elapsed time may not affect all questions equally. It is possible that questions with shorter reference periods, such as those asking about behaviors occurring "last week" in the food expenditures section, were more adversely affected by the elapsed time between interviews than questions with longer reference periods. Respondents may be answering the food expenditure questions about a different week during the reinterview than in the original interview.[10] Thus, the index may not be speaking to reliability in the food expenditure questions and may be correlating with the behavior coding data for the wrong reason. Given these caveats, our results suggest that respondent behavior coding is associated with one measure of reliability; however, its ability to predict reliability in our study was not uniform throughout the questionnaire. Additional research is needed to understand the characteristics of questions for which behavior coding is a valid indicator of reliability and those for which it is not.

[10] The questionnaire was modified during the reinterview to prompt respondents to report for the week before the original interview.

REFERENCES

Cannell, C., Lawson, S. and Hausser, D. (1975). *A Technique for Evaluating Interviewer Performance.* Ann Arbor, University of Michigan.

Cannell, C., Oksenberg, L., Kalton, G., Bischoping, K., and Fowler, F. (1989). "New Techniques for Pretesting Survey Questions." Report submitted to the National Center for Health Statistics. Ann Arbor, Survey Research Center, University of Michigan.

Hess, J., Singer, E., Ciochetto, S., "Evaluation of the April 1995 Food Security Supplement to the Current Population Survey." Report prepared by the U.S. Bureau of the Census, Center for Survey Methods Research for the U.S. Department of Agriculture Food and Consumer Service, Alexandria, VA, January 26, 1996.

McGuinness, R., "Reinterview Report: Response Variance in the 1995 Food Security Supplement." Report prepared by the U.S. Bureau of the Census, Demographic Statistical Methods Division/QAEB for the U.S. Department of Agriculture Food and Consumer Service, Alexandria, VA, forthcoming.

Perkins, Walter M., "On the Index of Inconsistency." Memo for the Center for Research and Measurement Methods, U.S. Bureau of the Census, 1971.

Presser, S., and Blair, J. "Survey Pretesting: Do Different Methods Produce Different Results?" *Sociological Methodology*, Vol. 2, No. 12, pp. 73-104.

APPENDIX A

Interviewer and Respondent Behavior Codes[1]

Interviewer Codes

E: Exact question reading
S: Slight change in question reading
M: Major change in question reading
V: Verify
O: Other

Respondent Codes

A: Adequate answer
Q: Qualified answer
I: Inadequate answer
C: Requests clarification
B: Break-in or interruption of question reading
D: Don't know answer to question
R: Refuses to answer question
O: Other

[1] For a fuller description of the behavior codes, see Hess, Singer, and Ciochetto (1996), Attachment E.

APPENDIX B

Interviewer and Respondent Behavior Coding Data and the Index of Inconsistency for Questions Included in the Regression Models

Question Number	Interviewer Behavior[2]	Respondent Behavior[3]	Index of Inconsistency
Food Expenditures			
1	97.1	86.0	68.7
1A1	93.2	67.0	97.5
1C	77.9	83.1	47.6
2	97.8	82.4	55.1
3	99.3	90.4	33.9
4	91.4	76.4	100.0
6	95.2	81.8	79.8
Program Participation			
9	96.6	92.1	9.6
9C	100.0	92.9	19.4
9D	93.3	86.7	32.0
9E	96.9	78.1	47.1
9F	95.4	88.4	15.1
Food Sufficiency			
11A	100.0	52.0	46.8
11	98.2	83.3	47.1
12	99.0	61.2	52.3
15	97.0	85.2	42.1
16	97.0	94.0	41.3

[2] Percent exact or slight readings.

[3] Percent adequate or qualified answers.

Question Number	Interviewer Behavior	Respondent Behavior	Index of Inconsistency
Coping Mechanisms and Food Scarcity			
17	95.8	95.8	43.4
18	96.8	93.6	35.0
19	97.5	100.0	35.9
20	98.9	87.1	43.5
21	100.0	93.6	35.6
22	98.9	92.5	39.5
24	96.8	96.8	41.0
25	92.9	50.0	56.1
26	100.0	92.9	46.2
28	97.9	94.6	54.2
32	100.0	94.6	36.0
33	100.0	93.9	49.2
35	98.9	91.4	47.4
38	98.9	98.9	48.2
Concern About Food Sufficiency			
53	81.7	77.4	54.1
54	90.3	82.6	48.7
55	92.5	79.6	54.2
56	95.0	87.5	50.1
57	97.5	85.0	65.9
58	97.5	75.0	48.0

A Discussion of Cutting Edge Research in Cognitive Interviewing and Behavior Coding

Robert F. Belli
University of Michigan

I would like to thank the speakers for presenting insightful papers that illustrate cutting edge research in cognitive interviewing and behavior coding. Before offering comments on this research, I would first like to review the purposes and problems of cognitive interviewing and behavior coding in order to frame my comments.

Cognitive Interviewing

The purpose of cognitive interviewing is to precisely assess the cognitive processes that affect the quality of survey reports. By engaging participants to explore in-depth their cognitive processes while answering questions, cognitive interviewing in principle is designed to determine whether survey questions pose problems in comprehension, retrieval, judgment, or answer formatting, and to specify the exact nature of these problems.

However, the findings of cognitive interviews may not generalize to actual field surveys. One issue is that cognitive interviews are largely based on convenience samples and small samples which do not mirror field surveys that typically involve large probability samples. A second issue is that the cognitive processes encouraged in cognitive interviews may not mirror those encountered in field surveys. For one thing, cognitive interviews are conducted in different settings and contexts than field surveys--they usually involve bringing participants into a controlled laboratory setting, whereas field surveys are administered to respondents who are in their own homes (if a household survey). In addition, the techniques used in cognitive interviews are largely unstandardized to allow the freedom to explore various cognitive processes, whereas the techniques of field surveys are typically standardized regarding the rules of interviewing. Finally, a third issue is that the interpretation of cognitive interviews are more subjective and based on the insights of the researcher rather than following publicly verifiable principles of scientific objectivity. Thus, conclusions drawn from various researchers and laboratories may not be reliable, a prerequisite for any ability to generalize outside the realm of the cognitive interviewing process.

Behavior Coding

The purpose of behavior coding is to identify those survey questions that pose the most threat to the ideals of standardized interviewing, both with respect to interviewer and respondent verbal behaviors. Interviewers are expected to read questions exactly as worded and to adequately and nondirectively address respondent misunderstandings with question content. Respondents, for their part, are expected in ideal conditions to be motivated to answer survey questions to the best of their ability and to express areas of misunderstanding if they occur.

Yet, with respect to data quality, it's not clear that the questions behavior coding prioritizes as problematic, that is, those questions that are most illustrative of being in variance with the ideals of standardized interviewing, are those that actually threaten the quality of survey report. Additionally, the problem codes, in and of themselves, do not precisely identify the kinds of cognitive and interviewing processes that are posing problems.

Beatty Paper on Cognitive Interviewing

Beatty conducts sensible and clever research in seeking to show that problems revealed in cognitive interviews are generalizable to field surveys. For the most part, Beatty seeks to discover whether the same problems revealed in cognitive interviewing also demonstrate themselves in survey situations. Importantly, it is demonstrated that both the problems and solutions revealed in cognitive interviews are (at times) mirrored in field surveys, and that the cognitive problems revealed in survey interviews do affect the quality of survey report even if the style of interviewing in field settings tends to mask these problems. Additionally, Beatty is working toward establishing that different interviewing techniques in cognitive interviewing does not affect the conclusions drawn, that researcher interpretations are not merely subjective, and that cognitive interviewing results are reliable across laboratories. Finding evidence in support of these hypotheses will go a long way toward demonstrating the utility of cognitive interviewing in improving the quality of survey report.

I have a couple of comments regarding this work that I believe characterizes its potential. As shown by Beatty in his presentation, there will be a need for objective coding measures in this work to assess issues of reliability and validity. Blixt, Dykema, and Lepkowski, in their presentation, have illustrated the benefits of using objective coding schemes in assessing which questions, and what aspects of questions, pose the greatest threat to data quality. Beatty also provides an illustration of the benefits of such coding with the analyses of the question dealing with respondents' assessment of how many days during the past month their health was not good. Beatty provides codes both for independent and dependent variables by coding whether the interviewer engaged in an elaborating probing style as is typical for cognitive interviews, or bu engaging in a re-orienting probing style as is more typical for field survey interviews, and by coding the precision of responses. Beatty finds that the elaborating probes revealed problems in cognitive processes that were masked by the re-orienting probes. No doubt that in extending this work the coding of cognitive interviews will be needed to assess whether different interviewing styles and different cognitive interviewing staff identify the same questions as problematic, and for the same reasons.

Related to this need for objective coding measures, the determination of whether laboratories that conduct cognitive interviews on small sample sizes will provide results that are generalizable to field surveys appears to remain as an intractable problem. Survey practitioners are not interested in any problem that may uniquely appear, after all, every survey question is likely to pose problems to some of the respondents some of the time. Rather, interest centers on those questions that pose the greatest threat, those that consistently reveal cognitive problems. As discussed by Blixt et al, the benefit of coding schemes is that they offer such an ability to identify the most problematic questions, but at the cost of requiring fairly large sample sizes (certainly beyond the tendency in cognitive interviewing to use sample sizes of 5-10 participants). Beatty in this research agenda will also require fairly large sample sizes to gain an understanding of the extent to which different interviewing techniques and different interviewers or laboratories are consistently finding the same problems.

Papers by Hess & Singer; Blixt, Dykema, & Lepkowski on Behavior Coding

Both of these papers indicate that variance from the ideals of standardized interviewing as revealed by behavior coding do affect the quality of survey reports. Importantly, quality of survey reports is measured in two different ways, by reliability of survey answers across the same survey questions administered on two occasions (Hess & Singer), and by the agreement of survey answers with external records (Blixt et al). Such a consistency of findings across different types of measures of data quality is reassurance regarding the authenticity of the results.

In comparing these different measures of data quality, both studies are able to ascertain the quality of factual data, but only the reliability measure (Hess & Singer) is able to determine whether there exist associations between behavior codes and the quality of answers to subjective questions. Interestingly, whereas both types of measures show that behavior coding is associated with survey quality with factual questions, Hess and Singer did not find reliable associations with subjective questions. Perhaps the fluid nature of subjective questions in the face of many competing contextual factors is responsible for the lack of findings.

Surprisingly, neither study found that interviewer question reading changes were associated with poorer data quality, in fact, Blixt et al. have counter intuitively found that interviewer variance from reading questions as written is associated with improvements in the exact matches between survey reports and medical records. In related work based on the same data, Belli and Lepkowski (1996) had not found any improvement in data quality associated with question wording changes. The difference between Blixt et al. and Belli and Lepkowski involves the manner in which comparisons of survey responses and external records were measured. Blixt et al. used a dichotomous dependent measure that distinguished between exact agreements and any disagreement, Belli and Lepkowski used a continuous measure based on the absolute value of the difference between reports and records. A possible explanation for the inconsistency of findings is that question-reading changes may be potent in affecting survey reports in opposite directions, on occasion being effective in leading to improved remembering, but at other times being counterproductive by encouraging poorer quality retrospective reports.

With regard to respondent behavior, both Hess and Singer and Blixt et al are consistent in showing that problem behaviors are associated with poorer quality reports. However, there are inconsistencies in that there are no reliable indications regarding the circumstances in which significant associations between the occurrence of problem behavior codes and data quality measures will appear. As one example, Blixt et al found that qualified answers are significantly associated with the occurrence of fewer exact matches in the reports and records for hospital stays over a 12 month reference period and office visits over a 4 week reference period, but qualified answers do not reveal significant associations in the auality of reports for office visits with either 6 month or 2 week reference periods. As another example, Blixt et al found that any respondent code problems were associated with fewer exact matches between reports and records for office visits that involved 12 month and 4 week reference periods, but not for 12-month hospital stays, 6- month office visits, or 2-week office visits. In short, there is no consistency in the appearance of significant findings based upon type of report or length of reference period, and just as Alice in Wonderland's Cheshire cat, the effects show themselves at unpredictable times. Another noteworthy inconsistency is that Blixt et al particularly found poorer data quality associated with qualified responses whereas Hess and Singer noted significant data quality decrements associated with problem codes other than qualified answers, since they treated qualified responses in the same way as adequate answers, as an indication of nonproblematic respondent behavior. Overall, these inconsistencies highlight the continuing problem as to how to interpret the precise relationships between behavior codes, interviewing dynamics, and the quality of survey response. Our present level of understanding only permits very tentative explanations for the associations that do appear.

One message that is particularly clear about this research is that respondent behaviors are more indicative of compromises to data quality than anything that the interviewer has direct control. Whereas respondent problem behaviors are associated with poorer data quality, interviewer problem behaviors are not. Belli and Lepkowski (1996) found additional evidence that interviewer behavior does not affect the quality of report by finding that with reports on 12-month hospital stays, regardless of whether an interviewer probed adequately or inadequately, the occurrence of probing was associated with greater discrepancies between reports and records. Apparently, it was the need to probe following respondent behavior that is driving this effect, since the manner in which interviewers probed did not matter.

Results suggest that improvements in data quality are less likely to be promoted by concentrating on interviewer adherence to standardized procedures, and more likely to be promoted by devoting attention to respondent needs that will facilitate the effective answering of survey questions. Such advances will depend on improvements in questionnaire design principles that maximize the ability of respondents to answer accurately and consistently.

Reference:

Belli, R. F., & Lepkowski, J. M. (1996). Behavior of survey actors and the accuracy of response. Health Survey Research Methods: Conference Proceedings (pp. 69-74). DHHS Publication No. (PHS) 96-1013.

Discussion: Validation of Cognitive Questionnaire Pretesting Methods

Theresa J. DeMaio
U.S. Bureau of the Census

I'd like to thank the authors for three very good papers. I enjoyed reading them all. And I think that the general topic of validating cognitive questionnaire pretesting methods is a very important one that deserves more attention than it receives. I'm going to focus my remarks today on the Beatty, Willis, and Schechter paper, since my research experience focuses more heavily on cognitive interviewing than on behavior coding.

I'd like to organize my comments around the four assumptions about the value of the methodology that were included in the paper. Beatty et al presented evidence from their work about these assumptions. I'm going to discuss the assumptions and present evidence relevant to them from my work at the Census Bureau. As with Beatty's examples, they are not derived from controlled experimental comparisons, but they are illustrative nonetheless.

Assumption #1. The cognitive interviewing method finds problems that will carry over to surveys. This is an important assumption, and one for which we have quite a bit of anecdotal evidence, I think. Beatty et al presented some evidence in their paper, and my work also substantiates this assumption. At the Census Bureau, we've done some testing of forms being developed for the 2000 census. The testing focused on the design aspects of the forms, rather than their content. The forms are self-administered, and that provides a bit of a twist to the average cognitive interview in ways that I will get back to later.

A well-planned and well-executed research program would incorporate preliminary stages of testing such as cognitive interviews prior to field testing. However, we all know how the constraints of operational schedules wind up squeezing the testing. In this case, I think there were definite advantages to the fact that the cognitive testing of three proposed census short forms took place simultaneously with a nationally representative field test that included these forms along with others.

Cognitive interviews showed that respondents thought two of the mailing envelopes were too flashy and didn't look official enough. The message that the census is mandatory, which was included on all the envelopes, was not imparted to respondents in some cases. There were differences in design aspects of the questionnaires, too, that were noted differentially as problematic by respondents, who completed all three forms. There was no roster on any of the forms and the item that requested the number of people living in the household had different, and in some cases unacceptably high, rates of item nonresponse. And the concept of the census including everyone in the household was not adequately conveyed on any of the forms.

When the nationally representative field test results came in (in the 1996 National Content Survey), the mail return rates for the envelopes that were viewed as flashy and unofficial suffered in comparison to the official envelope. Item nonresponse rates for the item requesting the number of household members were high. And many forms were received at the processing office that contained a single household member's name and information repeated in the answer spaces for up to five persons. In short, what we found in the laboratory was also experienced in the field.

Assumption #2. The response process when answering questions in a cognitive laboratory is more or less the same as in a survey interview. The very wording of this assumption assumes that cognitive interviews are conducted with interviewer-administered interviews, and all of the research reported here today deals with that type of interview. I agree that this assumption is inherent in the cognitive interview method, and the Beatty et al paper provides one clever attempt to provide evidence about this assumption. However, I think the assumption needs to be reworded to encompass self-administered interviews as well. In a self-administered interview, there is even more reason to question whether this assumption is a reasonable one. The respondent sits across the table from the cognitive interviewer, and while the respondent is completing the questionnaire, the interviewer in a concurrent interview is frequently asking probing questions.

One of the consistent findings we have noted in our research on self-administered questionnaires is that respondents invariably have problems with skip instructions. We have done interviews with different types of respondents, different questionnaire content, different formats for skip instructions and problems with skip instructions seem to be a constant. One possibility, of course, is that the cognitive interview situation, in requiring the respondents to focus both on the interviewer and the questionnaire, affects the respondent's ability to concentrate on the printed document and thus introduces skip pattern errors that would not occur otherwise. Although I can't say one way or the other whether this hypothesis is correct, I think it is an important research issue.

I and my colleague Cleo Jenkins have been considering this issue. While we haven't had an opportunity to collect information about skip instructions, we have collected data that may shed light on other aspects of the self-administered completion process. In the census form research that I mentioned previously, we built in a controlled experiment in which a random half of the interviews were conducted using concurrent think aloud methods and the other half were conducted using retrospective think aloud methods. (In retrospective interviews, the probing is conducted after the form is completed, while the probing in a concurrent interview takes place while the form is being completed.) We haven't had a chance to analyze these data yet, but I think it will provide a good opportunity to learn about the kinds of errors that respondents make in a concurrent vs. a retrospective interview, as well as the kinds of information that can be obtained through each. Unfortunately for us, but perhaps fortunately for the general public, the census short form does not contain skip instructions! A retrospective interview is not the same as completing the form at home, but at least we're taking incremental steps in the right direction.

Assumption #3. Cognitive interviewer behavior does not have an undue effect on the content of the interview. The Beatty et al paper notes that this assumption refers to two different things: first, that the interviewer's behavior affects the respondents' answers to the survey questions themselves, and second, that the interviewer's behavior affects the number of problems, types of problems, etc., that he/she encounters with the questions. I'm not sure that this assumption is really needed. I think the first aspect of the statement seems to overlap with assumption #1: that is, if cognitive interviewer behavior has an undue effect on the survey responses, it seems to me this would mean that cognitive interviewing results would not carry over to the actual survey. On the other hand, the second aspect refers to nonsystematic cognitive interviewer behaviors that could affect the research results they receive. And this seems to overlap with assumption #4, which I'll talk about next.

Assumption #4. The cognitive interviewing process is basically reliable--if repeated, it would yield similar results. This very important assumption is largely untested. Presser and Blair (1994) compared results across various pretesting methods, including cognitive interviewing, and within multiple trials of each one using a questionnaire that was a composite of various National Health Interview Survey supplement questionnaires in the early stages of development. They found that the results across three trials of cognitive interviewing were not totally consistent. They correlated the overlap between the questionnaire problems that were identified during the three independent sets of cognitive interviewing, and found that the correlations ranged from .4 to .6. This is the only systematic attempt I know of to compare the results of cognitive interviewing across interviewers or interviewing organizations. I know there have been other instances where, for example, the Census Bureau and the National Center for Health Statistics have conducted interviews on the same projects, but there has been no attempt to conduct comprehensive systematic analysis to compare the results. I think this is an area that is in need of future research, and I'm glad to see that Paul and his colleagues have some plans in that area.

Those are the assumptions that are presented in the Beatty et al paper. I also think there is another basic assumption that underlies the cognitive research we do that is mentioned but not given much prominence in the Beatty et al paper. I think it is important that we try to confront this issue, so I'd like to add another assumption to the list.

Assumption #5. Respondents have sufficient access to their thought processes that they can verbalize how they go about answering survey questions. This is in some ways related to assumption #1, but I think it goes deeper than that. We take what our respondents tell us as accurate renditions of their thought processes. Yet those of us who have conducted cognitive interviews know that there are distinct differences among respondents in their ability to verbalize what they are thinking about. Failure to verbalize a problem is not necessarily an indicator that no problem exists. Eleanor Gerber and Tracy Wellens (1996) have suggested that respondents may not be aware of cultural factors that come into play during the response process. And respondents may not appreciate the influence of the visual aspects of self-administered forms when they are completing them.

In some of our recent interviews on the census form, one of the objectives was to evaluate how respondents reacted to icons, or pictures with benefits messages, that were included on the form to provide information about why census questions are asked. Two kinds of things happened. In a few cases, respondents actually read some of the icons while they completed the form, but when asked whether they had noticed them, they said no. One respondent offered as an explanation, "Well, it might have gotten into my conscious but it never got into my subconscious." However, the more frequent occurrence was that respondents didn't appear to notice the icons at all, but when they were asked about them later, it was clear they had processed them, even though they never mentioned them during the think aloud. My point in bringing this up is to note that I think we need to investigate this assumption, like the others that are included in the Beatty et al paper.

In conclusion, I want to thank Beatty and his colleagues for their attempt to specify the assumptions that underlie the cognitive interview research that we do, and for giving me the opportunity to think about them.

REFERENCES

Gerber, E. and T. Wellens. 1996. “Perspectives on Pretesting: ‘Cognition’ in the Cognitive Interview.” Paper prepared for presentation at the Essex ‘96 Fourth International Conference on Social Science Methodology, July 1996.

Presser, S., and Blair, J. 1994. “Survey Pretesting: Do Different Methods Produce Different Results?” Sociological Methodology, Vol. 2, No. 12, pp. 73-104.