A COMPARISON OF COGNITIVE INTERVIEWING, EXPERT REVIEW, AND BEHAVIOR CODING: WHAT DO THEY TELL US?

Gordon B. Willis, Research Triangle Institute
Susan Schechter, Karen Whitaker, National Center for Health Statistics
Gordon Willis, Suite 420, 6110 Executive Blvd., Rockville, MD 20852

Key Words: Survey pretesting, cognitive methods

Introduction

It is common practice to evaluate the quality of survey questionnaires, prior to field administration. In particular, the past 15 years have witnessed the rapid growth of the use of the Cognitive Interview for evaluating potential sources of error in draft instruments (Forsyth & Lessler, 1991; Jobe & Herrmann, 1996; Lessler, Tourangeau, & Salter, 1989; Tourangeau, 1984; Willis, Royston, & Bercini, 1991). Separately, Cannell, Fowler, and colleagues have developed Behavior Coding techniques (Fowler & Cannell, 1996; Oksenberg, Cannell, & Kalton, 1991), which endeavor to chronicle the overt problems that occur in questionnaires. Finally, the use of the Expert Review has also been common practice in questionnaire development, simply because critical review by experts, prior to implementation, seems like a reasonable approach to improving quality (Forsyth & Lessler, 1991). In this paper we refer to these practices as questionnaire pretesting techniques (despite the fact that Expert Review is not empirical in nature, and might not be thought of as "pretesting").

The survey methodology field has passed the point of "trying out" pretesting techniques; these have become standard fare in the diet of survey development. Therefore, the current study builds on a small but growing body of research that attempts to answer several fundamental questions (see in particular Presser & Blair, 1994):

1) Are these techniques useful for pretesting questionnaires?

2) Do they give the same results when "stacked up" against one another?

3) If they fail to give the same results, what is different about them? Which one gives the best results?

4) If they *do* give the same results, which one is the easiest to conduct or most economical?

Willis, DeMaio, and Harris-Kojetin (1999) review the fundamental issues involved in comparing pretesting techniques, and conclude that a series of studies would be necessary to examine different facets of the general question, "Which pretesting method is best?" Therefore, our experiment addresses one, more limited research question: If the same questionnaire is subjected to Cognitive Interviewing, Behavior Coding, and Expert Review, will similar results be obtained, in the sense that the techniques detect the same qualitative types of problems?

Each of the pretesting techniques studied— Cognitive Interviewing, Behavior Coding, and Expert Review— consists of a number of inter-related features, and can be implemented in myriad ways. Therefore, we first describe those variants that we have chosen to study, and distinguish these from other, similar approaches, which we do not evaluate (see also Campanelli, 1997).

Cognitive Interviewing

The cognitive approach to the evaluation of survey questionnaires seeks specifically to identify problems that may be associated with respondents' cognitive processes (see Jobe and Herrmann, 1996). The conduct of Cognitive Interviews is a common means for applying the cognitive model in a manner that may ultimately improve the quality of survey questions, through the study of comprehension, retrieval, judgment, and response processes. Note that the term "Cognitive Interview" may represent as many identifiably different concepts as there are researchers who apply it; the procedure has limited common definition, and is in reality a family of related practices (see Blair & Presser, 1993; Conrad & Blair, 1996; DeMaio, Mathiowetz, Rothgeb, Beach ,& Durant,1993). We define the Cognitive Interview as the practice of using a limited degree of think-aloud instruction, combined with the judicious use of verbal probing by the interviewer. Probes may be scripted prior to the interview, or they may be spontaneously generated during the course of the interview. This procedure is described in detail by Willis (1994).

Expert Review

The defining feature of this technique is that the individuals applying it are considered to be expert in the critical appraisal of survey questionnaires. Beyond this, there are two major variables that produce differentiation in approach:

1) Individual versus group review. Very often, questionnaire "pretesting" consists of appraisal by an individual reviewer. Alternatively, a group-based Expert Review, resembling a focus group in which the object of attention is the survey questionnaire, may also be conducted. Presumably, the questionnaire has been reviewed before the group meeting by each of the individuals involved, so the group-review meeting may be seen as an extension of the individual review.

2) Informal review versus formal appraisal. A common means for conducting the Expert Review is to produce comments, pertaining to each survey question, in open-ended written form. However, several researchers have endeavored to develop more formal means for assigning explicit codes to problems that are found to exist in survey questions. For example, Lessler and Forsyth (1996) have developed the Forms Appraisal System, which contains more than 100 such codes. More recently, Willis and Lessler (1999) have developed a more compact coding system useable by relatively inexperienced questionnaire designers.

In this investigation, we focus on the individual form of review, conducted in an informal (note-taking) manner.

Behavior Coding

Behavior Coding was developed in order to focus on interviewer and respondent behavior in interaction, and therefore is also known as interaction coding. Behavior Coding relies on overt cues given during the administration of survey questions under field conditions (e.g., need to repeat question; respondent request for clarification). This technique is less invasive than Cognitive Interviewing, in that it represents a passive approach to pretesting (it involves no additional probing or other intervention, but simply the monitoring of the interview, e.g., through a tape-recording). The practice of Behavior Coding does not appear to exhibit the same degree of procedural variability as does Cognitive Interviewing or Expert Review. However, as this technique has developed, several variants have emerged, in terms of the number of separate coding categories used. The variant that might be referred to as the Cannell/Fowler system has typically contained roughly eight coding categories; this investigation used such a system.

Background studies

Fowler and Roman (1992) conducted a qualitative comparison of Cognitive Interviewing and Behavior Coding, and suggested that Cognitive Interviews were effective in identifying problems associated with question comprehension. Behavior Coding, on the other hand, was seen as detecting problems that may have been overlooked or not fully appreciated by interviewers. In the study most similar to the current one, Presser and Blair (1994) compared Cognitive Interviewing, Expert Review, and Behavior Coding, and used as a basic dependent measure the number of problems detected by each technique. Overall, the Expert Review panel identified the most problems, whereas Behavior Coding and Cognitive Interviewing produced similar numbers of identified problems. Presser and Blair also developed a qualitative coding scheme for use in describing the problems found in survey questions.

This present study follows the approach of Presser and Blair (1994) in determining the degree of overlap between Cognitive Interviewing, Expert Review, and Behavior Coding techniques. A subsidiary aim was to explore the degree of consistency *within* technique, following recommendations by Tucker (1997). In particular, we sought to answer the vital question of whether independent cognitive laboratories would produce similar results. To our knowledge, such a test has not previously been done in a formal manner (Presser and Blair compared *interviewers*, but not different *interviewing staffs).* A second departure from the Presser and Blair approach was to use larger sample sizes, for each of the pretesting techniques, in order to allow for a more statistically powerful assessment, and in the case of Behavior Coding, to produce a situation more representative of usual practice in survey organizations.

Further, rather than confounding pretesting technique with individual versus group-based implementation, our design involved individual-level evaluation for Cognitive Interviewing and Expert Review (that is, individual rather than group-based Expert Review, to be compared with individual-interview-level Cognitive Interviewing). Finally, the current design involved a system for identification of questionnaire problems that was more similar, across techniques, than that used in the Presser-Blair study. Specifically, as is done routinely for Behavior Coding, each reviewer or Cognitive Interviewer was forced to make a decision, for each survey question, of whether a problem existed for that question. This approach allowed a relatively straightforward system for data analysis and interpretation.

Mirroring the Presser-Blair approach, the current investigation did strive to determine whether, even if different pretesting techniques uncovered similar questions as problematic, they also detected similar qualitative *types* of problems. To this end, a coding system for the qualitative assessment of questions was produced and implemented.

Method

Questionnaire materials. The questionnaire that served as the standard basis across pretesting techniques consisted of 93 questions, mostly related to health. This questionnaire incorporated questions from both a draft of the NCHS National Health Interview Survey and the 1985 Canadian Survey on Aging and Independence. Paper-based (rather than computerized) questionnaires were used. Next to each question was placed a small box (a *Problem Box*), which the reviewer or interviewer was to check if a problem with the question was detected. Further, space was provided under every question for the reviewer/interviewer to enter written comments. For Expert Reviews, the reviewer was to check the box if he or she anticipated that one or more problems existed for that question. For Cognitive Interviews, the Problem Box was checked if a problem was observed during the interview, or in some cases, where the interviewer noted that a problem might occur (it was inferred), although this was not observed during the interview.[1] As an example, a reviewer might consider the term "biofeedback" to be too technical in nature for the typical survey respondent, or the Cognitive Interviewer might find that a laboratory respondent failed to understand that term, and enter a note to that effect.

Cognitive Interviewing procedures. Interviews focusing on the target questionnaire (as described above) were conducted independently by NCHS and NORC staffs. Forty-three Cognitive Interviews were carried out at the NCHS Questionnaire Design Research Laboratory by five experienced interviewers. Forty Cognitive Interviews were done by NORC staff members in Washington, D.C. and in Chicago, by four interviewers who were specially trained for this activity. Therefore, the NORC Cognitive Interviews were similar in approach to those of Presser and Blair (1994), who utilized relatively inexperienced interviewers, whereas the NCHS interviewing component was more representative of practice used in the everyday development and testing of questionnaires in Federal agencies (i.e., involving the efforts of a trained professional staff who routinely engage in these activities). Subjects responded to newspaper advertisements placed in the Washington Post (for NCHS interviews) or the Washington City Paper and the Chicago Reader (NORC interviews). The advertisement offered $30 to volunteers willing to spend an hour answering health questions. All subjects were 18 years of age or older.

Cognitive Interviewers were instructed to use both think-aloud and verbal probing. For NCHS interviews, interviewers developed their own probe questions, and asked these in either scripted or spontaneous fashion as they administered the interviews. Because they were less experienced as interviewers, NORC staff developed a series of scripted probes which interviewers were instructed to use, although they were also allowed to apply spontaneous probing. For each interview, the subject was first introduced to the task, and instructed to think aloud as they answered the survey questions. Interviewers administered the questions in face-to-face fashion, entered the respondents' answers, and probed as described above. Interviewers also "checked" the Problem Box as appropriate, and recorded written notes under each question found to produce problems. All interviews were tape-recorded, and lasted approximately 30-60 minutes.

Expert Review procedures. The questionnaire was reviewed independently by twenty-one staff members of five Federal agencies: NCHS, the Bureau of Labor Statistics (BLS), the Census Bureau, the General Accounting Office (GAO), and the National Agricultural Statistical Service (NASS). Expert Reviewers carried out activities similar to the Cognitive Interviewers, but through individual review of the questionnaires.

[1] There is little agreement whether Cognitive Interviewers should only note problems that they observe, or also record problems that they feel may exist (i.e., that they infer), even if not observed. A focus of the study not to be explicitly discussed concerns the degree to which problems were observed, as opposed to inferred, across the different techniques.

Behavior Coding procedures. Two Behavior Coding exercises using the target questionnaire were carried out. First, the questions were embedded in a larger paper questionnaire, as part of a face-to-face household pretest of the National Health Interview Survey (NHIS) conducted by Census Bureau interviewers in the Washington, D.C. metropolitan area (based on a convenience sample of households within unused Census Bureau Area Segments). Interviews were conducted with one adult household member who was 18 or older, at home at the time of the interview, and willing to be interviewed. A total of 29 interviews were tape recorded. The methodological section of each taped interview was then behavior coded at the Center for Survey Research (CSR) at the University of Massachusetts, and the tabulated coding results returned to NCHS.

The second Behavior Coding exercise was also done by CSR. In this case, CSR conducted 89 telephone interviews, using a random-digit-dial-based sample of residential phone numbers within the contiguous 48 States. Each respondent was again an 18+ household member who was home at the time that the call was made. Of the 89 interviews that were taped, the 83 that were intelligible were behavior coded by CSR staff.

For both studies, the basic behavior codes used by CSR were applied by three experienced coders who listened to the taped interviews. In analysis, if a question received any problem code, a Problem Box similar to the one used in the Cognitive Interviews and Expert Reviews was checked.

Results

The analyses described below centered on three basic issues: a) The number of problems identified by each pretesting technique, b) the correspondence within and between techniques with respect to the detection of these problems, and c) assessment of the qualitative types of problems identified by each of the techniques.

Number of problems identified. The first level of analysis involved identifying the frequency with which problems were found in the questionnaire by each exercise involving a pretesting technique (the NCHS Cognitive Interviews, NORC Cognitive Interviews, Expert Reviews, and the two Behavior Coding studies). In order to compute Problem-Box-based percentage scores, the percentage of times that the Problem Box was checked (or a problem code was assigned, in the case of Behavior Coding) for each item was calculated. For example, for the NCHS Cognitive Interviews, a question receiving ten checks across the 43 interviews received a score of 23.3% (10/43 x 100). Across the 93 questions, the percentage of Cognitive Interviews identifying a problem ranged from 0 (no Problem Boxes checked for the 43 interviews) to 41.9% (18 Problem Boxes out of 43 interviews were checked). These percentages were then averaged, across all questions, to produce the overall mean for each of the techniques, and are listed in Table 1.

Table 1. Problem percentage scores for each exercise involving the use of a pretesting technique.

Mean (%)	Standard Deviation	Minimum	Maximum
Cognitive Interviewing - NCHS (n=43):			
11.9	10.1	0	41.9
Cognitive Interviewing - NORC (n=40):			
12.3	10.0	0	42.5
Behavior Coding - household (n=29):			
20.7	17.0	0	81.0
Behavior Coding - telephone (n=83):			
26.1	23.7	1.2	100.0
Expert Review (n=21)			
27.0	20.5	0	71.4

Note: n refers to either the number of Expert Reviews, Cognitive Interviews, or Expert Reviewers.

Overall, the Expert Review produced the highest level of problems (27%), in accordance with Presser and Blair's (1994) findings. The Behavior Coding exercises also produced fairly high problem frequencies, although the household-interview-based pretest identified slightly fewer problems than the telephone survey (21% versus 26%, respectively). On average, the Cognitive Interviews reported the fewest reported problems. Notably, despite the considerable procedural differences between them, findings were similar between the two cognitive laboratories, with NCHS and NORC both producing values of approximately 12%.

These basic results are somewhat interesting, but clearly limited. Willis et al. (1999) in particular have argued that in the absence of information concerning question quality, the finding of more problems cannot be viewed as evidence that a particular pretesting technique is superior. In particular, "more is better" reasoning focuses solely on the issue of sensitivity, and ignores the possibility that a highly sensitive procedure may have poor specificity, and produce a large number of false positive results. Therefore, we endeavored to conduct further analysis that focused explicitly on the extent of measurable agreement between techniques.

Correlational Analysis. The second level of analysis ascertained the degree to which the same questions were found to be problematic, within and across pretesting techniques, through use of correlation analysis. Correlations were computed on an item basis, using as raw data the percentage score values described above, so that correlations were obtained for 93 data points (items) in pairwise fashion, for five sets of data. We made several hypotheses, concerning the expected magnitude of these correlations: a) all would be positive, and b) comparisons involving the same technique, either across site (NCHS versus NORC Cognitive Interviewing) or across replication (household versus telephone Behavior Coding) would produce the highest values.

Further, across different techniques, we predicted that the magnitude of correlation coefficients would follow a pattern that depended on the degree of similarity between techniques. In particular, we based our hypotheses on a model stipulating that the techniques exist on a *continuum of objectivity*, from most objective to most subjective, according to the ordering: a) Behavior Coding, b) NORC Cognitive Interviewing, c) NCHS Cognitive Interviewing, and d) Expert Review. Based on this model, we predicted the following specific set of results:

1) The correlation between Behavior Coding exercises (household versus telephone) would be the highest;

2) The correlation between NCHS and NORC Cognitive Interviews would also be high, but not as high as (1), because of the greater degree of subjectivity between Cognitive Interviewing variants;

3) The correlation between Expert Review and Behavior Coding (either household or telephone) would be the lowest, as these represent the most disparate methods;

4) The correlation between Expert Review and NCHS Cognitive Interviews would be higher than that between Expert Review and NORC Cognitive Interviews, because NCHS interviews presumably involved a greater degree of subjectivity than did those conducted by NORC, and more closely reflected an Expert Review type of activity;

5) On the other hand, the correlation between Behavior Coding and NCHS Cognitive Interviews would be lower than that between Behavior Coding and NORC Cognitive Interviews, based on reasoning similar to that expressed in (4) above.

These hypotheses, in sum, specify a strict ordering of correlation coefficients that can be easily tested. The relevant results appear in Table 2. Note first that as expected, all correlations are positive. As predicted, the highest correlation (.79) was between the two Behavior Coding studies. The NCHS and NORC Cognitive Interviews were also highly correlated, but somewhat less so (.68). This finding is indicative of fairly good between-method reliability; across laboratories, or across replication, the same techniques tend to identify the same items as having problems. This is a somewhat reassuring result, in that it addresses the concern expressed by Tucker (1997) that the results of these methods may be idiosyncratic in nature. Also as predicted, between-technique correlations were lower than within-technique correlations, but were still moderate to high.

With respect to other specific hypotheses:

1) The correlation between Expert Review and Behavior Coding (.54) was somewhat lower, but not the lowest values obtained, so this was in violation of expectation.

2) On the other hand, the correlation between Expert Review and NCHS Cognitive Interviews (.48) was higher than that between Expert Review and NORC Cognitive Interviews (.33), as predicted.

3) The correlations between the two Behavior Coding exercises and NCHS Cognitive Interviews (.49/.59) were lower than those between Behavior Coding and NORC Cognitive Interviews (.53/.73), again as predicted.

Overall, the correlation analysis was fairly supportive of our assertions concerning the continuum of objectivity that may exist between pretesting techniques.

Table 2. Correlations of item percentage scores, between pretesting techniques.

	NCHS Cognitive Interviews	NORC Cognitive Interviews	Behavior Coding (household)	Behavior Coding (telephone)	Expert Review
NCHS Cognitive Interviews:	-	.68	.49	.59	.48
NORC Cognitive Interviews:	.68	-	.53	.73	.33
Behavior Coding (household):	.49	.53	-	.79	.54
Behavior Coding (telephone):	.59	.73	.79	-	.54
Expert Review:	.48	.33	.54	.54	-

Note: All $\underline{p} < .001$

Qualitative analysis. The analyses described to this point are interesting, but provide no insight into a fundamental issue: Even if the different techniques find the same questions to be problematic, are they finding the same *types* of problems with the questions? In order to establish this, one must delve beyond the basic Problem Box level of analysis conducted above (the binary decision of whether a problem is identified or not), and further classify these problems according to a system that is able to assess the characteristics of these problems. The development of such a qualitative scheme has been of substantial interest in the survey methodological literature (Beatty, undated; Conrad & Blair, 1997; Lessler & Forsyth, 1996; Presser & Blair, 1994; Willis & Lessler, 1999; Willis, Royston, & Bercini, 1991).

Beatty (undated) has compared and contrasted coding schemes, and demonstrated that despite wide variations, these tend to focus on several key categories, depending in part on the extent to which they are inspired by a particular cognitive model of the survey response process. In particular, these systems tend to be grouped into: a) those that focus mainly on question comprehension and on features of question meaning, and b) those that apply relatively more equal emphasis to cognitive processes other than comprehension (recall, decision, response).

Following Conrad and Blair (1997), we chose the latter path, in an attempt to provide a generally comprehensive and evenly distributed system. Our coding scheme was initially based on a model of the survey response process (Willis et al., 1991) which emphasizes comprehension, recall, decision, and response processes, as well as logical problems (those not clearly associated with respondent cognitive processes, such as skip pattern errors, logical inconsistencies, and erroneous assumptions). This system is presented in Table 3.

Each of the 43 NCHS Cognitive Interviews, the 40 NORC Cognitive Interviews, and the 21 Expert Reviews were coded according to the five-category (CO, RE, BI, LC, LO) system (note that Behavior Coding was not directly amenable to such a system, and therefore was not included in this qualitative analysis). Specifically, for each interview or review, the coding scheme was applied to the written comments concerning the nature of the problems either found or anticipated to exist in the problematic questions. For example, the comments "subject couldn't understand the question" and "no one will know what this means - it's too long" received the code CO. Only questions that had the Problem Box checked by the reviewer/interviewer were coded using the scheme outlined above. Questions could be assigned up to three different problem codes, although use of more than two was rare.

Table 3. Coding system used to describe qualitative nature of problems through Cognitive Interviewing and Expert Review.

1) CO: Comprehension/Communication problems-
   a) IN: Administration problems for the interviewer
   b) LE: Problems due to question length
   c) TE: Problems with specific terms
   d) DI: Problems related to question difficulty
   e) VA: Problems related to question vagueness

2) RE: Recall-based problems
3) BI: Bias/Sensitivity
4) RC: Problems with response categories
5) LO: Logical/Structural problems with question

Inter-rater coding reliability was assessed for NORC Cognitive Interviews by having all interviews double-coded, and for NCHS Cognitive Interviews and Expert Reviews by having a second coder (a research assistant trained in the use of the coding scheme) also code a 50% subset of the data. For each of NORC Cognitive Interviews, NCHS Cognitive Interviews, and Expert Reviews, coding reliability was found to be adequate (kappa statistics exceeded .60, both overall and for each coding category, which is generally classified as "good" reliability).

For both Cognitive Interview exercises, and for the Expert Reviews, the number of codes assigned were then aggregated across all of the questions and reviewers/interviewers. The overall results of the coding analyses, in terms of code distribution, are presented in Table 4, for each of the major problem type codes. The basic result is immediately apparent; across methods, the "lion's share" of the codes were clearly assigned to one category: CO. In this light, the fact that the overall coding reliability may be high is of somewhat modest importance, as it may simply reflect the fact that this one dominant category captures the majority of identified problems. This result in itself may be significant, depending on the degree to which one is prepared to make generalizations based on these results. It could be that the questionnaire items selected for this study simply happened to contain more CO problems than others, or that the coding scheme we developed is best equipped to detect these types of problems.

Or, perhaps more interestingly, it could be that the nature of problems with survey questions may in large part involve problems in communication of meaning. Through in-depth analysis of the nature of Cognitive Interviewing, Gerber and Wellens (1997) have concluded that the practice of pretesting in large part involves the study of question meaning; that is, the communicative features of questions that are captured by our Comprehension/Communication coding category (see also Fowler & Roman, 1992). If Gerber and Wellens are correct, then our initial attempt at developing a coding scheme was somewhat misguided, in that it failed to take into account the extent and variety of communication and/or comprehension problems that may be inherent in problematic survey questions. To examine this issue further, we elaborated our CO problem type category into five subcodes (see Table 3), producing a system more consistent with the philosophy of Presser and Blair (1994). In brief, the results of an analysis that relied on this disaggregation, by several coders, again demonstrated a high degree of agreement in the use of sub-codes, between techniques, when applied to NCHS Cognitive Interviews and Expert Reviews (see Table 5).

Table 4. Overall percentage of problem type codes assigned, for each pretesting technique.

	CO	RE	BI	RC	LO	Total
	(Communication)	(Recall)	(Bias)	(Response Categories)	(Logical)	
NCHS Cognitive Interviews	70.5% (332)	11.0% (52)	1.9% (9)	12.1% (57)	4.5% (21)	100% (471)
NORC Cognitive Interviews	58.1% (358)	13.3% (82)	1.3% (8)	19.8% (122)	7.5% (56)	100% (626)
Expert Review	75.1% (414)	7.8% (43)	3.3% (18)	9.1% (50)	4.7% (26)	100% (551)

Note: The number of codes assigned in each technique is given in parentheses.

Table 5. Distribution of major codes and CO subcodes, by pretesting technique.

	NCHS Cognitive Interviews	Expert Reviews
1) CO*	70.5%	75.1%
IN	0.6%	1.7%
LE	1.7%	3.8%
TE	21.4%	22.9%
DI	11.9%	11.3%
VA	25.5%	24.5%
**	9.3%	11.1%
2) RE	11.0%	7.8%
3) BI	1.9%	3.3%
4) RC	12.1%	9.1%
5) LO	4.5%	4.7%
	100.0%	100.0%

* CO subcodes are expressed as percentages of the total number of codes assigned.

**Uncodeable

Discussion

We asserted at the beginning of this paper that no single study will answer the fundamental question: "Do different pretesting techniques work, and which one is best?" However, to the extent that our results may represent a useful piece of the puzzle, several conclusions seem warranted:

Pretesting technique reliability. As conducted within this study, pretesting techniques appear to exhibit a reasonable degree of consistency. Of particular interest is the result from Behavior Coding; researchers often wonder "how much is enough," when conducting a Behavior Coding study, and the issue of whether a relatively small sample is of use often emerges. In the current study, we found that the results from a pretest involving only 29 recorded interviews were very similar to those from one involving 83 respondents, indicating that useful information may be gleaned from a much smaller sample than is often involved in Behavior Coding exercises (Zukerberg, Von Thurn, and Moore, 1995, have made a similar argument).

Second, Cognitive Interviews that were conducted on different subjects, by different interviewers of disparate experience, and using somewhat different methods (NCHS versus NORC), revealed similar results. To some extent, this addresses the concern that once these variables are confounded, method reliability will suffer interminably. However, a more pressing concern may be the criticism that the current study did not engage Cognitive Interviewing in the manner normally applied by survey research organizations. In fact, Cognitive Interviewing procedures typically: a) make use of smaller testing round (often only 8-12 subjects), and b) involve multiple rounds of testing, in iterative fashion, with changes to questions made between rounds. We acknowledge these limitations; further analysis of our data, or further studies, will be necessary before we can adequately address these issues.

Relative confidence in techniques. The fact that different techniques appear to be somewhat similar in their results, depending on the closeness of these techniques, but are not completely overlapping, could be seen as positive result (see Esposito and Rothgeb, 1997). Presumably, different techniques have different strengths; for example, Behavior Coding is relatively objective, makes use of larger samples, and focuses on observation of overt problems. On the other hand, Cognitive Interviewing trades off strength in numbers (sample size) for intensity of focus; the effect of verbal probing is very likely to "bring out" covert problems that are otherwise not directly observable. Therefore, it is not surprising that there is some correspondence between the problems found, although the overlap is not total.

Competition between techniques. We make a final statement concerning the basic logic involved in "stacking up" pretesting techniques against one another, in order to determine which is best. Several authors (Campanelli, 1997; Willis et al., 1999) have challenged the notion that the methodological "holy grail" consists of identifying the technique that is superior in an absolute sense. Rather, these methods may naturally tend to insert themselves at particular points in the survey development process, simply through the nature of their constituent features.

Expert Review is easily and efficiently conducted very early in the development process, especially as the lack of coherent skip patterns, exact wording, and format may not be a major impediment. Cognitive Interviewing is a logical follow-up step (if only because it is sensible that a Cognitive Interviewer be a questionnaire design expert who has reviewed the questionnaire, prior to conducting these interviews). Finally, the subsequent Behavior Coding of a field pretest may take advantage of the fact that a relatively large sample is available, that interviewers will be (or, in most cases, should be) reading the questions and following instructions as scripted, and that the procedure lends itself best to passive observation, as opposed to invasive probing. Given these constraints and scheduling realities, there is no reason to believe that one must choose one technique over the others. The current finding that these techniques do seem to bolster, rather than to conflict with one another, may be sufficient justification for asserting that they provide the basis for an integrated package of techniques that in combination provide an effective means for contributing to the questionnaire development process.

References

Beatty, P. (undated). *Classifying Questionnaire Problems: Five Recent Taxonomies and One Older One*. Unpublished manuscript, Office of Research and Methodology, National Center for Health Statistics.

Blair, J., and Presser, S. (1993). Survey Procedures for Conducting Cognitive Interviews to Pretest Questionnaires: A Review of Theory and Practice. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 370-375.

Campanelli, P. (1997). Testing Survey Questions: New Directions in Cognitive Interviewing. Bulletin de Methodologie Sociologique, *55*, 5-17.

Conrad, F., and Blair, J. (1997). From Impressions to Data: Increasing the Objectivity of Cognitive Interviews. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 1-9.

DeMaio, T., Mathiowetz, N., Rothgeb, J., Beach, M. E., and Durant, S. (1993). *Protocol for Pretesting Demographic Surveys at the Census Bureau.* Unpublished manuscript, Center for Survey Methods Research, U.S. Bureau of the Census.

Esposito, J.L., and Rothgeb, J.M. (1997). Evaluating Survey Data: Making the Transition from Pretesting to Quality Assessment. In L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz, and D. Trewin, *Survey Measurement and Process Quality*, (pp. 541-571). New York: Wiley.

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

Fowler, F. J., and Cannell, C. F. (1996). Using Behavioral Coding to Identify Problems with Survey Questions. In N. Schwarz & S. Sudman (Eds.), *Answering Questions: Methodology for Determining Cognitive and Communicative Processes in Survey Research*, (pp. 15-36). San Francisco: Jossey-Bass.

Fowler, F.J., and Roman, A.M. (1992). *A Study of Approaches to Survey Question Evaluation.* Unpublished manuscript, Boston: Center for Survey Research, University of Massachusetts.

Gerber, E. R., and Wellens, T. R. (1997). Perspectives on Pretesting: "Cognition" in the Cognitive Interview? *Bulletin de Methodologie Sociologique, 55*, 18-39.

Jobe, J. B., and Herrmann, D. J. (1996). Implications of Models of Survey Cognition for Memory Theory. In D. J. Herrmann, C. McEvoy, C. Herzog, P. Hertel, and M. K. Johnson (Eds.), *Basic and Applied Memory Research: Practical Applications: Vol. 2*, (pp. 193-205). Mahwah, New Jersey: Erlbaum.

Lessler, J. T., and Forsyth, B. H. (1996). A Coding System for Appraising Questionnaires. In N. Schwarz and S. Sudman (Eds.), *Answering Questions: Methodology for Determining Cognitive and Communicative Processes in Survey Research*, (pp. 259-291). San Francisco: Jossey-Bass.

Lessler, J.T., Tourangeau, R. and Salter, W. (1989). Questionnaire Design Research in the Cognitive Research Laboratory. *Vital and Health Statistics* (Series 6, No. 1; DHHS Publication No. PHS-89-1076). Washington, DC: U.S. Government Printing Office.

Oksenberg, L., Cannell, C., and Kalton, G. (1991). New Strategies for Pretesting Survey Questions. *Journal of Official Statistics*, 7, 3, pp. 349-365.

Presser, J., and Blair, J. (1994). Survey Pretesting: Do Different Methods Produce Different Results? In P.V. Marsden (Ed.), *Sociological Methodology*, Vol. 24, (pp. 73-104). Washington, DC: American Sociological Association.

Sudman, S., Bradburn, N., and Schwarz, N. (1996). *Thinking About Answers*, San Francisco: Jossey-Bass.

Tourangeau, R. (1984). Cognitive Science and Survey Methods. In T. Jabine *et al.* (Eds.), *Cognitive Aspects of Survey Design: Building a Bridge Between Disciplines,* (pp. 73-100). Washington, DC: National Academy Press.

Tucker, C. (1997). Measurement Issues Surrounding the Use of Cognitive Methods in Survey Research. *Bulletin de Methodologie Sociologique, 55*, 67-92.

Willis, G. B. (1994). *Cognitive Interviewing and Questionnaire Design: A Training Manual.* National Center for Health Statistics: Cognitive Methods Staff (Working Paper No. 7).

Willis, G.B., DeMaio T.J, and Harris-Kojetin B. (1999). *Is the Bandwagon Headed to the Methodological Promised Land? Evaluating the Validity of Cognitive Interviewing Techniques.* In M. Sirken, D. Herrmann, S. Schechter, N. Schwarz, J. Tanur, & R. Tourangeau (Eds.), Cognition and Survey Research, (pp. 133-153). New York: Wiley.

Willis, G.B., and Lessler, J.T. (1999). *The Question Appraisal System: A Guide for Systematically Evaluating Survey Question Wording.* Final Report submitted to the Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion. Research Triangle Institute.

Willis, G.B., Royston, P., and Bercini, D. (1991). The Use of Verbal Report Methods in the Development and Testing of Survey Questionnaires. *Applied Cognitive Psychology*, 5, 251-267.

Willis, G.B., and Schechter, S. (1997). Evaluation of Cognitive Interviewing Techniques: Do the Results Generalize to the Field? *Bulletin de Methodologie Sociologique*, 55, pp. 40-66.

Zukerberg, A.L., Von Thurn, D.R., and Moore, J.C. (1995). Practical Considerations in Sample Size Selection for Behavior Coding Pretests. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 1116-1121.

Acknowledgments

This study was conducted over an extended time period, and fruition of this effort depended on the efforts of a number of individuals. Jeremy Morton conducted coding analyses and assisted in writing a preliminary version of this paper that was presented as a class project at the Joint Program for Survey Methodology of the University of Maryland/University of Michigan. Jon Gonzalez-Willis, Suzanne Moien, and Mark Bradbury served as research assistants, and carried out coding analyses. We would also like to thank 21 anonymous staff members of the Bureau of the Census, Bureau of Labor Statistics, General Accounting Office, National Agricultural Statistical Service, and National Center for Health Statistics, who served as Expert Reviewers.