Using Behavior Coding to Validate Cognitive Interview Findings

Johnny Blair[1], **Allison Ackermann[1]**, Linda Piccinino[1], Rachel Levenstein[2]
[1]Abt Associates, Inc, 4550 Montgomery Ave, Suite 800 North, Bethesda, MD, 20814
[2]University of Michigan, Ann Arbor, Michigan

1. Introduction

The fundamental assumption of survey pretesting is that questionnaire problems identified in pretesting will, if left uncorrected, occur in the fielded survey. When pretesting emulates the survey's interview procedures, the logical link between the pretest findings and the expectation of their reoccurrence in the survey is obvious. For example, often a pretest consists of a small number of interviews conducted using the planned data collection protocol. In that circumstance, any procedures used for problem identification that do not interfere with normal survey administration would produce valid indicators of problems. Such procedures might include post-interview group or individual interviewer debriefings, behavior coding, or respondent debriefing (assuming the respondents are not informed prior to the interview that they will be debriefed).

When cognitive interview pretests are conducted in a laboratory setting, the fundamental, if often unspoken, assumption is also that identified problems will occur in the actual survey. However, the conditions under which problems occurred and the conditions under which the survey will actually be conducted are quite different. Problems identified in the verbal reports of paid, volunteer, laboratory respondents may well also occur in the main survey, but there is little in the research literature to support this critical presumption.

Certainly, it would be a bold statement to claim that all laboratory-identified problems will cause response or other difficulties in the main survey. If one is not willing to make that claim, then the logical research question becomes: to what extent do problems identified in cognitive interview pretests actually occur in the field; and if they do occur, how frequently does that happen?

Willis and Schechter (1997) partially addressed this issue in a study that examined the field interview response patterns of a small set of questions that had been cognitively tested. Their analysis of response patterns that occurred in the field supported the hypothesis that the problems identified in the laboratory did affect respondent answers. One limitation on the generalizability of their findings is that the problems were of a type that, if they occurred, would be expected to produce specific response patterns. On the positive side, the specificity of the expected response patterns allowed Willis and Schechter to formulate and test exact hypotheses. This was an important study, however, considering the weight generally given to cognitive interview findings in decisions about the revision of survey questions, it is surprising that there has been scant follow-up Willis and Schechter's work.

It may be that the very success of cognitive interviewing as a pretest method is one reason for the lack of research on the validity of cognitive pretest findings. Problems identified in the lab are typically addressed. Therefore, opportunities for validation of cognitive interview findings in actual surveys is very limited.

The development of a health survey at Abt Associates presented an opportunity to do a validity check on a set of lab findings. For reasons described below, some questions in a telephone survey questionnaire were tested in the cognitive laboratory, but the question problems that testing identified in some items were not repaired and the questions were used in the survey in the same form in which they were tested.

In order to determine whether the laboratory-identified problems occurred in the telephone interviews, behavior coding was employed. Behavior coding is a possible validation tool because it does not affect data collection procedures and produces "results [that] are systematic, objective, and replicable" (Groves et al. 2004, p249). However, it was not clear whether behaviour coding alone would produce sufficient data for validation purposes.

Behavior coding is a widely-used pretest method in which interviews from the field are recorded and interviewer-respondent interactions are coded. The logic of behavior coding is that deviations from the ideal interaction whereby the interviewer reads the question exactly as worded and the respondent provides an acceptable answer are considered an indication of possible problems with the survey question. In general practice, a question is flagged as problematic if any particular type of problem occurs 15% or more of the times the question is administered. (Morton-Williams 1984, Fowler 1989, Zukerberg et al. 1995).

An oft-noted weakness of the method is that it does not provide explanations about likely causes of the identified problem. It may be, though, that this is the case because standard behavior coding procedures do not make use of

all the available data. In some interviewer-respondent interactions, additional comments or verbal exchanges occur that may provide clues about the nature of the question problems. For example, one type of problematic interaction is a respondent requesting clarification about a question's meaning. A comment of the form "What do you mean by [some part of the question]?" may help explain what difficulty the respondent is having. If this or other types of comments lead to further exchanges between the interviewer and respondent, more information about the question problem may be produced. This verbal data is not generally recorded as part of standard behavior coding. We investigated whether these verbal data used in conjunction with behaviour coding could be useful in detecting problems.

A second observation was based on a paper by Fowler and Cannell (1996) in which they show how standard behavior coding may identify some types of cognitive problems with survey questions. For example, a request for clarification might indicate an unclear term in the question. Respondent interruptions before the interviewer finishes reading the question may suggest that the question is worded in a manner that "results in a complete question before the question is finished" Fowler and Cannell, (p32). Their paper provides a useful expansion of the applicability of behavior coding as a pretesting method by addressing the standard claim that the method does not provide information about the nature of problems. And, although Fowler and Cannell do not comment on it, their result also suggests a use for behavior coding as a research tool. If cognitive problems identified in the laboratory are not effectively addressed and subsequently do occur in the field, they might be detected by behavior coding the field interviews. Our approach was almost the opposite of Fowler and Cannell's. They examined the standard set of behavior codes to come up with possible cognitive problems they might, in general, indicate. We examined a specific set of cognitive problems found in laboratory testing and predicted how they might manifest themselves in field interviews as interviewer-respondent interactions. This approach led to adding some codes to the set that is standardly used.

2. Study Design

The present study uses both standard behavior coding procedures as well as the actual interviewer-respondent verbal comments and exchanges to address two research questions.

1. To what extent do questions flagged as problematic in the cognitive laboratory actually exhibit problems in the field?
2. To what extent do the specific problems detected in cognitive testing occur in the field?

The research method involved selecting a set of questions identified as problematic during cognitive interviewing and collecting the two types of data during standard field administration to determine how often the laboratory-identified problems occurred in the field.

The questionnaire was designed for a general population telephone survey to identify both healthy and unhealthy respondents to estimate prevalence of a difficult-to-diagnose fatiguing illness. The questionnaire contained 61 questions. Eighteen cognitive interviews were conducted. Eleven of those respondents were recruited through local support groups for persons with the illness. The remaining seven respondents were recruited from the general population, using a screener to verify that they did not have symptoms of the targeted illness.

The cognitive interview protocol included instructions for respondents to think aloud as they answered the questions, and also contained scripted, question-specific probes, and generic probes. Across the 61 questions in the instrument, 82 individual problems were identified. In a report of the cognitive interview pretest, the identified problems were described, and recommendations were made for revising the questions to address these problems, which was done in many cases. However, some of the questions could not be revised because one survey goal was to compare results to other surveys in which those questions had been used. This subset of questions was not revised for the telephone survey. This presented an opportunity to observe the performance in the field of items that remained in the form in which they were tested. The results of the cognitive testing and subsequent actions are summarized in Figure 1. For each item in the questionnaire, either some types of problems were identified or not. Of those identified, some were revised and others remained in the form in which they were tested.

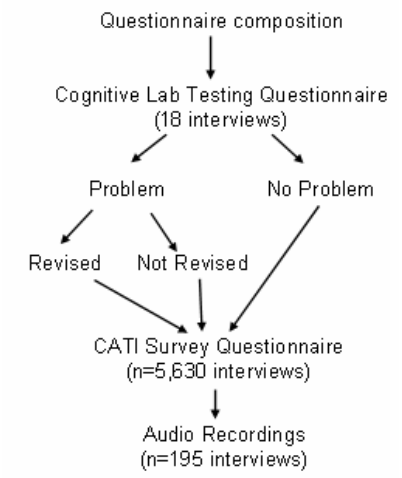

Figure 1: Study Design Flow Chart, Part 1

A subset of 29 of the problematic items was identified as likely to produce some interaction or other verbal evidence if the problem occurred during an interview. This meant, of course, that we faced a similar limitation as Willis and Schechter, in that only certain types of problems could be expected to produce the types of evidence that could be collected in a field setting. Therefore, the full range of laboratory problems was not validated in this research.

A combination of verbal exchanges and behavior codes (Table 1) were used to determine whether: a) a question flagged as problematic in the laboratory showed any evidence of a response problem in the field and b) a problem identified in the field was the same as the problem identified in the laboratory.

Table 1: Behavior Codes

Interviewer

- Reads exactly as worded
- Reads with minor changes
- Reads so that meaning is altered
- Has difficulty recording answer
- Gives inappropriate probe or interprets response inappropriately

Respondent

- Gives adequate/proper answer
- Qualifies answer
- Asks for clarification
- Answers "I don't know"
- Refuses to answer
- Gives inadequate/improper answer
- Interrupts question reading
- Asks for all or part of question to be repeated
- Says question is not applicable

The verbatim exchanges were systematically compared to the description of the problem in the cognitive interview pretest report to determine whether the field and laboratory problems matched each other. This procedure was applied to that set of 24 questions for which we judged the laboratory problem might be evident in the field interview. Four outcomes were possible for each of the 24 flagged questions in the potential 195 question administrations (Figure 2). First, no problem is evident. Second, a problem is evident but clearly is not the lab problem. Third, a problem is evident that clearly does match the lab problem. Fourth, there is evidence of a problem, but the evidence is not sufficient to determine whether it is the lab problem or not.

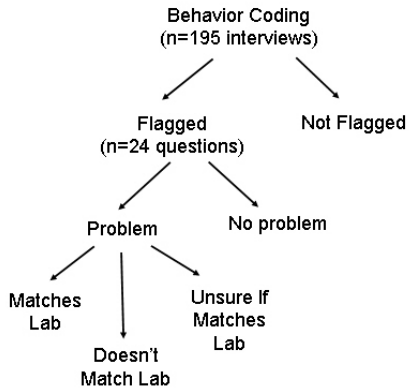

Figure 2: Study Design Flow Chart, Part 2

3. Findings

The unit of analysis is the question administration. The 24 questions could potentially have 4,680 (195 x 24) question administrations. In fact, there were 3,168 administrations of the selected set of items (Table 2). This difference is due to the fact that only portions of each interview were recorded. Sometimes the recorded segment did not contain particular flagged question. A total of 529 (17%) of these administrations showed evidence of problems. In nearly half these instances (47%), the combined behavior coding and verbatim evidence supported the conclusion that the field problem matched the problem identified in the lab.

Table 2: Results of Question Administrations (n=3,168)

	Number
Problematic Administrations	529 (100%)
Matched Lab Problem	247 (47%)
Didn't Match or Uncertain if Lab Problem	282 (53%)

The percentage of problematic question administrations varied greatly by item. Two items produced almost no evidence of problems, while one item had 60% problematic administrations (Figure 3). Eight items had less than 10% problem administrations and 13 items had evidence of problems over 10% of the time.

There was strong evidence that the specific laboratory-identified problems did, in fact, occur in the field (Figure 4). When examining the problematic administrations, for 21 of the 24 items, the field problem matched the lab problem more than 20% of the time.

4. Discussion

First, as to the research method, the combination of behavior coding with additional codes and verbatim interviewer or respondent comments and interactions was effective in identifying field problems and in determining whether the field problems matched those identified in the laboratory. This suggests the method would be useful for further validation studies, as well as for pretesting itself.

We have no way of assessing the impact of these problems on respondents' actual answers. So the impact

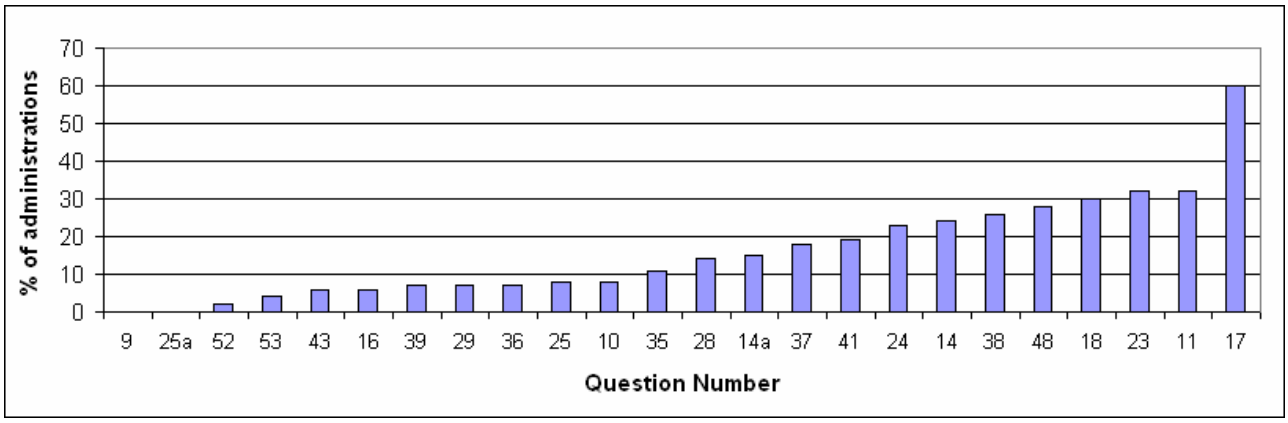

Figure 3: Percentage of Problematic Question Administrations

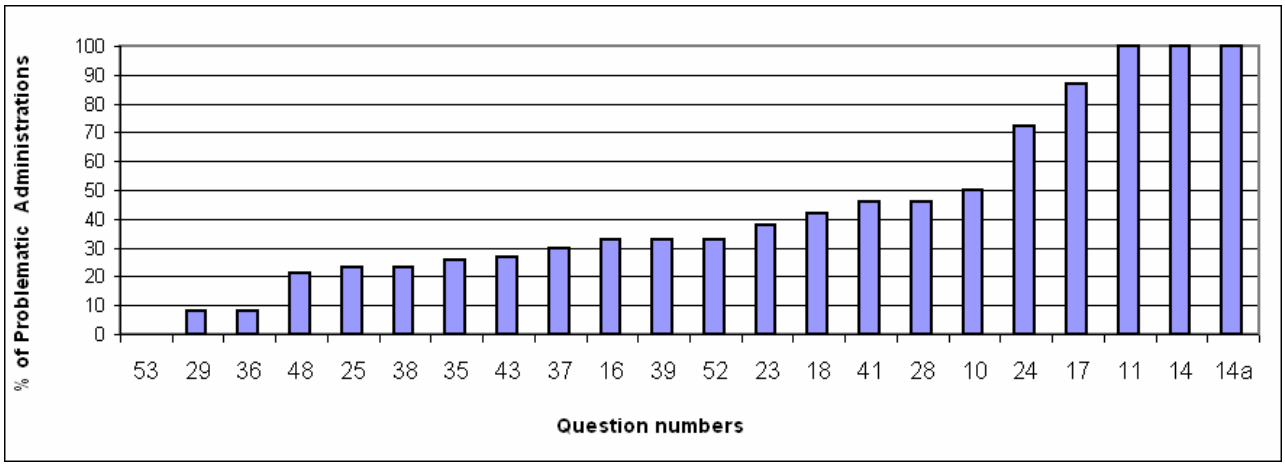

Figure 4: Percentage of Problematic Question Administrations That Matched Lab Problems

on measurement error cannot be determined. However, the percentage of problematic administrations (17%) suggests that these items may be cause for concern. In addition, it should be noted, that additional problems may have occurred that our method could not detect. So the 17% is an estimate—since it is based on only a sample of the 5,630 interviews—of the lower limit of problematic administrations. Moreover, in those problematic administrations the lab problem was verified almost half the time (47%). Again, this is a lower limit estimate, since there were instances where it could not be determined if the observed problem was the lab problem or not.

Additional planned analysis may provide further results. Because resource constraints limited the sample to less than 200 cases, sample size is an important limitation on the strength of any conclusions. The standard errors of these estimates need to be computed. It will also be useful to look at the nature of the problems that could be detected with this method. While the study provides additional grounds for the reliance on cognitive interviewing for pretesting, it also suggests that there is a lot yet to be learned about the method's validity. Further research is needed to validate cognitive pretest findings across a range of types of survey questions and kinds of question problems. Such research would also benefit from larger sample sizes. Blair et al. (2006), in a study of the impact of sample size on cognitive interview pretest results, found that substantial numbers of problems continued to be uncovered even after 20 or more cognitive interviews. The nature and prevalence of problems that required more extensive testing may differ from those more easily detected. The likelihood of these problems occurring in field administration may differ as well and should be included in further validation research.

References

Blair, J., Conrad, F., Ackermann, A., and Claxton, G. "The Effect of Sample Size on Cognitive Interview Findings," *Proceedings of the American Statistical Association,* 2006.

Fowler, J.F. Improving Survey Questions. U.S.A.: Sage Publications, 1989.

Fowler, J.F. and Cannell, C.F. "Using Behavior Coding to Identify Cognitive Problems with Survey Questions," *Answering Questions*, Schwarz, N. and Sudman, S., eds. San Francisco: Jossey-Bass, 1996.

Groves, R.M., Fowler, J.F., Couper, M., Lepkowski, J.M., Singer, E., and Tourangeau, R. *Survey Methodology*. New Jersey: Wiley and Sons, 2004.

Morton-Williams, J. and Sykes, W. "The Use of Interaction Coding and Follow-up Interviews to Investigate Comprehension of Survey Questions," *Journal of the Market Research Society, 26*, 109-127, 1984.

Willis, G.B. and Schecter, S. "Evaluation of Cognitive Interviewing Techniques: Do the Results Generalize to the Field?" *Bulletin de Methodologie Socialogique* (59 rue Pouchet, F 75017 Paris), June 1997, N. 55.

Zukerberg, A.L., Von Thurm, D.R., and Moore, J.C. "Practical Considerations in Sample Size Selection for Behavior Coding Pretests," *Proceedings of the Section on Survey Research Methods*, Joint Statistical Meetings, Alexandria, VA, 1995.