CAPI INSTRUMENT EVALUATION: BEHAVIOR CODING, TRACE FILES, AND USABILITY METHODS

James M. Lepkowski, Mick P. Couper, Sue Ellen Hansen, Wendy Landers, Katherine A. McGonagle, Jay Schlegel, Survey Research Center, University of Michigan; Fran Chevarley, National Center for Health Statistics
James M. Lepkowski, Survey Research Center, P.O. Box 1248, University of Michigan, Ann Arbor Mi 48106

Key Words: Interviewer effects, survey quality, questionnaire design

Introduction

The National Health Interview Survey (NHIS) has served as an essential element of the nation's health care statistics data collection system. The NHIS collects information each year from a probability sample of approximately 47,000 households containing 120,000 persons using personal interviews. The NHIS data are widely used by policy makers and others to study and chart the health of the U.S. population.

Increasing demand for data over the past decade has caused growth in the size and complexity of the NHIS. The length and inflexibility of the instrument made it difficult to modify. NCHS redesigned the NHIS over the period 1995-1997, with three principal changes: (1) increased sample sizes for important minority groups; (2) computer assisted personal interviewing (CAPI) replaced traditional paper-and-pencil data collection; and (3) the structure and content of the questionnaire was significantly changed. The sample design changes were implemented in 1995. CAPI programming was completed in 1996, and implementation followed in 1997.

NCHS required that the redesign impact on the quality of NHIS interviewing and data be assessed.. We report on the findings of three evaluations of the redesigned NHIS CAPI questionnaire: an analysis of trace files recorded by the CAPI software, coding of video tapes recorded in NHIS usability interviews, and coding of audio tapes of NHIS interviews conducted in the field. Two companion papers present an analysis of trace files from 1997 NHIS field interviews (Couper and Schlegel, 1998) and usability testing of the NHIS instrument in a laboratory setting (Hansen, Couper, and Fuchs, 1998). This paper presents methods and findings of an analysis of behaviors coded from audio taped interviews and compares the findings of the three methods, noting questions which the three methods identified jointly as well as uniquely. The paper also examines the characteristics of questions jointly identified by two or all three of the methods.

Beahvior Coding the Redesigned NHIS

Behavior coding (Fowler & Cannell, 1997; Mangione, Fowler, & Louis, 1992; Oksenberg, Cannell, & Kalton, 1991) is a technique which provides insight to the extent that survey questions tax the cognitive abilities of interviewers and respondents in personal interviews. Trained staff listen to audiotapes of survey interviews and code interviewer and respondent behaviors. These codes indicate the extent to which interviewers are seeking ways to clarify question wording and objectives or respondents find questions cognitively demanding.

A core set of 11 interaction codes which had been tested over a number of investigations (Cannell, et al. 1968; Cannell and Robinson, 1971; Mathiowetz and Cannell, 1980; Morton-Williams, 1979) were selected for the investigation of the redesigned NHIS instrument. These were supplemented with 16 experimental codes that had not been applied routinely in previous behavior coding activities. We reduced our coding scheme to those behaviors that were coded reliably (with inter-coder reliability values of kappa greater than 0.4; see below), examining seven core indicators and nine of the experimental codes, grouped into eight behavior summary indicators.

Indicators examined in this investigation consisted of three types. First, question-asking indicators are grouped under a single summary indicator, *major wording change*, in which the interviewer's reading or the question appeared to change the meaning of the question. Second, probing behavior is summarized by a single indicator, *failure to probe*, in which the interviewer does not probe at all or adequately enough to elicit a final codable answer. Third, 11 indicators of the quality of the answer provided by the respondent were grouped into six summary indicators. *Interrupts question* was assigned if the respondent interrupted with an answer during the reading of the question. *Multiple answers* occurred when the respondent gave more than one answer even though the question required a single answer. *Answer outside response frame* occurred when the respondent gave answers outside the options. Both these latter indicators were grouped as a single summary, *uncertain answer*. A *qualified answer* met the objectives of the question, but the respondent accompanied the answer with a qualifier such as "probably" or "about" or provides information that is not required by the question. *Definition request* and *repeat of question* were direct indicators of respondent initiated clarification of their task. *Interviewer intiated, respondent initiated*, and *think-aloud*

digressions indicated verbal responses which were not directed to answering the question. (These three indicators were grouped under the summary indicator *digression.*) A *don't know* response occured even though a final answer was obtained. These last 11 indicators were grouped under a summary *any respondent behavior.*

Thirteen persons with extensive interviewing experience were trained to code these indicators. They were given detailed instructions in two-day training sessions on how to apply the codes (see Blixt, et al. 1994). A total of 154 interviews consisting of 29,353 exchanges across 542 different screens were coded.

Eight interviews containing 1,178 exchanges were coded twice by two different coders. The Kappa statistic (Fleiss, 1981) was computed as a measure of inter-rater agreement. (Kappa is the ratio of the difference between the observed and expected levels of agreement to the proportion of agreement that is unexplained: $\kappa = (P_{obs} - P_{exp}) / (1 - P_{exp})$.) The reliability scores for most of the question-asking, probing, and respondent indicators fell in the fair to good range (kappa value greater than 0.4). The reliability of *failure to probe* and *multiple answers* (and hence the summary indicator *uncertain answer*) could not be measured on the small sample of exchanges coded twice. Coder reliability for *digressions* was in the poor range (0.30), but results are presented for *digressions* for the sake of completeness.

Table 1 presents the relative frequency of the core, supplemental, and summary indicators. Nearly one in three exchanges involved a major wording change. Half of those changes were due to fill errors, skipping the reading of a question, or incompletely reading a displayed fill in the CAPI instrument. In one exchange of 25 the respondent interrupted the question reading, while uncertain answers were given in nearly one in 20 exchanges. Respondents gave qualified answers in another one in 14 exchanges, a sign of difficulty formulating an answer. Don't know responses do not occur frequently, although there were questions that exhibited higher frequency. Some type of respondent behavior (including interviewer initiated digression) occurred in nearly one in five exchanges.

Percentage frequency of core, supplemental, and summary indicators were computed separately for each of 542 NHIS questions which appeared one or more times in the 154 interviews. Many of the questions were asked in very few exchanges, due to questionnaire skip instructions. In order to have reliable estimates of the percentage of times a behavior occurred for a question, 271 questions which had 25 or more exchanges were chosen for further review.

For the purposes of identifying questions that may be posing difficulty for the interviewer or the respondent, the questions were examined based on eight summary or core indicators: *major wording change, failure to probe, qualified answers, don't know* responses, *interrupts question reading, uncertain answer, digression,* and *any respondent behavior.* Six of the eight indicators concern respondent difficulty with the question, chosen because previous investigation (Belli and Lepkowski, 1996; Dykema *et al.*, 1997, Lepkowski *et al.*, forthcoming) suggested that respondent behaviors are more often associated with less accurate reporting than interviewer behavior.

Table 1. Frequency of NHIS Question-Level Behaviors for 29,253 Exchanges (154 Interviews)

Question-Asking	
Major wording change	31.2%
Incompletely read	7.8
Added word(s)	3.1
Deleted word(s)	4.2
Emphasis error	0.7
Fill error	16.7
Probing	
Failure to probe	0.8
Response	
Interrupts reading	4.0
Uncertain answer	4.7
Multiple answers	0.3
Outside response frame	4.3
Qualified answer	7.2
Digression	3.5
Interviewer initiated	2.2
Respondent initiated	1.5
Think aloud	2.1
Don't know	1.1
Any respondent behavior	18.6
Refuse to answer	0.2
Definition request	0.7
Repeat of question	2.8

A subset of questions which posed the greatest difficulty for interviewers or respondents were identified by an arbitrary criteria. Following follow Blixt *et al.* (1994) a standardized score (proportion with behavior minus average proportion across all 271 questions divided by the standard deviation of proportions across questions) was computed for each question. Scores greater than 2.0 for a single question on any indicator would have identified approximately two percent of questions for further review (more than 70 questions). This rather large numberwas reduced by selecting questions with standardized scores greater than 2.0 for two or more of the eight target indicators. A total of 24 questions met this criterion.

While the target indicators and frequencies suggest the nature of the interviewer or respondent difficulty with the question, question text and coder comments were examined to diagnose the source of the difficulty. For example, the question ADENLONG has a high frequency of interrupted question reading, and higher frequencies of qualified answers which contribute to a high frequency for the summary indicator, *any respondent behavior*. The reason for the interruptions in the question wording is clear when the text of the question is examined: *About how long has it been since you last saw or talked to a dentist? Include all types of dentists, such as orthodontists, oral surgeons, and all other dental specialists, as well as dental hygenists.* Respondents interrupted the reading because they assume that the question has been completed after the first sentence. They do not realize that further instructions come next. Answers may be based on an incomplete consideration of the types of providers that should be counted under "dental visits". An obvious remedy for this type of question problem is to provide the instruction first, giving the respondent an opportunity to consider all types of providers before formulating an answer.

ADENLONG also had a relatively high frequency of qualified answers. The respondent was given a card with the following response options:

(1) 6 months or less
(2) More than 6 months, but not more than 1 year ago
(3) More than 1 year, but not more than 3 years ago
(4) More than 3 years ago
(5) Never

The response options are straightforward time categories. Coder comments indicate that respondentsexpressed uncertainty about the category they chose, using modifiers such as "I think " or "It was probably ".

Comparison of Methods

Behavior coding can be a useful tool for the diagnosis of problems in questions that are being administered in a survey. Indicators highlight problems with wording of questions, the cognitive task expected of the respondent and the interviewer, and other aspects of the questionnaire design.

Other methods of identifying difficulties in a CAPI instrument may be used. Two were applied to the same NHIS instrument at the same time, an analysis of trace files from the CAPI instument (Couper and Schlegel, 1998) and a usability evaluation (Hansen, Couper, and Fuchs, 1998). The goals of these alternative methods were, in some cases, overlapping and, in others, distinct from those for behavior coding.

Trace files analysis examines the use of the function keys of the CAPI instrument itself, summarizing the history of the interviewer's interaction with the computer. Trace files are summary records of the functions used and answers recorded by an interviewer during the course of an interview. Function key use could overlap with some aspects of usability evaluation, but less often with behavior coding.

Usability evaluation is conducted in a laboratory where interviews can be directly observed by third parties to the interview and the interaction can be recorded on video tape. Video tapes provide a visual and audio record of interviewer, respondent, and computer behavior. Usability evaluation of the NHIS instrument addressed instrument design and functionality, identifying questions for which interviewers had difficulty interacting with the computer. Visual cues obtained from video taped interviews increased the sensitivity of the method (relative to behavior coding) for detecting respondent or interviewer difficulty with wording or the cognitive task. The usability evaluation data generated codes about two different types of indicators, behavioral and event. The usability behavior codes are similar to, and often a subset of, the behavior coding indicators.

Figure 1 is a depiction of the nature of the interviewer-respondent interaction in the CAPI setting. There are three elements in the interaction: the two human subjects (the interviewer and the respondent) and the CAPI application on the computer. All three have an impact on the quality of the interview results. Difficulty in the interviewer-respondent interaction can detract from (such as, for example, repeated digressions) or emphasize (such as, for example, the presentation of instructions to the respondent) the task. There is some expectation, however, that the nature of the interviewer-computer interaction can effect the interviewer-respondent interaction. For example, interviewers who have difficulty with function key placement for a particular question, or with distinguishing the question to be read to the respondent from the extraneous information on the computer screen, can become distracted from the question

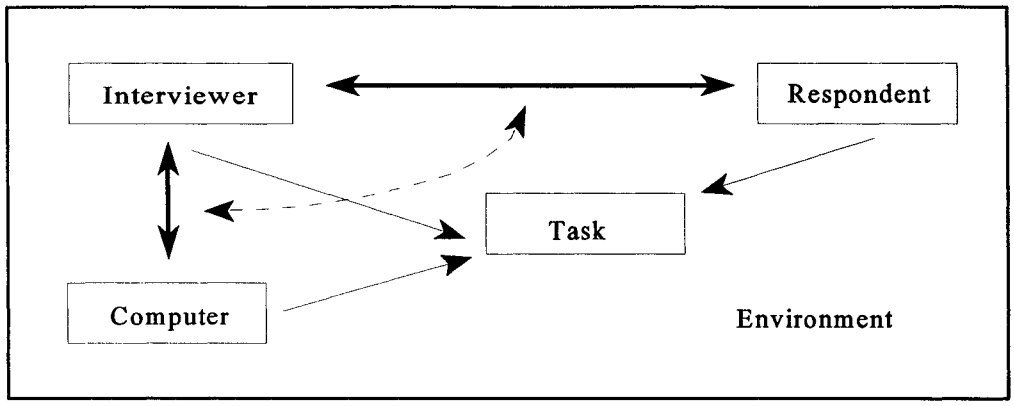

Figure 1. Interviewer, Respondent, and Computer Interactions and the Interviewing Task

asking and probing task. The distraction may decrease the quality of the interaction with the respondent because they are not concentrating on the interview. At the same time, respondents who digress may distract interviewers from the important task of entering data correctly and completely on the computer. The interviewer's task is to manage these two interactions (with the computer and with the respondent) simultaneously while keeping focused on the task of collecting accurate data.

While behavior coding can be used to examine the nature of the interviewer-respoondent interaction, itis not well suited to investigations of interviewer-computer interactions. It is difficult to code reliably indicators of this latter interaction based entirely on audio signals.

Trace files are useful for investigation of the "input" side of the interviewer-computer interaction in which interviewers attempt to record information from the interview. The "output" side, what the computer presents to the interviewer and how the interviewer deals with that information, is not contained in trace files. Trace files provide an indication of the extent to which interviewer-computer interaction is impeded by the design of the instrument. For instance, a change in the use of function keys for a particular question may confuse the interviewer, resulting in the repeated use of an incorrect key.

Usability evaluation can be used to examine both the interviewer-computer and the interviewer-respondent interactions simultaneously. Visual evidence provides a more complete indication of the nature of the interviewer-respondent interaction than audio alone. Usability evaluation also provides visual evidence of the use of functions and features of the computer, since a complete picture of the use of the keyboard as well as of the information entered is obtained. For example, a trace file analysis may show that a given function key was used, but it does not indicate what other keys (besides function keys) may have been incorrectly used. Such additional information can possibly identify placement problems for function keys (see Hansen, Couper, and Fuchs, 1998). Usability evaluation can be particularly useful for identifying where the interviewer's attention is focused, and which of the interviewer-computer or interviewer-respondent interactions is dominant at any time.

All three methods were applied to the same survey. Behavior coding was conducted on 154 interviews covering 542 unique screens. Behavior coding scores identified 24 questions which posed a difficulty in the interviewer-respondent interaction. Trace file analysis was conducted on more than 16,000 field interviews containing some 418 different screens (screens with interviewer instructions only were not included in the trace files analysis). Trace file analysis methods identified 51 screens which posed difficulty in the interviewer-computer interaction. Usability evaluation was conducted on 38 laboratory interviews with 475 unique screens. Usability evaluation identified 35 total screens for which there was either a difficulty for the interviewer-computer interaction or the interviewer-respondent interaction. Seventeen of these screens were uniquely identified by event coding, 11 uniquely by usability evaluation behavior coding, and eight by both event and behavior coding.

Across all three methods, a total of 86 questions (screens) were identified as having difficulty in either interviewer-respondent or interviewer-computer interactions. Four (4.7%) of these were identified by all three methods. Another 19 (22.1%) were identified by two methods, while the remaining 63 (73.2%) screens were identified by only one method. Among methods, four of 24 behavior coding questions, or 16.6%, were identified by both of the other methods, a rate nearly identical to that of usability event coding (four of 25 or 4.0%). Trace files had the least complete overlap with only four among 51 or 7.8%.

Another 13 questions identified by behavior coding, or 54.2%, were also identified by either trace files analysis or usability event coding, with the greatest "pairwise" overlap coming with trace file analysis. For usability event coding, the "pairwise" overlap is also high, 10 of 25 or 40%. Again, trace files analysis had the least overlap in pairwise comparisons, 11 of 53 or 20.8%. Of course, these comparisons are somewhat misleading since trace files analysis generated more total screens for comparison than either of the other two methods. Still, it appears that trace files generate a different set of difficult screens than either behavior coding or usability event coding. Behavior coding and usability event coding overlap as often as behavior coding and trace files analysis, while both of these pairs overlap more than trace files analysis and event coding.

Some of the overlap between methods could possibly be due to screen characteristics such as the presence of instructions to the interviewer, multiple item questions on a screen, open or closed question format, and the use of hand cards. Screen characteristics were examined for each of the 86 questions identified across the three evaluation methods. The screens were grouped by functional purpose such as household enumeration, checkpoints, questions seeking frequency or time estimation, questions concerning limitations of activity, questions about the sample child, questions about income, questions about health insurance, questions about injuries, questions about physical activities, recontact information, and other types of questions. There was no discernable pattern in overlap across these groups. The four questions which all three methods identified difficulty are in four different groups. Given the small number of questions spread across the 11 groups, it is not surprising that no particular pattern of "pairwise" overlap was found either.

Since nalysis of overlapped questions revealed no discernable patterns, it is useful to examine whether there are certain characteristics of the questions themselves that are uniquely identified by each method. Preliminary analysis (detailed coding of question characteristics is in progress) suggests that question characteristics are uniquely related to evaluation method. That is, question characteristics do not provide a strong indication of three way or pairwise overlap. For example, the four questions identified by all three methods have 11 different question characteristics among them. The presence of header information, interviewer instructions, and help indicators, or use of multiple items on a screen were present for three of these four screens. The remaining seven characteristics occur singly for a question, except for text enhancements which appears twice among the four. The nature of the four characteristics that occur most often in four overlapped screens suggest that these four were anticipated by questionnaire developers and CAPI instrument designers to be more complex since they added information to the question text and response options. A more complete anlaysis will examine the extent to which the three way and two way method overlap is likely based on these questions characteristics.

Discussion

Three methods of evaluation of the NHIS CAPI instrument have been illustrated in this and companion papers. Each method focuses on a different aspect of the interviewer-computer-respondent interaction. It might thus be expected that the methods should be expected to overlap very little with one another in identifying questions that pose difficulty in the interview. However, there may be problems which arise during an interviewer-computer or interviewer-respondent interaction which "spill over" to the other domain. For example, a question may pose a particularly difficult task for a respondent, which leads to a complex interviewer-respondent interaction. In addition, the complex interaction from this question may lead to increased difficulty for the interviewer during interaction with the computer.

The evidence indicates that the methods do not overlap. Behavior coding has the greatest overlap with the other two methods, but trace files analysis and usability evaluation identify essentially unique screen problems. The overlap remains incompletely explored in the present investigation, and analyses of the types of questions for which overlap occurs continues. Preliminary review suggests that questions with greater complexity reflected in more frequent use of instructions and multiple items may be identified by all three methods more often, but further coding and analysis remains to be completed.

Among the three methods, usability evaluation may appear to be more expensive and time consuming than the other two methods. On the other hand, usability evaluation yields rich data on both interviewer-computer and interviewer-respondent interaction, something neither of the other two methods can do as effectively. Behavior coding is less expensive, although it is time consuming for data collection staff to administer. Behavior coding focuses on the interviewer-respondent interaction, collecting useful information about interviewer and respondent behavior that can reflect indirectly usability and design problems in an instrument. Trace files are the cheapest data to collect among the three methods, yielding data on the entry task of the interviewer. Extension of trace files to keystroke files and item level time stamps increase the utility of the method. Trace files can be a useful supplement to the other two methods,

bringing greater insight into the problems of the interviewer-computer interaction.

The present investigation indicates that each of these methods can make a unique and valuable contribution to the evaluation of a survey instrument. It is unlikely that a survey organization will have time or resources to apply all three on a given survey, though. Instead, all three methods can be used simultaneously in a laboratory setting. A usability laboratory allows the collection of data that can be used for behavior coding and for trace files analysis. While a laboratory setting is artificial, and may not uncover certain kinds of behavioral problems that arise in field administration, it may be the most useful and complete method for collecting a full set of indicators on the interviewer-respondent and interviewer-computer interaction. While a laboratory evaluation is a compromise solution, allowing collection of data on all interactions at one time, but in an artificial setting, it may provide the most cost effective means for survey organizations to conduct a thorough review of the properties of a survey instrument.

(Support for research was received from the National Center for Health Statistics under Cooperative Agreement S278-15/15 and from the Survey Research Center at the University of Michigan.)

References

Belli, Robert F. and James M. Lepkowski. "Behavior of Survey Actors and the Accuracy of Response," *Proceedings of the Sixth Health Survey Methods Research Conference.* Washington, D.C.: Agency for Health Care Policy and Research, 1996, pp. 69-74.

Blixt, S., Lepkowski, J.M., Belli, R.F., Cannell, C., Giamalva, L., and Buckmaster, J., "Behavior Coding Results for the Health Field Study," Research Report, Survey Research Center, The University of Michigan, 1994.

Cannell, C.F., and Robison, S., "Analysis on Individual Questions," in L. Lansing, S. Withey, and A. Wolfe, (eds.), *Working Papers on Survey Research in Poverty Areas*, Chapter 11, Ann Arbor, MI.: Institute for Social Research, The University of Michigan, 1971.

Cannell, C.F., Fowler, F.J., and Marquis, K.H., "The Influence of Interviewer and Respondent Psychological and Behavioral Variables on the Reporting in Household Interviews," *Vital and Health Statistics*, Series 2, No. 26, Washington, DC: U.S. Government Printing Office, 1968.

Couper, Mick P., and Jay Schlegel, "Evaluating the NHIS CAPI Instrument Using Trace Files," paper presented at the American Association for Public Opinion Research meetings, St. Louis, MO, May 17, 1998.

Dykema, Jennifer, James M. Lepkowski, and Steven Blixt. "The Impact of Interviewer and Respondent Behavior on Data Quality: Analysis of Interaction Coding in a Validation Study," Chapter 12 in *Survey Measurement and Process Quality*, Martin Collins *et al.*, editors. New York: Wiley and Sons, Inc., 1997.

Fleiss, J.L., *Statistical Methods for Rates and Proportions*. 2nd edition, New York: Wiley, 1981.

Fowler, F. J., & Cannell, C. F. Using behavior coding to identify cognitive problems with survey questions. In N. Schwarz and S. Sudman (Eds.), *Methods of determining processes used to answer questions.* San Francisco: Jossey-Bass, 1997.

Hansen, Sue Ellen, Mick P. Couper, and Marek Fuchs, "Usability Evaluation of the NHIS Instrument," paper presented at the American Association for Public Opinion Research meetings, St. Louis, MO, May 17, 1998.

Lepkowski, James M., Sally A. Sadosky, and Paul S. Weiss. "Mode, Behavior, and Data Recording Accuracy," *Computer Assisted Survey Information Collection*, Mick P. Couper *et al.*, editors. New York: Wiley and Sons, Inc. (forthcoming, 1999).

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

Mathiowetz, N.A., and Cannell, C.F., "Coding Interviewer Behavior as a Method of Evaluating Performance," *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 1980, pp. 525-528.

Morton-Williams, J., "The Use of 'Verbal Interaction Coding' for Evaluating a Questionnaire," *Quality and Quantity*, Vol. 13, 1979, pp. 59-75.

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