EVALUATING THE NHIS CAPI INSTRUMENT USING TRACE FILES

Mick P. Couper and Jay Schlegel, University of Michigan
Mick P. Couper, Survey Research Center, P.O. Box 1248, University of Michigan, Ann Arbor Mi 48106

Key Words: CAPI; Trace files; Instrument evaluation

1. Introduction

Computer assisted interviewing (CAI) has brought fundamental changes to the process of survey data collection (see Couper and Nicholls, 1998). This is especially true for computer assisted personal interviewing (CAPI) in which the interviewer uses a laptop computer to conduct interviews in respondents' homes, and hence is removed from the immediate support of technical staff. Survey instruments have increased in complexity and in the number of actions and operations required by the interviewer. Effective design of the CAPI instrument is critical, and tools to assist in the evaluation of CAPI instruments are needed.

Trace files or keystroke files are automatic byproducts of many CAI software systems. While the primary purpose of such files are for diagnosis and debugging of instrument errors during development and testing, they can also be used to evaluate other aspects of the survey process, such as interviewers or the instrument. These data have previously been used to evaluate interviewer performance in CAPI (Couper, Hansen, and Sadosky, 1997) and respondent performance in CASI (Caspar and Couper, 1997). In this paper, we focus on the use of trace files in the evaluation of a CAPI survey instrument for the National Health Interview Survey (NHIS).

Trace files vary in the level of detail they provide. In transaction-based CAI systems (e.g., Surveycraft), every key that is pressed by the interviewer is recorded in the keystroke file. This includes those keys that have no function, or produce no reaction from the computer. In contrast, execution-based systems (e.g., UC Berkeley's CASES system) capture only those entries that are executed by the system. In other words, if an unmapped function key was pressed, or if several changes were made to a text entry before submitting it, these key presses would not appear in the trace file. Most CAI systems capture item identifiers as part of the trace file, permitting item- or screen-level analyses as presented here.

We include an example of a CASES trace file segment in Figure 1. In this example, the interviewer entered a last name of "JOHN," realized on the following screen it should be "JOHNS," used the backup function to return to last name and typed "S." However, the default CASES approach is to overwrite rather than append to existing text entries, requiring the interviewer to backup again and enter the full name.

Figure 1. Example of CASES Trace File Segment

```
MORPER@       : 5::an:1
     : 5::db:[goto MORCK]
NEXTNM@FNAME    : 6::an:PETER
NEXTNM@MNAME    : 6::an:M
NEXTNM@LNAME    : 6::an:JOHN
NXTLIV@       : 6::an:1
     : 6::db:[goto NXTLIV@1]
NXTSEX@       : 6::co:b
NXTLIV@       : 6::co:b
NEXTNM@LNAME    : 6::an:S
     : 6::db:[goto NXTLIV@1]
NXTLIV@       : 6::co:b
NEXTNM@LNAME    : 6::an:JOHNS
```

Trace files also have a number of limitations for instrument evaluation. They capture only one part of the interaction in a CAI interview—that between the interviewer and computer. Even this is incomplete—we only see completed actions, not intentions, or failed attempts. We also have no information on functions interviewers should have used but did not.

Despite these drawbacks, a key benefit of trace files is that they are virtually costless to collect, and are available almost instantaneously, making them very useful for evaluation of the instrument during pretesting, or as a source of relatively quick feedback on the effectiveness of interviewer training. In addition, they can supplement methods such as usability testing (see Hansen, Couper and Fuchs, 1998) and behavior coding (Lepkowski et al., 1998). While these latter methods are rich data sources for understanding interviewer, respondent or instrument difficulties, they are expensive and time-consuming to collect. Trace file analysis is most usefully done in combination with one of these other methods to identify potentially problematic items or screens in a CAI instrument.

2. Data and Analysis

The data we examine are from the NHIS, an ongoing survey of health-related issues in the United States, collected by the Bureau of the Census on behalf of the National Center for Health Statistics (NCHS). The survey has undergone a three-phase redesign over the

past few years as it converted from PAPI to CAPI. The CAPI instrument is programmed using version 4.2 of the CASES software. Phase I of the redesign, conducted during the first six months of 1996, involved a small number of interviewers (16) using the CAPI instrument. In Phase II, during the second half of 1996, all NHIS interviewers conducted about half their sample cases using CAPI, while the balance was done using PAPI. Finally, in Phase III, beginning in January 1997, the entire NHIS sample was switched to CAPI, with no more paper interviews being conducted. The trace file data are from Phase II of the CAPI implementation We have trace files from over 16,000 completed interviews, containing over three million unique screen occurrences. This represents an average of about 182 completed questions or items per interview.

For this paper, we have restricted the analyses to those CAPI screens common to the usability test and behavior coding reported on in this session. There are 418 screens in the trace file dataset which also occur in the behavior coding and usability data, and we focus our analyses on these. Screens may contain several items, and the same screen may appear several times in the same interview (once for every household member, for example).

This means that the number of times a screen appears varies. Some screens, based on complex skips, are relatively rare, while others may occur several times in each interview. In order to account for this variation, we examine the ratio of the occurrence of various events on a particular screen to the total number of times the screen appears in the dataset.

3. Results

There is considerable variation in how often different functions are used. For example, 74% of all interviews have at least one use of the backup function [F1], and on average the function is used almost 14 times per interview. In contrast, the two jumpback functions, [F4] and [F5], each occur in less than 3% of interviews. We focus here on three types of functions: backing up from screen to screen, use of online help, and recording of interviewer notes.

3.1 Backups

We first identified the screens that were the most frequent targets of backups. The screen with the highest ratio of backups was HIKIND, with 3,353 backups in 7,892 occurrences of the screen, for a ratio of 0.42. In other words, almost once in every two times this screen appeared in the survey, it was the target of a backup action. HIKIND is a multiple response question on the kinds of health insurance members of the family have. The format of this item has changed over the course of Phase II data collection. In the first half of this period, the screen appeared as in Figure 2. This format was consistent with other multiple response items in the NHIS, in that a number was entered for each different response, and an "N" entered when no more options were selected.

Figure 2. HIKIND: Old Version

```
Item: HIKIND@1

                              Subject:    WILMA FLINTSTONE
                                          Respondent: WILMA  FLINTSTONE
      What kind of health insurance or health care coverage do
      you have?  MARK ALL THAT APPLY. (Anything else?)

      ENTER (N) FOR NO MORE AFTER THE LAST TYPE.       (H)
      (1)  Private Health insurance plan from employment
      (2)  Private Health insurance plan purchased directly
      (3)  Medicare
      (4)  Medi-Gap
      (5)  Medicaid
      (6)  Military Health Care/VA
      (7)  CHAMPUS/TRICARE/CHAMP-VA
      (8)  Indian Health Service
      (9)  State-sponsored health plan
      (10) Other government program
```

However, in the second half of Phase II, the HIKIND screen was changed to that illustrated in Figure 3. Here the interviewer presses [Enter] at each item not selected to move down the column, then places an "X" alongside the selected items. The interviewer can use [F1] or the up arrow key to move back up through the answered items, but cannot proceed forward with the down arrow. In order to remove an unwanted X, the interviewer must use the backspace key or [F6] to clear the entry. In introducing this change, the HIKIND screen became inconsistent with the many other multiple response screens in the NHIS instrument.

Figure 3. HIKIND: New Version

```
Caseid: 00011022
Item: HIKIND@a

                                  Subject:    FRED FLINSTONE
                                  Respondent: FRED FLINSTONE
    What kind of health insurance or health care coverage do
    you have?  EXCLUDE private plans that only provide
    extra cash while hospitalized or pay for only one type of
    service (nursing home care, accidents, or dental care).
    FR: MARK "X" ALL THAT APPLY.  (Anything else?)
                                                          (H)
    (1)  Private health insurance plan from employer or workplace
    (2)  Private health insurance plan purchased directly
    (3)  Medicare
    (4)  Medi-Gap
    (5)  Medicaid
    (6)  Military health care/VA
    (7)  CHAMPUS/TRICARE/CHAMP-VA
    (8)  Indian Health Service
    (9)  State-sponsored health plan
   (10)  Other government program
```

Nearly half of the backups on this screen (45%) are movement within the screen. Detailed analysis of these within-screen backups reveal that in 61% of these cases, the interviewers are changing a number or a blank to an X. In other words, it appears that interviewers are attempting to enter information on this screen in a manner

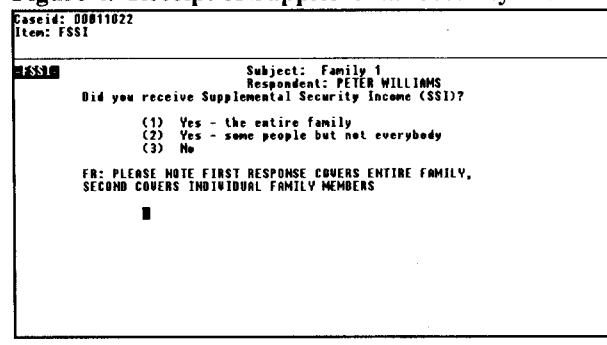
consistent with the other multiple response items in the instrument, but inappropriate for HIKIND. Furthermore, these results may underestimate the extent of this problem, as such backups are likely to occur predominantly in the second half of Phase II after the design change was implemented.

There are two lessons for instrument design from this example. First, changing the design of questions in the middle of an instrument does lead to interviewer difficulties. When introduced, such design changes should be applied consistently across all similar screens in the instrument, and interviewers need to be alerted to such global design changes. Second, the answers to the HIKIND question appear to serve as useful contextual information that could be a valuable information display for interviewers on later related questions. Many interviewers appear to be backing up to this item to review the information collected, without changing anything.

Another item with a high ratio of backups (about 13 backups for every 100 screen occurrences) is one which collects date of birth for each member of the household (DOB). A large proportion (45%) of the backups to this screen are from an age verification screen (AGEVER) that follows it. Once date of birth is provided, the computer calculates the respondent's age and feeds it back for confirmation in AGEVER. However, if the displayed age is not correct, the system does not go back to DOB for correction of date of birth. Instead, the instrument takes the interviewer to a screen (AGEGES) where he/she is asked to estimate the person's age. Many interviewers apparently realize on reading the age in AGEVER that they had miskeyed the year of birth, or the respondent corrects them directly ("No, my age is *x*, not *y*"). In other words, the interviewer knows the correct year of birth and simply returns to the previous screen (DOB) to correct the information, rather than proceeding to record an estimate. This appears to be an example of interviewers making sure they record the correct information in the initial question, rather than relying on estimation to change the response later. This suggests placing these two items on the same screen to permit easy correction. (Subsequent versions of the NHIS now ask respondents for both their age and their date of birth on a single screen, and resolve any inconsistencies on a following screen.)

Another item with a relatively large number of backups, also about 13 for every 100 times the screen is used, is FSSI (see Figure 4). This item is the fifth in a series on different sources of income. The previous four items in the series all have two response options: (1) Yes and (2) No. The fifth item (FSSI) changes format, with three response options: (1) Yes, the entire family, (2) Yes, some people but not everybody, and (3) No. We found that 87% of the backups to FSSI are from a followup question asked only of those who answered (2) to FSSI. Of these, 97% returned to FSSI and changed the answer from (2) to (3).

Figure 4. Receipt of Supplemental Security Income



Changing the pattern of response options in this series of questions violates a key design principle of consistency and produces the problem seen in the trace files. The note to interviewers attempts to reduce the problem by alerting interviewers to the inconsistency, but certainly does not eliminate the error. Furthermore, it is possible that errors such as this may go undetected if the follow-up question was not closely linked to the item in question. In other words, we may be underestimating the incidence of this type of error.

Several of the screens with higher frequency of backups contain multiple questions, while others are single questions with multiple responses (e.g., HIKIND). Table 1 shows the number of screens and average ratio of backups to total screen occurrences for all screens of each of these types. Multiple-item screens or forms have significantly ($p<.01$) more backups on average than single-item screens. Similarly, multiple-response items produce significantly more backups than single-response items.

Table 1. Mean Backup Ratios by Question Type

	Number of screens	Mean backup ratio
Multiple-item screen (form)		
yes	85	0.059
no	333	0.037
Multiple-response item		
yes	45	0.065
no	373	0.039

One reason for putting a group of related items on a single screen is that it facilitates navigation and correction of such items. This finding suggests that interviewers are availing themselves of the opportunity to navigate around

forms and correct answers on the same screen. In multiple-response items, it is possible that interviewers may select the same response more than once, requiring a correction. Alternatively, respondents may be changing their minds more on these types of questions, or providing answers in a different order than that presented on the screen.

The use of [F1] to backup is one way for interviewers to navigate around the instrument. On forms, or multiple-item screens, interviewers can also use cursor keys (up, down, left, right, home, end, page up, page down). In the NHIS instrument both the up and left arrows are mapped to the [previous field] function, while the right and down arrows are mapped to the [next field] function. The items with high numbers of backups also tend to have a lot of cursor movement. This suggests that on multiple-item screens where both [F1] and cursor keys can be used to move around the screen, both strategies are used by interviewers. Further, it suggests that both backups and cursor movement are indicators of potentially problematic screens. Either respondents are changing their minds, or interviewers are making errors that require backups and corrections.

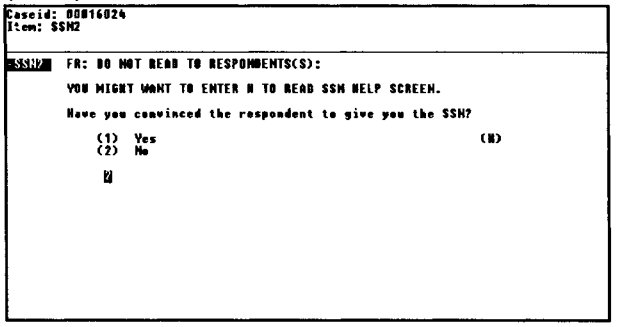
3.2 Help Screens

Online help is rarely accessed in the NHIS, with only 9% of all interviews having any help access, and an average use of 14 times per 100 interviews. Not every screen in the NHIS instrument has a corresponding help screen. Only 241 of the 418 screens we examined have an associated help screen. Of these 241, only 124 (51.5%) help screens were ever accessed during Phase II data collection, and only 78 (32%) were accessed more than once. The relatively infrequent use of online help parallels findings by Baker (1992), Couper, Sadosky, and Hansen (1995) and Sperry et al. (1998). This suggests that the infrequent use of help screens is not unique to the NHIS — the use of online help is apparently rare across systems and surveys.

The screen with the most use of online help is one on which interviewers are expected to record the two-character abbreviation for the respondent's state of birth (USBORN). If they do not know this, the help screen presents a list of the states and their associated codes.

The next most frequent help item is for SSN2, illustrated in Figure 5. It follows a screen (SSN) on which respondents are asked to provide a social security number. SSN2 is reached only if the response is "refused" or "don't know" on SSN. The help screen provides additional justification for the collection of SSN. It is interesting to note that the same help screen can also be accessed from SSN, but was never accessed from there. It seems that if the answer entered on SSN is "don't know" or "refused," the interviewer has failed to obtain the information. The SSN2 screen may thus appear unnecessary to interviewers, and we speculate some are looking to the help screen for guidance on how to complete this item.

Figure 5. Social Security Number Followup Screen (SSN2)



Several of the most frequently accessed help items provide further definitions for questions. These include MOD and VIG (definitions of moderate to light, and vigorous physical activity, respectively), FSSI (definition of supplemental security income), and FSPEDEIS (definition of special education or early intervention services). NATOR and RACE, questions about Hispanic origin or ancestry and race respectively, permit multiple responses. Three of the help items are interviewer checkpoints.

Like USBORN, PLBORN provides several screens for country of birth for those not born in the U.S. Neither of these lists (of states and countries) are ordered alphabetically (for example Iran is on the first help screen for PLBORN, while Iraq is on the third). We suspect that some of the help use may be interviewers trying to find the appropriate code on the help screen.

3.3 Interviewer Notes

Another indicator of potentially problematic items may be those on which interviewers enter notes, using the [F7] key in the NHIS. Like help access, the use of question-specific notes is relatively rare in the NHIS, appearing in about 9% of interviews. Some of the items identified as problematic in terms of backups (e.g., HIKIND) and help access (e.g., SSN2) also show relatively high levels of note use. While we do not know the content of the notes from the trace files, the prevalence of notes on particular items may again point to potentially problematic items that deserve further investigation. The most frequent use of item-specific notes occurs for RPAGEGES (4 times for every 100 screen occurrences). This item is similar to AGEGES mentioned earlier, but refers to the respondent rather than

other household members. This may suggest interviewers are explaining the reason for the age discrepancy.

It is odd that 59 times interviewers have used [F7] on the INOTES_2 screen. This is a screen to review and add additional interview-level notes, simply by pressing [1]. It may suggest interviewers are confusing the use of item-level versus interview-level notes. Examining the contents of the notes may reveal further information about the reason for notes on these particular screens.

4. Summary and Discussion

In this paper we have examined the utility of trace file analysis for identifying potentially problematic items in a CAPI instrument. These analyses, complemented by an in-depth investigation of particular items, identified several problems in the NHIS instrument.

Examining items that are the target of an unusually high frequency of backups reveal several sources of potential problems. One is a change in the entry format across items. Some interviewers fail to notice the change in format, and attempt to enter responses in similar fashion to previous items, necessitating corrections. Similarly, an inconsistency in the numeric labels assigned to response options produced a relatively large number of backups to correct an incorrect input. Both of these examples suggest changes to the survey instrument to reduce the incidence of this type of trace file activity.

By examining the items on which help screen usage or the entry of interviewer notes occur, trace file analysis can help identify items that are candidates for revision or at least may be targeted for additional testing.

While the information provided by CASES trace files is somewhat limited, we believe that trace file analysis is a useful complement to other methods for identifying potentially problematic items in an automated instrument. By themselves, trace files do not reveal the cause of a problem, but they allow one to focus in on specific questions that generate unusually high frequencies of CAI function use. If used in combination with usability testing and behavior coding, trace files can be especially helpful in confirming the prevalence of certain problems identified using the other methods. Usability tests and behavior coding are more expensive methods that must of necessity be limited to a small number of cases. Trace files can verify whether problems observed in a laboratory setting, for instance, also occur in the field, as well as the frequency with which they occur. Another benefit of trace files is that one can examine the effect of successive instrument changes during the course of an ongoing study. A number of the problematic items or screens we identified in this paper have subsequently been (or will soon be) changed in the NHIS. Analysis of trace files from later versions of the instrument will reveal whether the change has indeed reduced the need for interviewer actions such as backups and help screen access.

References

Baker, R.P. (1992), "New Technology in Survey Research: Computer-assisted Personal Interviewing (CAPI)," *Social Science Computer Review*, 10 (2): 145-157.

Caspar, R.A. and Couper, M.P. (1997), "Using keystroke files to assess respondent difficulties with an ACASI instrument." Proceedings of the American Statistical Association, Survey Research Methods Section.

Couper, M.P. (1998), "The application of cognitive science to computer assisted interviewing." In M.G. Sirken, D.J. Hermann, S. Schechter, N. Schwarz, J. Tanur, and R. Tourangeau (eds.). *Cognition and Survey Research.* New York: Wiley, forthcoming.

Couper, M.P., Hansen, S.E., and Sadosky, S.A. (1997), "Evaluating Interviewer Use of CAPI Technology," in L. Lyberg, P. Biemer, M. Collins, E. De Leeuw, C. Dippo, N. Schwarz, and D. Trewin (eds), *Survey Measurement and Process Quality,* New York: Wiley.

Couper, M.P. and Nicholls II, W.L. (1998), "The history and development of computer assisted survey information collection." In Couper, M.P., et al. (eds.) (1998), *Computer Assisted Survey Information Collection.* New York: Wiley, forthcoming.

Hansen, S.E., Couper, M.P., and Fuchs, M. (1998), "Usability evaluation of the NHIS instrument." Paper presented at the Annual Meeting of the American Association for Public Opinion Research, St. Louis, May.

Hansen, S.E., Fuchs, M., and Couper, M.P. (1997), "CAI Instrument Usability Testing," Paper presented at the Annual Meeting of the American Association for Public Opinion Research, Norfolk, VA, May.

Lepkowski, J.M., Couper, M.P., Hansen, S.E., Landers, W., McGonagle, K.A., Schlegel, J., Wright, T., and Chevarley, F. (1998), "CAPI instrument evaluation: Behavior coding, trace files, and usability methods." Paper presented at the Annual Meeting of the American Association for Public Opinion Research, St. Louis, May.

Sperry, S., Edwards, B., Dulaney, R., and Potter, D.E.B. (1998), "Evaluating Interviewer Use of CAPI Navigation Features," Chapter 18 in M.P. Couper et al. (eds.) *Computer Assisted Survey Information Collection*, New York: Wiley, forthcoming.