MEASURING INTERVIEWER PERFORMANCE USING CAPI

Mick P. Couper, Sally A. Sadosky, Sue Ellen Hansen, Survey Research Center, University of Michigan
Mick P. Couper, Survey Research Center, 426 Thompson, Ann Arbor, MI 48106-1248

1. Introduction

With the change from paper and pencil interviewing (PAPI) to computer-assisted personal interviewing (CAPI), new ways need to be found to evaluate interviewer performance, particularly in using the new interviewing tools. The introduction of CAPI has by no means diminished the need for critical interviewing skills (Couper and Burt, 1994). However, the focus here is on the new skills required of interviewers in computer-assisted interviewing (CAI). This paper is a brief summary of a larger research project to use mock interviews and keystroke files to evaluate interviewer performance in a CAPI survey.

Traditional measures of interviewer performance in PAPI surveys are no longer sufficient in a CAI environment. In many PAPI surveys supervisors review completed questionnaires for legibility, completeness of responses, answers within range, correctness of skips, and so on. Not only is this no longer possible in a CAI instrument, but it is also less necessary as the instrument ensures completeness, follows the correct skip patterns and performs range and other error checks automatically. This does not mean, however, that the interviewer does not make errors in using the CAI instrument. What it may mean is that many of the errors have become harder to detect, or that different kinds of errors are being committed.

One set of tools for evaluating interviewer performance that works on both PAPI and CAI surveys is behavior coding of taped interviews (for face-to-face surveys) and monitoring (for telephone survey interviews). These methods yield rich detail on the interaction between the two humans engaged in the interview (the interviewer and the respondent), but provides little if any insight into the interaction between the human (primarily the interviewer) and the computer. Ideally, a CAI system (both the hardware and software) should be an unobtrusive presence in the interview, a tool to facilitate the smooth and successful completion of the interview. To the extent that interacting with the computer causes difficulty for the interviewer, it may intrude in the interaction between interviewer and respondent. In the same way that interactional difficulties in the interviewer-respondent interaction may harm the data obtained (see Suchman and Jordan, 1990), so too may the interaction between the interviewer and computer in a CAI survey.

How can the interaction between interviewer and computer in CAI effectively be studied? Laboratory tests, commonly used in software evaluation to observe the interaction between user and computer (Dumas and Redish, 1993), create a largely artificial environment, threatening the generalizability of the findings to field settings. Observing natural interactions in a field setting also poses problems, as each interviewer is faced with a wide variety of different situations, making comparisons across interviewers difficult. Scripted mock interviews offer advantages over both these approaches. They provide a relatively natural setting and a consistent set of stimuli to all interviewers (Rustemeyer, 1977). They also allow tests to be embedded in the instrument that can be used to evaluate interviewer performance on aspects of CAI use.

CAI provides an additional source of data in that a record can be obtained of all keys pressed by the interviewer as s/he moves through the instrument. One feature of keystroke files that has mitigated against their use in evaluation is that they tend to produce large volumes of free-format data that are difficult to reduce to meaningful levels. It is thus important to develop methods for reducing and analyzing interviewer keystroke behavior in order to aid our understanding of how interviewers interact with the laptop computer and what types of errors they make.

This study combines the use of mock interviews with an analysis of keystroke files to investigate interviewer use of a CAPI instrument. The use of a consistent set of questions and answers for all interviewers was used to facilitate the analysis of keystroke files. Little is known about the types and frequencies of errors interviewers make when using a computerized instrument. The focus of this paper is on an evaluation of the utility of these tools for measuring and evaluating interviewer performance in CAI.

2. Errors in Human-Computer Interaction

Why should we be concerned about interviewer errors in a CAI survey? There is evidence from a number of CAI surveys (see Baker, 1992; Weeks, 1992) that CAI improves data quality over paper-and-pencil data collection, and indeed this is why many have adopted the new technology. If attention is focused only on interviewer keying errors as recorded in the final data set, initial evidence suggests that such errors are extremely rare, particularly on fixed-response

or closed-ended questions (Dielman and Couper, 1994; Kennedy, Lengacher and Demerath, 1990). Interviewers are clearly able to comply with the requirements of CAI, but there is little information on the extent to which they are doing so efficiently or in the way intended by the survey or instrument designer.

Despite the low error rates reported for CAPI surveys, evidence from other fields suggests that users of computer systems make mistakes far more frequently than might be expected (Shneiderman, 1992; Card, Moran and Newell,1980; Brown and Gould,1987). Even in relatively closed systems like CAI where the user (interviewer) has little control over the process and a relatively limited range of actions to perform, errors can have an impact on data quality, efficiency and user satisfaction. To the extent that an interviewer is thwarted in the attainment of a particular goal (e.g. changing a previous answer), the result may range from increased frustration to abandonment of the original goal (e.g. leaving the answer unchanged). Focusing only on the data recorded in the completed interview may underestimate the problems interviewers face, whether through shortcomings in the instrument itself, interface problems between the interviewer and the system, insufficient training, or inadequacies on the part of the interviewer.

Following Zapf, et al. (1992), we employ a broad view of errors in this paper. By error we mean any (temporary) non-attainment of a goal. We focus on all types of difficulties interviewers may experience as they interact with the computer during a survey interview, whether or not these are productive of incorrect data at the end of the process.

Errors have been the focus of a large number of studies in the field of human-computer interaction, resulting in a variety of error classification schemes (see for example Reason, 1990; Rasmussen, Duncan and Leplat, 1987). Many of these follow Norman's (1983) distinction between mistakes and slips. A slip occurs when a correct intention is executed wrongly (e.g. deleting a file accidentally), while a mistake is an incorrect intention where the action conformed to the intention (e.g. deleting a file intentionally although it is still needed) (see Frese and Altmann, 1989: 67). This error classification scheme, as does any other, must at some point make inference about the intended goals of the users, which makes it difficult to distinguish between the two classes of errors.

Many of the available error typologies are based on further distinctions of the levels at which the error occurred. For example, Zapf et al. (1992) distinguish between three levels of action regulation, the intellectual level (at which thought errors, memory errors and judgement errors are made), the level of flexible action patterns (at which errors of habit, omission or recognition are made) and the sensorimotor level (which includes typographical errors). They note that in systems of low complexity, errors at the lower levels (such as sensorimotor errors) are more likely to occur (p. 321).

In contrast to many of the systems studied, most CAI systems are relatively restricted. There is a limited range of possible actions available to the interviewer, limiting the utility of existing error typologies for use in a CAI application. Furthermore, imposing an error classification scheme on interviewer behavior in CAI may be premature without first exploring the types of keystroke behavior that occur in a CAI application, and the frequency with which various errors occur. Finally, the intentions of the interviewer may not be revealed in the examination of keystroke behavior, although in some cases such intentions may be inferred from surrounding keystrokes. We thus use a detailed examination of the keystroke files from the mock interviews to explore the nature of the interviewer-computer interaction and to identify the types of errors that interviewers make in using a CAPI system.

3. Design and Data Collection

This research was conducted as part of the first wave of the study on Asset and Health Dynamics of the Oldest Old (AHEAD). This was a national survey of adults aged 70 and older conducted by SRC. The sample consists of approximately 9,500 households and 12,000 individuals. Both telephone and personal interviewing was used by field interviewers, using laptop computers to collect the survey data. We refer to both interviewing modes as CAPI to emphasize the dispersed nature of the data collection. The AHEAD instrument was programmed using Autoquest.

A total of 137 interviewers were trained for the AHEAD study in Fall, 1993. All newly-hired interviewers received training on general interviewing skills, after which all interviewers attended four days of study-specific training. This included an introduction to computer basics and the CAPI software, round robins and role play exercises, individual practice time and homework exercises using the laptop computer.

At the start of the AHEAD training sessions, interviewers completed a questionnaire designed to elicit background information, including both survey and computer experience, and attitudes toward the use of computers for interviewing.

Immediately after training, and before the start of production interviewing, each interviewer was asked to complete a scripted mock CAPI interview over the telephone with their field supervisor, all using the same

script and instructions. The scripted mock interviews included a number of tests of various CAI functions (e.g. entering a "don't know" or refusal, changing an answer on a previous screen). In addition to the specific tests, a wide variety of other interviewer keystroke behaviors during the mock interviewer are examined.

Interviewers and supervisors were instructed to treat the mock interview as a "live" practice, and each mock interview was tape-recorded by the supervisor to ensure that instructions had been followed. Both the datafile containing substantive responses and the keystroke file from each mock interview were transmitted to Ann Arbor via modem as part of their regular transmissions.

Out of the 137 interviewers trained for the AHEAD study, all completed the interviewer questionnaire during training. Although all interviewers did the mock interview, usable keystroke files were received for only 132 interviews, of which an additional 7 were incomplete, leaving 125 complete mock interview keystroke files for analysis.

4. Analyses

The keystroke files for the mock interviews were examined in two ways. First, a coding scheme was developed to summarize interviewer performance on each of the 13 tests embedded in the mock interview. Second, a WordPerfect macro was written to produce summary counts of a variety of function key presses and key combinations, which were compared against a "template" or model keystroke file representing the most efficient route through the instrument.

These two approaches to keystroke file analysis complement each other. The first, in similar fashion to behavior coding, yields rich data on what happened at various points in the interview. The second, less time-consuming and labor intensive, yields data that are less rich in detail. This approach aggregates selected keystroke behaviors across the entire interview, and includes any use of function keys outside of the specific tests which may have called for their use. The aggregate counts can reveal how many times a key was pressed by an interviewer, but does not inform us whether the use of that key was appropriate or successful. However, this approach can help to identify interviews that may require closer attention using more detailed procedures. Each method of analysis thus focuses on different types of keystroke behavior.

The function key and other key combinations for special functions are presented in Table 1.

Table 1.
Function Keys Used in AHEAD Instrument

Key	Function
[F1]	Question-specific help; [Esc] to exit
[F2]	Comment or note
[F3]	Pop-up menu in household roster screen
[F4]	NO FUNCTION
[F5]	Suspend interview and save data
[F6]	NO FUNCTION
[F7]	Next unanswered question (after [F9])
[F8]	NO FUNCTION
[F9]	Backup one response
[F10]	Restore response (after [F9])
[Alt D]	Don't know
[Alt R]	Refused
[Tab]	Move forward to next item on screen
[Shift Tab]	Move back to previous item on screen
[Up],[Dn], [Left],[Rt]	Cursor keys used to move through roster screens
[Backspace]	Destructive backspace
[Esc]	Cancels entry or exits from help screen

5. Evaluation and Coding of Individual Keystroke Files

The tests embedded in the mock interview were designed to test the use of some of the CAI functions listed in Table 1. Some tests required a combination of CAI functions and interviewing skills, making it difficult to disentangle the source of any error.

For example, in the first test, the "respondent" answered "11 years of school in Norway" to the question on father's education. Interviewers were expected to look up the question by question (QxQ) specification using [F1]. There they would find that they should enter a note or comment to describe the answer (using [F2]), then a code of "97" (other). Failure to do a QxQ lookup may be due to interviewing skills (either the interviewer knew what to do without consulting the help screen, or assumed that this was not necessary), or CAI skills (not knowing how to invoke the help screen or enter a note in Autoquest).

Almost half of the interviewers did not use a note to enter the response. The vast majority of these simply entered "11" for 11 years of schooling. Note that whereas the balance (52%) entered the correct response, only 6% invoked help on this question.

There was a more explicit test of the use of the help screen in which the respondent specifically asked for a definition. In this case 86% of the interviewers looked up the definition on the help screen. In a third test of the help screen, a question asking for Medicare number included an interviewer instruction on the CAPI screen to use [F1] if the respondent needed persuasion.

In reaction to respondent hesitation, 68.5% of interviewers used the help screen. Of the 124 interviewers who completed all three of these tests, almost all (96%) successfully used the [F1][Esc] key combination at least once to look up information on the help screen. Only 5 interviewers were unable to access question-specific help screens on any of these three tests.

We can conclude that knowing how to access online question-specific help is not a problem for most interviewers. However, they tend to use this utility sparingly, even during an interview where they may still be unfamiliar with the survey instrument. We do not have comparable data for paper-and-pencil interviewing, but suspect that the use of the interviewer manual is equally rare.

This raises questions for training on interviewing skills. If we expect interviewers to use QxQs, we should do a better job training them to do so. The problem in the first test (father's education) appears to be one of interviewer judgement as to the appropriateness of the response, rather than lack of knowledge of CAI functions. Interviewers need to know both when to look up additional information for a question, and what kind of information they will find when they do so.

The first test included both the use of context-specific help, and the use of notes to record verbatim responses. As already mentioned, 52% of interviewers correctly used [F2] to enter a note. In another test using notes, all but 20 interviewers correctly used [F2] to enter a note. Combining these two tests, it is found that only one interviewer did not successfully enter a note or comment in either question. This again suggests that the problems (if any) that occurred on these tests were not primarily caused by a lack of knowledge of CAI functions.

Another set of CAI functions that interviewers were trained to use was backing up to change answers to earlier questions. Essentially [F9] is used to back up one question at a time. Once the required question is reached, the interviewer could simply type in the correct response, or press the [F10] key to redisplay the previous answer and make any necessary corrections. The interviewer could then either press [F7] to jump forward to the next unanswered question, or proceed forward one question at a time through the answered questions. Neither the [F10] for redisplay nor [F7] to jump forward again are necessary actions. Nonetheless, it would be informative to see how many interviewers made use of these features.

Three tests involved backing up and changing previous answers, one going back a single question, and the other two going back two questions to change answers. The successful use of the function key [F9] to back up to a previous question ranged from 82% to 95%. Combining the results of all three tests, all 124 interviewers who completed the three tests used [F9] to back up at least once. However, 20 interviewers failed to enter the correct response at least one of the three times. A total of 88% of the interviewers used [F7] on at least one of the two tests on which it could have been used.

Three of the tests examined the entry of a "don't know" response. Again, the percentage of interviewers who successfully used this CAPI function is high: 98% of all interviewers completed at least one of the three tests requiring the entry of an [Alt-D] for a don't know response.

A test of the use of a refusal response was included in a series of items in which interviewers were expected to use [Alt R] to terminate the series if the respondent could not continue. Of the 125 interviewers who completed this test, 79.2% correctly used [Alt R], while 12% initially used [Alt D] then [Alt R], and 8.8% did not use [Alt R] at all.

Generally the keystroke files from these tests reveal that interviewers have little difficulty using the most common CAI functions, even shortly after training. However, this may be missing the difficulties they are having in getting the tests completed. A more detailed examination of the individual keystroke files is being undertaken to explore the things that interviewers may be doing in attempting to reach a correct response.

6. Aggregate Counts of Keystroke Behavior

In addition to evaluating the specific tests, the keystroke files were used to produce aggregate counts for certain keystroke sequences for each interviewer. This was done to determine how many times interviewers used various function keys, regardless of whether or not these were explicitly tested during the mock interview. For purposes of comparison, the selected keystrokes were also counted in a model or template mock interview.

Interviewers were expected to invoke the help screen four times during the scripted mock. This would mean that the [F1] would be pressed four times, and the [Esc] key pressed four times to exit the help screen. On average, interviewers used help fewer times than expected. In fact, 4% of the interviewers never pressed the [F1] at all during the mock interview, while a further 7% did not press [F1] and [Esc] in combination to invoke the help screen and then exit.

It was also found that 30% of interviewers never used the optional [F7] to restore all responses and return to the next unanswered question. The use of this function is not critical if the interviewer goes back only

a few items. From these tests we do not know whether these interviewers did not use [F7] because they didn't know how to or didn't feel to need to do so. Unless such a test is explicitly included in a mock interview, we cannot distinguish between these two reasons.

The two tab keys, [Tab] and [Shift Tab], are used to move forward and backward respectively across the items on a roster screen to make corrections. Most of the interviewers (82.4%) did not use these keys in the mock interview. However, only a single interviewer did not use the arrow (or cursor) keys, which are similarly used to correct errors. On average these keys were used 28.5 times (standard error=1.66). The backspace key is also heavily used by interviewers, on average 16.7 times (standard error=1.65). This suggests that interviewers are frequent users of cursor keys to correct typographical errors. To the extent that such errors are noncritical (i.e., they don't change data values), such editing can be considered inefficient behavior on the part of interviewers.

To determine how many unnecessary keystrokes on average are used by interviewers, the mock interview keystroke files can be compared to the template file. The template file includes a total of 1,603 keystrokes or keystroke combinations (e.g. [Alt-D]). On average, interviewers used 1,563 keystrokes to complete the mock interview, fewer keys on average than in the model interview file. This could reflect the fact that the model interview included correct recording of all respondent and interviewer comments, probes, etc. An indicator of possible inefficient keystroke use is that interviewers on average used cursor keys 45.2 times, while their use was required only once in the template file. Thus, we may be underestimating the level of inefficient keystroke behavior by interviewers. Nonetheless, it seems that interviewers are not grossly inefficient in their use of the CAPI instrument.

How can inefficiencies be reduced in computer-assisted interviewing? On approach may be to tell interviewers to ignore typographical errors in text entry. This was in fact done during training for this study, but it appears that interviewers will insist on editing text responses. This may be inevitable, since most organizations do not permit them to get back into completed CAPI interviews. An alternative approach is to assume that such editing will be done during the interview, and give the interviewers tools and training to facilitate this task. However, is it an efficient use of training time to teach interviewers editing skills? Will interviewers actually benefit from such additional training, or will they suffer from training overload?

The aggregate keystroke files are also useful for identifying cases that warrant closer examination. There were a number of outliers in the counts: [F2] was used 45 times by one interviewer, [Alt D] 63 times by another, [Esc] 59 times, [Right] 98 times, and so on. An examination of these cases may reveal particular difficulties experienced by some interviewers in dealing with the CAPI instrument.

Some evidence of differential use of function and cursor keys by computer experience was found. Interviewers with extensive computer experience used function keys (F1-F10) and special keys (backspace, tab, cursor keys, etc.) significantly more often than those with no previous computer experience. However, those with extensive computer experience are also significantly more likely to press erroneous function keys, suggesting a greater willingness to experiment with the CAPI instrument, or errors produced through incorrect transfer of knowledge from other systems. Similar trends are found for typing skills,although not reaching traditional significance levels.

No significant relationships were found between survey experience and the use of various keystrokes, although the use of marginal notes shows a slight positive relationship with interviewer experience, and interviewers with no prior survey experience used fewer total keys and fewer function keys on average.

7. Next Steps and Conclusions

This paper has presented a brief overview of some of the analyses we have conducted on the AHEAD mock interview keystroke files. Much remains to be done. The keystroke files provide us with a wealth of data on interviewer-computer interaction during the survey interaction. Our work is focussed on understanding such behavior, and finding ways to use these files in a systematic way to evaluate aspects of interviewer performance on CAI, and to identify needed improvements in CAI instruments or interviewer training to minimize the errors or inefficiencies committed by interviewers in using CAI.

The keystroke files from production interviews will be examined using aggregate keystroke counts. Variation in interviewer performance in both the mock and production interviews will be explored further. Additional work will be done on the refinement of coding schemes for the detailed analysis of keystroke files, and for the identification of keystroke sequences to be included in the aggregate counts.

The Autoquest system has been updated to insert question numbers in the keystroke file. Collection of keystroke data from other surveys using this feature will enable us to examine interviewer behavior by question. For example, we will be able to identify on which questions the help function is most likely to be invoked, or where interviewers often have to go back

to change answers, or other places they may be experiencing difficulties.

Despite the successful implementation of CAPI on a number of studies, and interviewers' positive reactions to the use of a laptop computer for interviewing, there is much we still do not know about how interviewers use the systems and instruments we provide them. For instance, we may spend a lot of resources and effort to provide interviewers with online context-specific help, but we do not know how often they make use of such facilities. Analysis of keystroke files from production interviews will permit us to determine the extent to which interviewers use the various CAPI functions that are provided, and the success with which they do so.

Although their limitation should be acknowledged, keystroke files are useful tools in the evaluation of interviewer performance in using CAI. Coupled with mock interviews that provide a standard set of stimuli to all interviewers, they permit an evaluation of how well interviewers are able to use the CAI instrument in most common interviewing situations they are likely to face in the field. Keystroke files should be seen as one tool among many available to survey managers to evaluate the performance of their interviewers. They could be used in conjunction with other measures to assess interviewer competence and skill in carrying out CAI surveys.

Acknowledgements

This work is supported in part by the National Agricultural Statistics Service of the USDA, and by internal Center funds. The authors are grateful to the AHEAD staff and Field and Computing Sections of SRC for their support in the collection of these data, and particularly to Qian Yang for assistance in the preparation and analysis of the keystroke files.

References

Baker, R.P. (1992), "New technology in survey research: Computer-assisted personal interviewing (CAPI)." Social Science Computer Review, 10 (2): 145-157.

Brown, P. and Gould, J. (1987), "How people create spreadsheets." ACM Transactions on Office Information Systems, 5: 258-272.

Card, S., Moran, T.P. and Newell, A. (1980), "The keystroke-level model for user performance with interactive systems." Communications of the ACM, 23: 396-410.

Couper, M.P. and Burt, G. (1994), "Interviewer attitudes toward computer-assisted personal interviewing (CAPI)." Social Science Computer Review, 12 (1): 38-54.

Dielman, L. and Couper, M.P. (1994), "Data quality in a CAPI survey: Keying errors." Journal of Official Statistics, forthcoming.

Dumas, J.S. and Redish, J.C. (1993), A Practical Guide to Usability Testing. Norwood, NJ: Ablex.

Frese, M. and Altmann, A. (1989), "The treatment of errors in learning and training." Pp 65-86 in Bainbridge, L. and Ruiz Quintanilla, S.A. (eds.), Developing Skills With Information Technology. Chichester: Wiley.

Kennedy, J.M., Lengacher, J.E., and Demerath, L. (1990), "Interviewer entry error in CATI interviews." Paper presented at the International Conference on Measurement Errors in Surveys.

Norman, D.A. (1983), "Design principles for human-computer interfaces." Proceedings of CHI '83: Human Factors in Computing Systems. New York: ACM, pp. 1-10.

Rasmussen, J., Duncan, K., and Leplat, J. (eds.) (1987), New Technology and Human Error. Chichester: Wiley.

Reason, J. (1990), Human Error. Cambridge: Cambridge University Press.

Rustemeyer, A. (1977), "Measuring interviewer performance in mock interviews." Proceedings of the American Statistical Association, Social Statistics Section, 341-346.

Shneiderman, B. (1992), Designing the User Interface: Strategies for Effective Human-Computer Interaction. (2nd. ed.) Reading, MA: Addison-Wesley.

Suchman, L. and Jordan, B. (1990), "Interactional troubles in face-to-face survey interviews." Journal of the American Statistical Association, 85: 232-241.

Weeks, M.F. (1992), "Computer-assisted survey information collection: a review of CASIC methods and their implications for survey operations." Journal of Official Statistics, 8 (4): 445-465.

Zapf, D., Brodbeck, F.C., Frese, M., Peters, H., and Prümper, J. (1992), "Errors in working with office computers: A first validation of a taxonomy for observed errors in a field setting." International Journal of Human-Computer Interaction, 4 (4): 311-339.