The 1997 U.S. Residential Energy Consumption Survey'*s Editing Experience Using BLAISE III*

Joelle Davis and Nancy L. Leach, Energy Information Administration (USA)

Introduction

In 1997, the Residential Energy Consumption Survey (RECS)Xone of the major energy consumption surveys in the United StatesXbegan collecting data using Computer Assisted Personal Interviewing (CAPI) techniques rather than the traditional Paper and Pencil Interviewing (PAPI) method. The move to CAPI was motivated by a corporate goal to decrease the time between the collection of data and the publication of results (in previous RECS, there was typically a two to three year lag), and still maintain the high data quality that has historically been associated with the survey. Additionally, a secondary goal was to choose a CAPI software program that allowed the questionnaire to be programmed by subject-matter program staff rather than systems designers.

The RECS is conducted by the Energy Information Administration (EIA), the independent statistical and analytic agency within the U.S. Department of Energy (DOE). Prior to selecting the CAPI software that would be used to program the RECS questionnaire, EIA previewed several CAPI software packages and determined that BLAISE best matched EIA=s requirements of a package that would not only affect how the data were collected, but would also impact the subsequent data processing steps, such as coding the responses, editing the data, and imputing for missing values.

This paper will focus primarily on the impact of the use of CAPI on three survey processing steps: the coding of comments, post-interview editing of the data and imputing for missing data.

Background

The RECS is a national statistical survey of households in the United States. The survey, which is mandated by the U.S. congress, is sponsored by the U.S. Department of Energy. RECS was conducted annually from 1978 to 1982 and then in 1984, 1987, 1990, 1993 and 1997. RECS collects detailed information about the physical characteristics of the occupied housing unit, the demographic characteristics of the household, the types of energy used in the home and detailed information about energy-using equipment and appliances. These data are collected during a voluntary on-site 30-minute personal interview.

Prior to the 1997 RECS, data were collected using PAPI. In 1997, data collection was changed to CAPI techniques using BLAISE III version 1.12. The questionnaire was programmed by technical (not systems) program office staff and contained a maximum of 348 questions with 29 range edit checks and 19 consistency checks imbedded within the questionnaire. These consistency checks consisted of both hard and soft edits. The 1997 RECS collected data from 5,900 households between April 1997 and August 1997, with a response rate of over 80 percent. Over 200 field interviews were trained to use the BLAISE III CAPI questionnaire.

CAPI Impact on Data Processing Steps

Comparisons can be made between the PAPI and CAPI versions of RECS at several stages during the data processing. These key stages are coding the comments, editing the data, and imputing for missing data.

Coding the Comments

Even though BLAISE III was used for the 1997 RECS, some manual coding was necessary in the cases where the interviewer had to key in an "Other/Specify" response or when the interviewer recorded notes in the comment fields. In both cases, the keyed responses had to be reviewed and judgements had to be made about how to code the responses. "Other/Specify" responses were recorded into CAPI during the interview. Following the interview, they were printed out along with the questionnaire wording and reviewed by coders who determined the appropriate response and then entered the response into the CAPI database. This CAPI coding process is many times more efficient than the PAPI coding method, in which editors were required to sort through stacks of questionnaires in search of comments that might need to be coded, at which point they would continue with basically the same steps as in the CAPI coding.

In the case of interviewer recorded comments in CAPI, the comments were printed out by questionnaire identification number and compared with the respondents' answers to relevant questions. Corrections were made if necessary and entered into the database. Table 1 shows that there were savings during the coding phase of the survey in terms of both staff level of effort and timeliness. The time needed to complete the coding was decreased by over seventy-five percent, so although it is true that the 1997 questionnaire was about one-half the size of the 1993 questionnaire, this is still a substantial savings in time.

Table 1. Comparison of Coding Effort, 1993 and 1997 RECS

ITEM	1993 RECS (PAPI)	1997 RECS (CAPI)
Staff Level of Effort .... Number of Coders.................. Number of Supervisors...........	200 Person Days 3 Full-Time 1 Full-Time	45 Person Days 1 Part-Time 1 Part-Time
Elapsed Time from Beginning to End of Coding Task	3.5 Months	18 Days (5 Days for Other/Specify Coding, 13 Days for Interviewer Comment Coding)

Source: Energy Information Administration, 1993 Residential Energy Consumption Survey and 1997 Residential Energy Consumption Survey.

Post-Interview Editing

Range and consistency checks normally occur in a PAPI questionnaire after the field work is completed and all cases have been coded and entered into a database. In the 1993 RECS collected using PAPI, this editing task took approximately 6 months to complete. The Edit Plan, which identified and programmed every possible skip pattern and consistency check, began before the data were available. The plan was 250 pages long with about 50 editing programs. The multistep editing procedure consisted of printing a list of cases that failed a particular edit, printing a list of variables that would be helpful in resolving any inconsistencies, examining the paper questionnaire, determining the necessary corrections to the inconsistencies, recording any corrected information on the questionnaire and on the error listing sheets, and finally, refiling the questionnaires.

In 1997, with the use of CAPI, the post-interview editing was streamlined. Most importantly, many of the edits from the 1993 post-editing plan were incorporated as edits in the BLAISE questionnaire. Also, because the data were immediately available, editors and analysts could begin examining the frequencies and crosstabulations as the data were being collected. This enabled analysts to identify consistency errors during the early stages of data collection and to write and program edit specifications only for identified errors. The Editing Plan and programs were updated and modified, as necessary, throughout the data collection period.

The 1993 Editing Plan of 250 pages and 50 editing programs was reduced to approximately 20 pages and 6 editing programs in 1997. These 1997 edit programs produced lists of cases that failed a particular edit check. The editor examined the data and any interviewer comments that might help resolve the edit failures and determined the necessary corrections, which were then reentered into the CAPI database.

The actual savings in terms of person hours as they relate to the post-interview editing phase of the survey are difficult to determine, partly because of the differences in the way specific task hours were tracked in the 1993 and 1997 RECS, and partly because the survey instrument was shorter in 1997 and did not include a section that had required extensive editing in 1993.

Nevertheless, the fact that the data were immediately available to the survey contractor in 1997 was extremely beneficial. The CAPI data could be reviewed much earlier than the PAPI data, which, in turn, allowed identification of potential data problems during the data collection phase, rather than during the editing phase which previously had occurred after the survey was out of the field. Table 2 shows statistics for the 1993 and 1997 post-interview editing phase of the RECS.

Table 2. Comparison of Post-Interview Editing Effort, 1993 and 1997 RECS

ITEM	1993 RECS (PAPI)	1997 RECS (CAPI)
Staff Level of Effort		
Number of Editors	3	0
Number of Editing Managers	1	1
Number of Computer Programmers........	3	1
Research Assistant	0	1
Elapsed Time from Beginning to End of Coding Task	5-6 Months	3 Months Interspersed With Other Tasks

Source: Energy Information Administration, 1993 Residential Energy Consumption Survey and 1997 Residential Energy Consumption Survey.

Imputing for Item Nonresponse

Item imputation is a statistical process used to generate values for missing items. It is designed to minimize the bias of estimates based on the resulting data set. In the RECS, missing data items were generally treated by a technique known as hot-deck imputation. In hot-decking, when a certain response is missing for a given housing unit, another housing unit with similar known characteristics (the donor) is randomly chosen to furnish its reported value for that missing item. The value is then assigned to the building with item nonresponse (the nonrespondent, or receiver). This procedure is often time-consuming and generally occurs after field work is completed and the data have been edited.

In the 1993 PAPI version of RECS, 378 variables were imputed. About 50 variables were missing data for 10 or fewer cases and about 40 were missing data for 100 or more cases. The vast majority of the missing variables were due to "No Answer" (that is, the interviewer did not record a response or the response was inconsistent with other responses), rather than a "Don=t Know" or "Refusal."

In comparison, the 1997 CAPI version of RECS had only 145 variables imputed. Over half were missing data for 10 or fewer cases and only 6 variables were missing data for 100 or more cases. Most of the missing data in 1997 was due to a "Don=t Know" or "Refusal" response. Because BLAISE does not allow a necessary field to be left blank, this source of error was eliminated. Table 3 provides imputation statistics for the 1993 and 1997 RECS.

Table 3. Comparison of Data Imputation Effort, 1993 and 1997 RECS

ITEM	1993 RECS (PAPI)	1997 RECS (CAPI)
Number of Household Questionnaire Items	559	348
Number Of Variables Imputed	378 (68 percent)	145 (42 percent)
Number of Variables Not Imputed	181 (32 percent)	203 (58 percent)

Source: Energy Information Administration, 1993 Residential Energy Consumption Survey and 1997 Residential Energy Consumption Survey.

Instrument design errors can occur during the use of both PAPI and CAPI. In the 1997 RECS there were some errors that occurred because of CAPI programming errors. However, whenever this type of error occurred, it occurred consistently. In a PAPI instrument, an error or lack of clarity in a questionnaire skip instruction can cause discrepancies as to which follow-up questions are asked; each interviewer may interpret the instruction in a different way. In a CAPI instrument, however, programming errors cause consistent problems to occur in the data. For the 1997 RECS, in some cases the data were retrievable or correctable; in other cases decisions were made to ignore certain data items. Because the errors were consistent, it was easier to decide what could and should be done with the data when such programming errors occurred.

Summary

Both major goals–increased timeliness and high data quality–were met using BLAISE. In 1993, the elapsed time between data collection and EIA's first look at any data was approximately two months; in 1997, analysts were able to view partial data even before the interviewing was complete. The 1997 RECS data was actually published five months sooner than the 1993 RECS. Much of this increased timeliness was due to the fact that the RECS data collected in 1997 using the CAPI questionnaire were much cleaner than were the 1993 data collected using PAPI. The use of CAPI resulted in less nonresponse, and a considerable savings of time during the editing and imputation phases of the survey. Item nonresponse due to interviewer skip pattern errors was greatly reduced. There is little doubt that the built-in data edits nearly eliminated item nonresponse due to interviewer error. The hard and soft edits greatly reduced the number of inconsistencies in the data, thereby resulting in a much cleaner data set. When programming errors did occur, they occurred consistently throughout the 5,900 cases and allowed for consistent decisions to be made on how to handle the problem cases.

Based on the success of the 1997 RECS using BLAISE to program the questionnaire, the 1999 Commercial Buildings Energy Consumption Survey (CBECS), a more complicated survey with many more skip patterns, is currently being programmed in BLAISE 4 for Windows as a Computer Assisted Telephone Interview. This instrument is also being programmed in-house by an EIA technical analyst.