DATA VALIDATION IN THE CAI WORLD

Charlene Walker, Ann Brown, Richard Veevers - Statistics Canada
Charlene Walker, Special Surveys, Jean Talon Bldg (5-C6), Statistics Canada, Ottawa, ON, K1A 0T6

Key Words: Income Assistance, development, verification

1.0 INTRODUCTION

This paper presents and discusses the data validation approach developed for the computer assisted interview (CAI) surveys of Statistics Canada's Self-Sufficiency Project (SSP). For the purposes of this paper, data validation is defined as the process of ensuring the high quality of survey data. CAI permits a continuous and iterative data validation technique not possible with paper and pencil interviewing (PAPI).

SSP is a large-scale multiple cohort study. It uses monthly data collection in a distributed (decentralized) interviewing environment. CAI technology is in widespread use, but not with many surveys of this scope. Therefore, validation approaches for such surveys are just developing. This paper presents one approach.

The SSP validation process encompasses a Total Quality Management approach similar to that proposed for editing by Granquist (1995). (However, it is important to note that editing is only a small part of this data validation approach.) Even though the validation approach is continuous and iterative, it can be divided into two phases: development and verification.

Data validation in the CAI world can begin long before data are available, through efficient development. The objective of the development phase is to design a high quality, user-friendly computer assisted application. The objectives of the verification phase are to ensure high quality of collected data and to provide feedback to the development phase to improve the survey process.

Challenges which are unique to CAI are highlighted in this paper. In conclusion, the paper describes how the experience of the first iteration of validation was used to fine-tune the validation of subsequent survey waves.

2.0 THE SELF-SUFFICIENCY PROJECT

2.1 Overview

The Self-Sufficiency Project (SSP) is a research demonstration project studying the effect of a short-term employment-based income supplement offer on the self-sufficiency of single parent Income Assistance (IA) recipients. Eligible IA recipients qualify for a limited-time monetary supplement if they find full-time work within a given time period. The data will indicate whether the supplement offer is effective in encouraging single parents to become independent of income assistance.

The effects of the supplement are being studied using a multiple cohort study. Statistics Canada is developing and administering a baseline survey and three follow-up surveys between 1992 and 1999 to approximately 9400 project participants. Sample intake for the baseline survey took place monthly over a two year period (2126 in the first year and 7284 in the second). Dependent interviewing is used for follow-up contacts approximately 18, 36 and 54 months after the initial contact. The data dependency across waves is an advantage only if the data are accurate. This is one of the reasons that data quality is such a concern.

SSP is a social experiment, that is, it uses a random assignment of project participants (randomly selected from the target population) to program and control groups to determine the effect of a treatment (the offer of the income supplement). For more details on SSP, see Lui-Gurr *et al.* (1994). Reliable observations concerning the effect of the earnings supplement offer can be made by comparing these two groups, only if non-sampling errors are controlled. The experimental design and small sample size mean that any small data anomalies can erroneously indicate a difference between program and control groups or conversely, no difference. Any bias introduced by inaccurate data can be detrimental to the analysis. This is another reason that it is crucial to ensure that the data are of high quality. For more details on quality assurance in SSP, see Brown *et al.* (1995).

2.2 The CAI Application

The Canadian Labour Force Survey (LFS) was converted to CAI in 1993. It was Statistics Canada's first CAI survey in a distributed interviewing environment. It was decided to allow interviewers to familiarize themselves with CAI technology using the LFS before other surveys began using CAI. Thus, it was not possible to conduct the SSP baseline survey using CAI technology.

The first follow-up survey was the first opportunity for SSP to utilize CAI technology, computer assisted telephone interviewing, specifically. The SSP team was interested in CAI technology for several reasons. One of the primary reasons was that information about the household composition and employment status can be readily returned to the interviewer in subsequent waves. Another is the increased complexity of question flows and edits that a CAI application allows over a PAPI questionnaire.

The CAI application developed for SSP comprises a case management component, a roster module and a core questionnaire. The case management system was designed by Statistics Canada to plan the monthly assignment, assign outcome codes, and encrypt and transmit data. The roster, also designed by Statistics Canada, introduces the survey to the respondent and allows confirmation and updating of demographic information. The SSP core questionnaire collects information about: marital history; employment history; job search strategies; education; child care arrangements; the respondent's home; restrictions due to disability; types of non-employment income; quality of living conditions; attitudes toward IA and work; and knowledge about SSP.

There are several features of the CAI application that have proven to be very important to the data validation process. One of these features is the ability to enter comments using the F4 function key. If the respondent has difficulty understanding the question or provides information that is relevant to but does not directly answer the question, this can be noted in a comment screen. This provides a technique to monitor questions more closely than possible in a PAPI questionnaire, where the difficulty and cost of data capturing comments would be prohibitive. The information can be used to modify question structures in future waves as well as provide a qualitative context to questions requiring a categorical or numerical answer.

Another feature that is important to the validation process is the fact that modules in the core questionnaire are time stamped upon entry and exit. In addition to providing budget information, these start and stop times are used to verify the questionnaire outcome codes (Section 4.1).

A feature that greatly improves the data quality in a CAI application is the ability to perform edits during the interview process.

Soft edits can be used in conjunction with numeric or date questions, when the respondent provides a response that would appear to be out of range. The application generates a message asking the interviewer to verify the answer with the respondent. If the respondent changes his or her response, the new response can be entered but if the respondent confirms the answer, the interviewer can simply override the edit and move to the next question. Soft edits have the advantage of improving data quality without the limitation of requiring the respondent to provide a response that passes the edit rule.

Hard edits, on the other hand, must be used sparingly since they require a response within the acceptable range before the interview can proceed. There are some instances when a hard edit is very useful. For example, Employment History Question 26 asks "How many weeks per month did you usually work at that job or business?" The answer can be in the range of zero to four. Responses outside this range are neither possible nor accepted. This reduces the number of invalid cases due to response, transcription or data capture errors.

3.0 DEVELOPMENT PHASE

The first step in data validation is development. As Kovar (1995) noted, it is important to prevent data quality problems rather than to try to correct them after the fact. With this in mind, the SSP team invested a great deal of effort in questionnaire design, application testing and interviewer training for the first follow-up (year 1) survey.

3.1 Questionnaire Design

Questionnaire design is an important component of data validation. It is important to ensure that questions are understandable and unambiguous. With a CAI application, it is possible to create complicated questionnaire flows thereby minimizing the need for editing skip patterns after collection. Automated edits can be incorporated to minimize errors.

Questionnaire design is an iterative process. In CAI, as opposed to PAPI, the iterative process can be continued into the collection phase, but there is a risk of carrying it too far. After three months of collection, the first follow-up core questionnaire was changed, with significant consequences to post-processing (Section 4.4).

3.2 Testing

A very structured approach to testing is important in order to maintain control over the testing process and to ensure that all possibilities are covered in testing. Predetermined scenarios with predetermined outcomes are keyed in by testers. Problems are documented and communicated to the programmer. After problems are corrected, all

scenarios are retested to ensure that nothing was inadvertently changed during the debugging process (Nowlan, 1995).

Application testing is a complicated process consisting of three stages: modular testing, integrated testing and end-to-end testing.

Modular testing divides the application into small manageable pieces to make testing and debugging easier. Attention is paid to question wording, flows and edits. For SSP, the division into modules was a natural one because of the grouping of questions. Each module is tested independently of the others before the modules are combined, or integrated, into a complete package.

During integrated testing, testers simulate interviewing in the field. Each tester enters a variety of scenarios on a laptop to ensure that the modules work well together. The scenarios follow the interview process from opening the case and completing the roster, through the questionnaire, to the end when a final outcome code is assigned.

End-to-end testing expands integrated testing to simulate the situation in a regional office. Cases are routed to many interviewers, scenarios are keyed in, cases are finalized and transmitted back to a supervisor. Tracing, transmission and post-processing are tested at this stage.

3.3 Interviewer Training

Extensive interviewer training is an essential component of data validation. A comprehensive training manual and formal training in the Regional Offices provide the interviewers with a solid knowledge base. In addition, the application contains resettable practice cases allowing the interviewers to practice using the computer application and to become familiar with the subject matter.

4.0 VERIFICATION PHASE

The second step in data validation is verification. After data are collected, they are verified to ensure they are complete and accurate and to improve the questionnaire and the survey process.

The major steps in the data verification of the first follow-up (year 1) survey data were verification of outcome codes, removal of duplicate records, comment file analysis, data reformatting, consistency checking and interviewer retraining.

4.1 Outcome Codes

Outcome codes represent the response status of an interview. Outcome codes were analysed to ensure they accurately represent the status of the data. The application automatically assigned a complete code when an interview had reached the end of the application. A partial code was manually assigned if the interview had at least begun the Education module. All other records were manually assigned a non-response code.

Unfortunately, manually assigned codes are not always accurate. In some instances, dockets were assigned non-response codes when there was sufficient data to warrant a partial code. Conversely, some records were coded as partials without sufficient data.

Because of the possibility of error, all outcome codes were examined. Each record was subjected to a test using modular start and stop times to ascertain whether or not that record met the requirements for that particular outcome code. For instance, complete records should have non-blank values in all start and stop times. All records which did not meet the requirements for their outcome code were examined and resolved manually.

4.2 Duplicates

The goal was to provide one record per respondent with the most complete information possible. However, at the end of each monthly data collection period, interviewers transmit all records to Head Office. Records with non-response codes are returned to the Regional Office, along with the new sample, for another attempt at contact (for up to three months). As a result, each record may appear up to three times on the data files. These duplicates were resolved automatically.

Even though a respondent may have more than one record across months, they should not have more than one record per month. This would mean that it had been duplicated during transmission or post-processing. Analysis revealed that only one record had a duplicate within the same data collection month. It was resolved manually.

After duplicates were resolved, 1911 complete and partial records remained.

4.3 Comment File Analysis

Use of the F4 function key in the application allows interviewers to enter comments which are saved to a comment file. The ability to enter comments has many

potential benefits to CAI data quality over PAPI data quality. Explanations of complicated situations could greatly help researchers in analyzing data. In a PAPI data capture system, there is no facility to capture comments and if there were, the cost would be prohibitive.

A flag on the comment file indicates the type of comment. The first type are verbatim comments (text field write-ins saved automatically to the comment file). The second type are comments entered using the F4 function key (available for any question on the questionnaire). The third type of comments contain 'other specify' write-ins which are automatically saved to the comment file. (Since these accompany the data file, they were not analysed.)

An extensive and lengthy analysis of the comments was undertaken to make improvements to the questionnaire for the next survey wave and to indicate possible data problems. For the first recipient follow-up (year 1), 1911 respondents had a total of 4364 verbatim comments in the comment file and 2398 F4 comments. There were several reasons for the vast number of comment file entries.

One reason for many comments was the learning curve involved. Time is needed for the interviewers to learn the most efficient use of the comment field.

The second reason is that interviewers are reluctant to move around in and experiment with the application because of a fear of new technology. If they discover additional or conflicting information about a prior question, they tend to enter it as a comment rather than move back to the prior question to change the data.

The third reason for many comments was the use of hard edits for dates. Even though every attempt was made to make the application user-friendly, it was still more rigid than PAPI. Sometimes, hard edits prevented the interviewer from entering the date provided by the respondent (especially for job start dates), so they entered the hard minimum date and a comment.

The final reason for many comments was the one-way communication between the software packages. It was not possible to return to the roster once in the core questionnaire. Therefore, if the respondent gave the interviewer additional demographic information during the core questionnaire, the interviewer had to enter the information as a comment.

In addition, the software allows one to specify where verbatim write-ins are saved - the data file, the comment file or both. After examining comments, it became evident that all verbatim questions were not handled consistently. Thus, it was necessary to analyse verbatim comments in order to standardize where they were saved.

As a result of the number of comments, the analysis was extremely time consuming. In order to minimize the subjectivity and time needed, a conservative approach to changing data was adopted. In addition, a well-structured analysis procedure and extensive quality control were implemented to ensure data quality was high.

The following sections present the results from each of the validation steps based on comments.

4.3.1 Verbatim Comments

Questions in the employment (industry, occupation) and education modules (major field of study) are the only verbatim fields in the questionnaire. Since these fields were extracted for coding, it was necessary to ensure that the data file contained complete information. Therefore, verbatim comments were analysed to determine whether they contained valuable information missing from the data file. Comments were classified according to whether there was an entry in the data field and if so, whether the comment was different from the data field.

For 78% of verbatim comments, the corresponding data field was blank. A manual review of the comments revealed that the necessary information was expressed concisely at the beginning of the comment field and so could readily be transferred to the (shorter) data fields.

In 21% of verbatim comments, the data field was not blank. For 20%, the comment field added no additional information and so the data field was left as is. For the remaining 1%, changes were made manually as necessary. Manual changes were needed when valuable data appeared at the end of the comment.

A duplicate category was introduced when more than one comment for the same question were found. (If a comment was too long, it wrapped to the next line in the comment file.) The comment with the most useful information was retained. The others were deleted and counted as duplicates (0.9% of verbatim comments).

4.3.2 F4 Comments

Since comments entered using the F4 function key did not accompany the data file to the analysts, it was necessary to determine if they had any impact on the data or on future questionnaire design. Therefore, F4 comments were analyzed using a similar classification system as was used for verbatim comments.

Most F4 comments (76.2%) did not prompt data changes because they elaborated upon the answer, for example, a comment of 'approximate'. In some cases, the comments contradicted the data field. These comments were ignored because of the difficulty in interpretation and because changes were made conservatively. In the Personal Attitudes module, none of the comments necessitated changes because the comments generally indicated reasons for the given answers.

Simple recoding was done when comments for 'mark one' questions necessitated recoding of only the associated question and did not affect the questionnaire flow (1.5% of F4 comments). These were recoded automatically to a value which had been entered on the database.

Complex recoding was necessary (13.4% of F4 comments) for 'mark all that apply' questions, numeric or date questions and if skip patterns were affected. For the most part, changes were made manually (8.9% of F4 comments). This rate was much higher for Marital History (17.9%) because respondents sometimes reported marital status changes that had not been reported in the roster and for Non-Employment Income (18.9%) which had numerous 'mark all that apply' questions. It was possible to automatically recode cases when respondents reported a monthly rather than semiannual income, by multiplying the number of months that the respondent received the income by the amount reported in the comment (4.5% of F4 comments). These all occurred in the Non-Employment Income section.

Approximately 9% of F4 comments corresponded to verbatim questions. Most occurred in the employment history module. For half of these verbatim comments, the data field was blank. As with the other verbatim comments, these comments were copied to the data field (after verifying that no other comment existed containing more information). The remaining changes (4.5% of F4 comments) were handled by manual intervention because of difficulty in interpretation.

4.4 Reformatting

Reformatting was necessary in order to present the data in a logical, straightforward manner. There were two reasons that the data initially were not in the best format possible. The first reason was mid-collection questionnaire changes and the second was software shortcomings.

Due to the questionnaire change, it was necessary to combine the data from two questionnaires to form a concise package by reformatting the early data to resemble the later data. For example, a code was introduced to indicate that a question should have been answered but was not. The SSP experience indicates that even if changes are well documented, they make post-processing and data verification extremely difficult.

The application had several shortcomings that had to be solved through reformatting. 'Mark all that apply' questions were reformatted to indicate whether a respondent skipped the question. In addition, for 'mark all that apply questions,' responses of 'don't know' and 'refuse' were moved to their proper positions in the answer subset. Finally, partial records were reformatted to make it evident when the interview ended.

4.5 Consistency Checks

Due to the version change and the changes resulting from comment analysis, it was necessary to verify all question values and flows before finalizing the data. This was done by checking frequencies of each question to verify that all the values were valid and that the correct number of respondents were following the skip patterns. This revealed errors in manual editing due to comments and it uncovered some questions that had not yet been reformatted.

At this point, the interview date was also verified by examining the values to certain questions in the non-employment income module where specific categories are dependent on the interview date. Since the interview date is generated by the system on the last day the record is accessed before transmission, it sometimes was not the actual interview date and had to be recoded.

4.6 Interviewer Retraining

Based on the comments and outcome codes received, interviewer instructions were sent out to clarify questions and correct problems in interpretation. For example, Education Question 21 asks "For how many weeks did you take/have you been taking this training?" A common comment was "one day," or "2-3 days." Detailed instructions were sent out describing how to code these cases. In addition, interviewers were instructed to use the comment field more sparingly.

5.0 THE FUTURE

The data validation process was an excellent learning opportunity. The SSP team gained an understanding of data validation itself and also a better understanding of

CAI technology. Out of this knowledge, many changes were made prior to the next wave of validation.

5.1 Development Phase

Data validation of the first follow-up data resulted in changes to the second follow-up questionnaire (a computer assisted personal interview). It became obvious from comments that certain questions were ambiguous or difficult to understand. They were clarified or expanded. For example, in the first follow-up questionnaire, Your Home Question 5 asks "What do you pay for your monthly rent or mortgage?" A common comment would explain that a boarder or a common-law spouse pays some or all of the rent. As a result, this question was split for the second follow-up - one part asks what the respondent pays, one asks if other people pay part of the rent or mortgage, the third part asks for the amount that other people pay. Also for the second follow-up, many hard edits were softened.

To reduce the problem of invalid outcome codes, the second follow-up application automatically assigns a partial code if the interview reaches a certain point.

To avoid version changes for the second follow-up survey, more time was spent fine-tuning and testing the application to minimize changes after the start of collection. In addition, for the second follow-up survey, a version number was added to help with version control.

5.2 Verification Phase

In order to reduce the amount of manual intervention as much as possible, much of the verification process has been automated. For the most part, verification takes place in a database query system containing one record per docket with a clean outcome code (duplicates and invalid status codes are resolved outside the system). The query system allows an operator to classify comments and edit data, while keeping a record of those changes. It also allows for ad hoc queries and can be used to examine data from previous survey waves by linking databases through a common identification number. Finally, the query system can be used for consistency checking by automatically verifying flows and question values.

As a result of the comment analysis, verbatim comments will not be analysed in subsequent survey waves. Instead, verbatim comments with a corresponding blank field in the data file will automatically be copied to the data field and all others will be ignored. This eliminates the need to manually classify verbatim comments while not introducing any significant error into the data.

6.0 CONCLUSION

Data validation is an important process for any CAI survey, but especially for complex surveys such as SSP. CAI has the potential to vastly improve data quality, but only if a continuous improvement cycle is used to measure and validate data quality, and to improve the survey-taking process. CAI technology is opening up a universe of possibilities for complexity in questionnaire design, but, that complexity must be balanced by suitable data validation in order to ensure high data quality.

7.0 ACKNOWLEDGEMENTS

The authors wish to thank the SSP project team members at Statistics Canada, both past and present, who made this paper possible: Mike Garcia and Lia Gendron; and for their helpful comments on this paper: Scott Murray and Geoff Hole at Statistics Canada; and a special thanks goes out to Dave Dolson for all his help.

8.0 REFERENCES

1. Brown, A. and Veevers, R. (1995), Quality Assessment and Improvement in the Self-Sufficiency Project, Proceedings of the International Conference on Survey Measurement and Process Quality.

2. Granquist, L. (1995), Improving the Traditional Editing Process, Chapter 21, Business Survey Methods, Wiley Series in Probability and Mathematical Statistics.

3. Kovar, J. (1995), Data Editing Methods and Techniques, Volume 2, Chapter 5, What to Do When an Edit Fails, Conference of European Statisticians, Work Session on Statistical Data Editing.

4. Nowlan, S. (1995), Computer Assisted Interviewing Testing Guidelines, Internal Memorandum, Statistics Canada.

5. Lui-Gurr, S., Currie Vernon, S. and Mijanovich, T. (1994) Making Work Pay Better than Welfare: An Early Look at the Self-Sufficiency Project, Social Research and Demonstration Corporation.