MANAGING THE QUALITY OF ESTABLISHMENT SURVEYS IN AN ENVIRONMENT OF RAPIDLY CHANGING SURVEY-TAKING TECHNOLOGIES

Mary March, Business Survey Methods Division,
Statistics Canada, Ottawa K1A OT6

KEY WORDS: Reengineering, quality, integration, employee involvement, automation, large units.

1.0 INTRODUCTION

Recent technological advancements, particularly in the domains of telecommunications and of personal computing, are enabling dramatic changes in processes used to conduct establishment surveys. Because of these changes, surveys can be reengineered to reduce respondent burden, speed up survey-taking, cut down on paper usage and handling of questionnaires and, ultimately, reduce costs and improve efficiency. In addition, it is possible to improve the quality of the survey products and of the manner in which they are delivered. Reengineering also can potentially have a positive social effect- an improved quality of work life for employees of the survey-taking organization and a better interaction with respondents.

Examples of reengineering of survey processes include: computer assisted survey interviewing (CASI); selective approaches to editing; data collection using handheld computers; data collection using touch-tone telephones; data collection using EDI(Electronic data interchange); survey systems made up of ready-made generalized components; integration of survey operations (an example of organizational reengineering); centralized frames; etc. In another era, automation of editing and imputation (using methods such as "hot deck" and "nearest neighbour") and of record linkage were examples of reengineering.

The purpose of this paper is to examine the impact of the recent major changes to the nature of establishment survey processes from the point of view of approaches and methods to be used for quality assurance. The paper will begin by explaining why, in managing quality of surveys, statistical organizations focus on processes. Then, traditional survey processes and the tactics that have been applied to assure their quality will be characterized. Then, the paper will examine some of the reengineering trends, analysing them from the point of view of controlling the quality of survey processes. It also will recommend logical approaches to quality assurance - based on experiences of survey takers and on strategies used in industry.

2.0 MANAGING THE QUALITY OF SURVEY PROCESSES

While technology-induced advances are an undoubted boon in today's tight money circumstances and have the potential to improve the quality of survey processes, they can introduce a challenge for survey managers. They must still be able to guarantee acceptable quality levels and to prove that those quality levels are being met. There is no doubt that a reengineered survey process can produce, in many circumstances, better quality. However, sometimes it is more difficult to measure and control quality as the process takes place. This can be due to reorganization of the work in a way that results in less control (for example, moving a task from a central location to the interviewer or to the respondent), to elimination of paper records (possibly removing a physical audit trail), to the process becoming much more complex or to its being accelerated because of automation. Reengineering solutions can make it more difficult to observe a process or to obtain measures of its quality.

Some managers may be tempted to conclude that being able to observe a process to verify quality is no longer necessary. After all, there is often a capacity to build in automated controls or edits. This conclusion is valid if it is possible or practical to specify edits that can detect all of the important errors. Otherwise, some other form of quality assurance is still necessary.

Managers could also decide to approach measurement and control of quality levels from the point of view of assessing the final products. They may reason that if the final product is acceptable to users, there is no need to delve into the way it is being produced. Experience has shown that it is rarely possible for users to be certain that they have a good survey product by assessing the end product alone. Although survey estimates can be compared to historical data or with estimates from administrative sources or from other surveys, the alternative estimates are themselves usually subject to error. "Performance testing" of estimates is not a straight-forward task. Fellegi(1981) noted that "there are no objective and quantitative yardsticks by which a client could observe and on the basis of which he/she could determine whether statistics will perform according to his/her expectations". He added that "quality assurance . . . inexorably leads one to a review of the design and implementation of statistical processes."

It follows that the only reasonable approach to quality assurance of survey products is to focus on the survey processes. Therefore, as processes are reengineered it must continue to be a priority to build in tools or mechanisms to control their quality so that they stand up to review.

3.0 TRADITIONAL ESTABLISHMENT SURVEY PROCESSES AND QUALITY ASSURANCE ASSOCIATED WITH THEM

Quality has long been an important concern with surveys. There is a significant amount of risk associated with a survey (unless it is a small study) because it is usually an expensive undertaking and it is not always possible or practical to test every aspect of it beforehand. Also, some important factors that determine success, such as respondent cooperation and availability of skilled staff are not always in the control of the survey-taking agency. It has long been understood that it is necessary at the planning and development stages to ensure that concepts, procedures and tools are workable and will produce valid results. Once the survey begins, operations must be monitored attentively, with the hope of detecting any major problems early.

The purpose of this section is to enumerate the series of survey processes that have comprised an establishment survey in recent times, to assess their nature and to discuss tactics that have been employed to assure quality.

3.1 TRADITIONAL SURVEY PROCESSES

Not every survey requires the same processes. A first-time survey needs planning and design work that will not be repeated the second and subsequent times that the survey is run. Initial steps can include: determining if a survey is necessary, determining the survey's target population, investigating availability of information from administrative sources, choosing collection methods, designing the questionnaire and collection procedures, selecting or building a frame, designing and selecting the sample and its associated estimation procedures and putting together the survey systems. Once the survey has been run one or more times, design flaws will have been ironed out and the initial steps become unnecessary. For every survey, processes include: data collection (possibly using one or more methods), data editing at various stages, follow-up to resolve inconsistencies and non-response, data capture, imputation of missing information, estimation (including measures of reliability), dissemination of the estimates, data analysis, interfacing with users and, finally, survey evaluation including recommending modifications to prepare for subsequent runs.

Methodologies used for establishment survey processes have evolved over time as have approaches to assuring their quality. Reengineering could be said to characterise at least some parts of this evolution. (Although there may not have been as much motivation to economise as there is now.) Two developments in the past that could be considered examples of reengineering are: the introduction of automated edit and imputation systems in the 70's and 80's, and integration efforts such as the one at Statistics Canada in the last decade, which consisted of developing a centralized business survey infrastructure including common concepts and an integrated frame.

As mainframes and mini-computers have become available, most survey processes have been at least partly automated. Processes such as sample selection, imputation, estimation and production of statistical tables have become mostly automated - consisting of running appropriate batch computer programs one or more times depending on the size and complexity of the survey. Interactive systems with on-line edits have been used for data entry and editing since the late 70's. More recently, interactive systems have become available on a variety of platforms to enable computer-assisted interviewing.

3.2 QUALITY ASSURANCE OF TRADITIONAL PROCESSES

Quality assurance of survey computerized systems has traditionally been at the front end- at the design and development stages. Specifications have been thoroughly reviewed, checked and approved by users before a system is built. As the system was built, the documentation was verified to ensure conformance to specifications and, as much as possible, the system was acceptance tested by the organizations and individuals who would eventually use it. Once production began, the development team monitored the system closely- standing ready to "troubleshoot" if problems occurred and to correct the system if required. After that, the systems have stood as implemented except in cases where use of the system for a longer period identified previously unseen problems or deficiencies and revisions or modifications were required. (For Statistics Canada systems built by the Systems Development Division for internal use, correction of deficiencies and minor modifications have normally been considered to be "maintenance" and are usually the responsibility of the original development organization.)

For systems acquired from another survey-taking organization or from a commercial organization, acceptance testing and correction of errors and deficiencies have not been as straightforward. Purchase of a maintenance contract can help, but, the original interface with the supplier - including specifications and agreements if a system was built on contract and careful assessments of requirements or needs versus available functions for a system bought "off the shelf"- have assumed critical importance.

Quality assurance of human components of survey processes has been an important consideration at every stage of implementation. Before beginning the process, quality assurance methods that apply include: careful review and testing of concepts, thorough testing of questionnaires and procedures, selection of capable staff and effective training. Once processes were implemented, commitment to quality assurance manifested itself as competent supervision (including coaching and re-training as appropriate); as monitoring and inspection; and as mechanisms such as debriefing sessions and evaluations that enable employees to raise questions and problems and to provide feedback about the effectiveness of procedures. Where feasible, processing operations have incorporated systematic inspection of work applying statistical quality control concepts and methodologies and including feedback to employees, managers and users. Statistical quality control, a means by which managers can readily monitor and control quality levels, has been used for straightforward repetitive operations, such as clerical editing or data entry when they are carried out in a controlled environment in a central office (either the headquarters of the organization or a regional or district office). In such situations, quality control methodologies with roots in manufacturing have been and continue to be applied - inexpensively and with minimal disruption to the flow of operations.

Acceptance sampling, by which lots of work are subjected to a statistical test - consisting of checking a sample of units to obtain a measure of the quality of the lot which is compared to an acceptance limit - and reworked if rejected, has been used with Statistics Canada's most important manual coding, data entry and editing operations. The methodology, if applied rigourously, can guarantee a specified level of outgoing quality. The power of the statistical test can be varied, for example, by reducing sampling rates for employees who show a history of low error rates. As, to the present, even the most stable surveys experience error rates at units per hundred levels for human processes, acceptance sampling is still a viable approach. (In the manufacturing sector, where targets for defective rates are at units per thousand levels or lower, acceptance sampling no longer is appropriate. These error rates must be much lower because, for example, to have one in one hundred telephones not working would imply a huge number of disatisfied customers while error rates at that level in survey units will have a minor effect on survey estimates compared to those of response and sampling errors.)

Editing has been and continues to be a very important part of quality assurance strategies in survey-taking organizations. An edit, when applied to survey information, can be defined to be any logical check or test that is applied to detect possible errors. Checks of this nature are applied throughout the course of a survey. Edits can be applied at the field or record level or they can be applied to several records at the same time or to information that has been derived from aggregating records. Editing is, in fact, a form of verification. Juran(1988) mentions 100% "automated inspection" which, for surveys, could correspond to running an automated editing program.

With establishment surveys, all units being processed are not equal in terms of their impact on estimates. Acceptance sampling methods by which each unit within a lot has the same probability of selection for correction may not sufficiently control outgoing quality (although from the point of view of providing feedback on types of errors and of detecting errors in procedures they are very useful). Editing all records, even those with very little impact, can be inefficient. At Statistics Canada, survey practitioners have long been aware of the importance of certain units identified as "outliers", "take-alls", "top contributors","large companies", etc. because of their significant impact on survey estimates. These units are usually accorded special treatment, it being deemed essential that their information be checked for validity on a 100% basis, usually manually.

Activities carried out by humans that are more complex or that do not take place in a controlled environment have always presented more of a quality control challenge. Unfortunately, also, they are processes such as certain design activities or face-to-face interviewing where there is a potential of serious modelling, conceptual or response errors being introduced. Errors at these stages could have an even more serious impact than would arise from errors in the simple to control operations such as data entry or editing.

Most professionals and technical staff at Statistics Canada work at a central location. In those circumstances, supervision, peer review processes and consultation with advisory committees can help to assure quality. Designers also receive feedback from the operational staff and from end data users regarding workability of concepts, methodologies and systems.

Quality in face-to-face interview situations has been much more difficult to address. Quality assurance at earlier stages of the process is important - i.e. more attention to cognitive research, to planning and to training beforehand. Also, all materials and procedures used must be prepared and tested carefully. Public communications from the survey organization to the respondent before the interviewer arrives also can have a positive impact. Application of any quality control or measurement procedure during the interviewing process or after the fact is not easy. Problems arising during an attempt to use verification by reinterviewing during a Statistics Canada business profiling study in the mid-'80's illustrated this difficulty. The interviews were very complex. The businesses being reinterviewed did not appreciate the burden of a second interview nor did they seem to care for the agency's apparent mistrust of its interviewers. A supervisor can accompany interviewers to verify interviews but this is only feasible on a limited basis.

With most collection processes, quality assurance during the process can be difficult and often costly to implement. Therefore, most of the effort has had to be at the front end - before the process takes place. Current quality management thinking's assertion that it is more efficient to build in quality up front than to "inspect in" quality applies in the survey-taking environment - especially at the collection stage.

Given that budgets are limited, the extent to which resources have been applied to quality assurance is determined from the manager's previous experience in developing surveys and the amount of risk that there is. If the survey design is a replicate or almost a replicate of a previous survey, quality assurance is less of an issue.

4.0 SURVEY REENGINEERING TRENDS, SOURCES OF POTENTIAL ERRORS ASSOCIATED WITH THEM AND QUALITY ASSURANCE STRATEGIES

Do current developments affecting the nature of processes that make up establishment surveys require that traditional methods used for quality assurance must also be rethought? In this section, an attempt is made to identify a group of trends and to assess each from the point of view of appropriate approaches to quality assurance.

1. There is a trend to focusing effort on the large or important units because they have the most impact on the estimates.

As mentioned, large units have always been subject to special attention. In recent years, however, survey managers have had to identify priorities as part of finding ways to reduce costs. Protecting the quality of data for the large units is usually assigned a high priority as methods are rethought. Selective editing and concentration of survey frame maintenance and collection efforts on larger units (relying on administrative information for smaller units) are all part of this trend. Also, processing steps associated with these units are also unlikely to be completely automated because of complexity and, therefore, per unit costs can be quite high. Quality assurance associated with these units is an extremely important issue from the point of view of quality and cost. Errors in processing these units could have serious consequences. Also, since the cost of processing each unit is high, it is essential that causes of errors be determined and feedback provided to reduce the likelihood of rework being required. Quality is a important issue because errors in these units can have a major impact. The cost Currently, there is a tendency to 100% verify the manual work on these units, at least their processing if not the collection of their data. It seems apparent that the quality of processing of these units should be a main focus of quality management strategies for establishment surveys. It seems worthwhile to focus process error measurement, analysis and improvement efforts on these units.

2. Integration approaches are being used extensively in the interests of efficiency.

There is a move to combine collection of data for several surveys where the same respondent is involved, to combine previously separate survey processes and to use the same staff, systems and/or facilities to handle more than one survey. One advantage of integration (i.e. the moves to combined surveys, centralised processing, generalized systems, etc.) from the quality assurance point of view is that there are likely to also be less unique processes to be controlled from the point of view of quality.

A more integrated approach to quality assurance also should be possible. It might be expected, also, that it would take less time for processes to be brought into control since less learning is necessary. On the other hand, up front quality assurance will increase in importance. Developing tools for multiple use will require increased attention to assuring their quality before they are made available.

3.There is increased automation - more use of personal computers and of telecommunications technology.

Collection processes are being affected dramatically. Mailed-out questionnaires and paper questionnaires completed by face-to-face or telephone interviews are being replaced with data entered directly into a computer by the respondent or interviewer making use of computers and available telephone technology. Data collection has traditionally been the most expensive stage of a survey and the most difficult to control. Information required by establishment surveys is often complex and there can be a considerable amount of numerical data to be handled. For respondents, participating can be time-consuming and rewards for cooperation are minimal.

With the advent of smaller and increasingly powerful laptop computers, hand held computers, greater touch tone capability and with the development of business communication technologies (such as FAX, EDI, electronic forms, etc.), it is becoming possible to reduce the burden and the cost of collecting information while improving timeliness. Almost every establishment survey can be reengineered to some extent to take advantage of new developments and trends in technology. Survey taking organizations are investing in new technologies with the reward of decreased survey costs.

Quality assurance of systems - especially of system requirements - is more important than ever. Also, testing should be as efficient as possible. Systems are likely to be used widely, so they must be thoroughly tested. In addition, shortages of funds do not allow for a "trial-and-error" approach to starting up with the expectation of being able to handle numerous problem notices. Where outside suppliers are involved, it is important to establish close customer-supplier partnerships with them. Suppliers selected should have demonstrated their commitment to quality by applying appropriate quality control procedures and providing dependable after-sales service.

For the actual survey process that has been automated, it is important to ensure that adequate quality control procedures are still in place. By reducing the human component, there may be less opportunity for errors to be introduced. When redesigning a process, possible sources of errors should be carefully assessed and quality measures such as special training programs, built-in edits and/or some form of monitoring capacity included.

4. Paper is being eliminated or reduced considerably.

With the introduction of personal computers and with user friendly inter-active systems and increasing storage capacity, it is becoming possible to virtually eliminate paper records. Not only are questionnaires being replaced by a system data record but the numerous computer printouts, coding forms, etc. that have been produced during processing and editing operations can also be phased out. The impact of this trend is a reduction in survey costs, including costs of paper, printing, shipping and storage and less waste of a scarce natural resource.

Elimination of paper has a considerable impact on some traditional quality assurance methods which take advantage of the existence of paper. For example, data entry quality has been measured by re-entering information from the same questionnaires and comparing it with previously entered information. Without the existence of questionnaires, it is no longer as straightforward to measure data entry error rates. Similarly, editing actions cannot be evaluated as easily. In fact, the use of acceptance sampling methods to control the quality of survey operations may no longer be efficient or practical given that processes are moving more quickly and are being decentralised. (Stopping at some point to accept or reject work is more likely to cause a bottle-neck.) The only alternative appears to be to rely mostly on up front quality assurance - that is, assuring the quality of the design, procedures, systems and training, to make better use of edits (automated verification) and to work to minimize resources and time required for inspection. It appears more appropriate to monitor processes than to allow time for inspection after the fact.

5. Handling of documents and of data is being reduced.

This trend is the consequence of other trends such as focusing on larger more important units, integration, automation and eliminating paper. The impact of this trend is that there are likely fewer different sources of error and, hence, errors will occur less frequently. However, the manual intervention, when required, may require more skill and errors will tend to have more impact.

As a result of this trend, survey managers will find that development and training of staff will take on more importance as will early detection of errors and feedback.

6. The jobs of survey-taking staff are becoming more "professional".

As the amount of human intervention decreases, fewer persons are required to carry out a survey but their jobs are becoming more complex and involving. It could be said that they are being "professionalised". Their jobs are becoming more interesting and perhaps more worthy of commitment.

If employees feel more responsibility for their jobs, they are more likely to be interested in improving quality. As reengineering redefines jobs and gives them more substance, it is possible that quality assurance can become much more employee-driven.

The employees themselves will decide where more training is needed and be prepared to participate in activities such as continuous improvement teams and in the redesign of processes to better suit the people involved (an activity known as sociotechnical engineering).

As surveys take less time, workers may find themselves performing more functions or working on more surveys. For some workers, there may be increased stress associated with survey work and perhaps a requirement for support from the organization in dealing with it.

Managing the increased complexity of survey jobs including the training of staff and designing workable procedures will be a key component in assuring the quality of surveys.

5.0 CONCLUSIONS

After examining trends associated with reengineering of establishment survey processes, it becomes clear that approaches to quality assurance must be rethought as well. Many of the trends such as integration and automation require more preparation up front. Most of the effort associated with quality assurance must also be moved up to earlier stages of the process. Inspection and correction of errors needs to be replaced with more effort at planning stages. Inspection to any extent should only be necessary at the beginning of a process to quickly ensure that it is under control and performing as expected.

For any human process, it is still important to be able to guarantee that expected quality levels are being achieved, because the user will not be able to judge the quality of a product using an external yardstick. With integration, automation and less handling it should be easier to obtain such measures.

The trend to professionalising employees should make management of the quality of survey processes easier. "Professional" employees should be more committed to improving quality and therefore ready to participate in managing quality.

Reengineering in an organization usually involves more thinking ahead. The accommodation of the management of quality must be part of this thinking.

6.0 REFERENCES

Colledge, M. (1990), Integration of Economic Data: Benefits and Problems. Presented at the Statistics Canada Symposium on Measurement and Improvement of Data Quality, October, 1990. Proceedings.

Fellegi, I. (1981), A Note on Quality. Unpublished.

Informatics Services and Development Division (1988), Application Development and Support Methodology, Internal Document. Statistics Canada.

Juran, J.M. and Gryna, F.M., Jr.(1988), Juran's Quality Control Handbook, Fourth Edition, New York: McGraw-Hill.

Juran, J.M. and Gryna, F. M., Jr.(1980), Quality Planning and Analysis: from product development through use. New York: McGraw-Hill.

Pasmore, William A.(1988), Designing Effective Organizations : the sociotechnical systems perspective. New York: Wiley.

Statistics Canada (1987), Quality Guidelines, Second Edition. Minister of Supply and Services Canada.

A METHODOLOGY FOR EVALUATING CATI/CAPI DEVELOPMENT ENVIRONMENTS

Douglas Yeo, Joan Hocken, and Sylvie Michaud, Statistics Canada
Douglas Yeo, R.H. Coats 11-C, Tunney's Pasture, Ottawa, ON K1A 0T6

KEY WORDS: Computer-Assisted Personal Interviewing, Computer-Assisted Telephone Interviewing, evaluation criteria, survey models.

ABSTRACT

Statistics Canada is making increasing use of simultaneous data collection and capture with computer assistance. Attention is being focused on field interviewing. As well, the use of standard CATI/CAPI packages must be balanced with meeting custom survey requirements. To help meet these two related yet sometimes conflicting objectives, a standard set of two survey models and an evaluation criteria has been constructed. The models are hypothetical surveys that emulate questionnaire design and data collection requirements, for business surveys and for social surveys. These tools test the capabilities of a new product. This paper describes the uses made of the methodology to date and the conclusions drawn from the evaluations.

1. INTRODUCTION

Statistics Canada initiated the Generalized Survey Function Development (GSFD) team to provide a more general approach to development of production systems. The GSFD team has developed generalized systems for sampling, data collection and capture, edit and imputation, and estimation.

The Generalized Data Collection and Data Capture System, DC2, is the system most closely involved with Computer-Assisted Interviewing (CAI). It provides for mixed mode collection and capture in a centralized facility. It has a sophisticated management support system, and provides interactive and batch editing for capture from forms and for collection/capture from telephone interviewing. DC2 is a graphics-based system running on a UNIX platform in a client-server model.

Because of the swift pace of CAI technological advances, the GSFD team formed a multi-disciplinary team to keep abreast of New Collection/Capture Technologies. This team, which was formed in 1989, has several objectives:

- to identify new collection/capture technology, both hardware and software. As well, the team was to use a methodological approach to evaluate the potential of these new products for generating cost benefits,
- to identify classes of surveys capable of using the new technology,
- to develop prototypes either to automate the interview or to provide the respondent with ways to capture our data for us, and
- to make recommendations based on prototype testing results.

2. REASONS FOR THE DEVELOPMENT OF AN EVALUATION METHODOLOGY

The team has completed several research projects since its inception. Each project has tested a specific survey/technology combination. For example, touchtone telephone response was tested using a small sample of respondents to the Survey of Employment, Payroll and Hours. The team has learned a tremendous amount about each technology being tested and about the best strategies to employ for prototype development and testing. However, it became apparent that a more formal evaluation strategy was needed. One of the first conclusions reached is that, along with hardware evaluations, the evaluation of *the development environment itself*, and its capabilities, must be stressed.

The development environment includes all tools used to design and develop applications. Tools include application generators and screen painters, which are used to speed up the development process and reduce the resources required.

The development environment, and the time needed for its implementation, must be balanced against the ability to meet survey requirements and deadlines. At Statistics Canada, these survey requirements are complex and application generators often do not have the required flexibility. On the other hand, the use of a standard development environment may reduce resource requirements, important in this period of cost and time constraints. It is vital that Statistics Canada control the high costs of developing systems for rapidly changing technologies.

In order to meet these conflicting objectives, the New Collection/Capture Technology team has been working on methods to evaluate development environments consistently and methodically. Evaluating an environment for one survey is usually insufficient, since there are significant differences in requirements between the various surveys to be automated. It is necessary to look at a combination of surveys. This is especially true in production, where the use of a single new technology for several surveys makes it much more likely to recoup development costs.

For the past year, the team has been developing a standard methodology for evaluating development environments. The methodology has two major components: a survey model component consisting of a Business Survey Model and a Social Survey Model, and an evaluation criteria component. The use of survey models allows the evaluation of the requirements of several surveys at the same time. This methodology has currently been focused on decentralized interviewing. In this scenario, field interviewers would use portable computers both for personal visits and for telephone interviews.

The potential benefits of this methodology are many. It will provide the ability to contrast development environment features with survey requirements. The advantages and disadvantages of each development environment can be identified based on the requirements of a particular survey. It will provide measures on development learning curves, application development times, and prototyping aids. It will identify functions that must be built outside the development environment and integrated, such as editing or case management features. It also will identify functions that cannot be provided at all.

Another benefit will arise in production. When the survey models are developed on different environments, it will be possible to compare performance between the production applications. The production models also can be used to benchmark various hardwares. Testing strategies built for the models may be used wherever appropriate, with little additional time needed to tailor the data.

3. THE TOOLS OF THE METHODOLOGY

3.1 Survey Models The survey models are composite surveys that incorporate typical questionnaire design and data collection requirements at Statistics Canada. In developing the survey models, the requirements from different types of surveys were examined. A list of requirements that would satisfy most surveys was produced. As well, when current applications are re-implemented using new technologies, new requirements are created. These requirements can be added to the survey models. Based on these requirements, two survey models were developed, one for social surveys and one for business surveys.

The survey models are represented as paper questionnaires, derived from a combination of questions from different current surveys. A paper questionnaire does not provide a complete picture of the collection process. For example, it does not necessarily show edits that should be implemented. For this reason, a procedures manual is included with each survey model, which describes in greater detail the collection process, edits, and other features.

In addition to the survey application itself there is a need for case management. Case management includes all the tools that must be provided to an interviewer in order to do a survey. For example, it includes such things as displaying the list of the cases that have to be done, assigning outcome codes, and making appointments. Statistics Canada is currently developing a general case management system. The description attached to the survey models includes basic case management requirements. However, the environment in which the survey models are built should be able to link to the generalized case management system.

3.1.1 Business Survey Model (BSM) The BSM was designed to be representative of business surveys at Statistics Canada. Because of the focus on decentralized interviewing, a mixed-mode approach is assumed. The BSM is a repeated survey, with the first contact generally done by personal visit. Subsequent contacts may be done by telephone from the interviewer's home.

Including numerous features in the BSM satisfied the requirements of a variety of business surveys. The BSM uses several question types, such as open-ended, closed, multiple choice, mark all that apply, and numeric. Validity, historical, and inter-field edits are specified. There is a simple skip pattern in the questionnaire, representative of the kind of branching often seen in business surveys. The BSM also covers automated coding of text, such as location and commodity coding. Provision has also been made for other features such as variable transaction length and variable ordering of questions.

Six sections comprise the BSM. The first, on shipments and inventories, collects numeric data. Next, the agricultural section contains multiple choice and numeric fields. This section includes several skips. This section also asks for a global unit of measure, but allows for a different unit of measure for specific questions.

Sections three and four cover manufacturing and employment. These sections are used to test the ability of a CAI package to navigate inside a grid; the interviewer needs the ability to move around an on-screen table of fields. As well, section three requires the pre-filling of some fields with previously collected information.

Section five, on transportation, contains multiple choice and mark-all-that-apply questions. The CAI package is tested here for its ability to handle variable transaction length. This is a form dynamic rostering. where records are added and subtracted dynamically. A number of transaction records must be completed, and this number varies for each company.

Section six collects consumer price information, with a requirement to be able to change the question order dynamically. This variable ordering of questions is one of the potential requirements of the Consumer Price Index survey, if it is to be converted to CAI. The interviewer needs to see the items on-screen in the same order they appear in the store. As well, it must be possible to change this order, if the store makes any changes.

This section-by-section approach to the survey models is useful for testing different functions. For example, if handheld computers are to be evaluated, the testing would concentrate on section 6 of the BSM.

3.1.2 Social Survey Model (SSM) Like the BSM, the SSM was based on the requirements from several different surveys. One typical requirement is that the sampling unit differs from the surveyed unit. Many social surveys actually sample dwellings, while the surveyed unit is the person. Some surveys also interview everyone in the household. For these surveys, the number of selected dwellings (which corresponds to the sampled units) is known before collection, but the number of interviews that will be done in each dwelling is not known. For this reason, the SSM also includes dynamic rostering in its list of requirements.

The SSM also includes different eligibility criteria for its different sections. Some questions are intended for all household members, some for all the household members within an age group, and one particular section is intended for one person in the household randomly selected.

Different types of questions were included in each section of the SSM: open-ended and closed, single entry, multiple choice, mark all that apply, dates (complete and month/year only), and numeric. A question has also been provided to simulate on-line coding: the question requires selecting an item from a very large list. Additions to the list are also allowed, if the answer is not on the list. The SSM also specifies the derivation of variables that are used later in the interview process (for example, deriving spells, or time intervals, based on dates).

The SSM contains several types of edits: range edits, consistency between variables in the same questionnaire, consistency edits between variables in different questionnaires (for example, consistency between two persons living in the same dwelling), and edits involving dates.

The SSM specifies that questions may use some historical information in the interview. Pre-filled information coming from previous answers in the questionnaire also may be used.

Finally, several new longitudinal social surveys are being implemented at Statistics Canada, with a new tracing requirement. To reflect this within the SSM, some sections are intended for everyone in the dwelling. If a person moves out, a new dwelling has to be created to be able to trace and interview that person.

The SSM has four basic components. The first one is a contact and demographic section. It includes introductory questions, questions to establish that the right household is contacted, and some demographic questions asked of everyone living in the dwelling. It is assumed that for some cases, previous information may exist, and so the information may only have to be displayed and updated if required.

The second component collects labour information. This section applies to all persons living in the dwelling aged fifteen and over. Each person is asked some general activity questions, and some detailed job information, for the two "main" jobs. In this section, the different date questions are used to derive spells of employment, unemployment, and inactivity.

The third component collects some information on income and wealth. Again, this component is asked of each person aged fifteen years and over. In this section, some edits are done between the different income questions, and also between persons in the dwelling.

Finally the fourth component contains some questions about health. The questions are asked of one person living in the dwelling, who is chosen at random.

3.2 Evaluation Criteria The evaluation criteria is based on a standard check list of requirements for CAI software. It is a complement to the two survey models. In concordance with the team's goals, its stress is on the evaluation and comparison of development environments. However, it also contains some hardware requirements. The focus of the evaluation criteria is again on decentralized CAI. It is intended to extend this set to include Computer-Assisted Self-Interviewing (CASI) or Computer-Assisted Data Entry (CADE) requirements when necessary.

The format is based on an earlier set of evaluation criteria produced by the DC2 team (1988). Like the DC2 version, it is divided into three main sections.

The first is a general section, covering such categories as user support, cost, and documentation. This section is used to record basic information about a particular CAI package. This information can be critical. For example, one category covers the stability of the software package being evaluated.

The second section encompasses work station requirements. The main categories are navigation (such as branching and tracking), editing attributes, call scheduling, and case management. Other topics are also covered, such as screen presentation, backup, and security.

The final section is entitled "Developer's Environment", in the sense of systems issues for developers. More technical categories are covered, such as the attributes of the screen painter and the ability to link to other software packages.

It is possible to rank the importance of many items in the evaluation criteria. As in the National Agricultural Statistics Service document, there is a range of rankings, such as two (minimal need) and five (needed for all surveys). The rank also may be B or S, indicating a requirement in business or social surveys respectively. Such rankings are especially valuable for the work station requirements.

3.3 Integration of tools Requirements in the survey models obviously are reflected in the evaluation criteria. However, work must continue to more closely integrate the survey models and the evaluation criteria. From the integration of the tools should come a formal testing strategy. Future tasks are to determine priorities or a score to the different criteria that have been developed, to assess the potential for a new technology, and to be able to compare different new technologies that are tested.

4. CURRENT APPLICATIONS

The tools described above have been put to several uses. The following sections describe some of these applications.

4.1 Research into Pen Computer Development Applications Several years ago, the first pen computers came on the market. Two years ago, the high-technology world recognized the potential and several vendors began announcing new hardware and software products. Changes in this market are occurring with ever-increasing speed, and the cost is rapidly approaching competitiveness with keyboard-based notebook computers. Several areas in Statistics Canada feel that the use of pen computers is the way of the future. This is especially true in decentralized CAI, where the informality of the pen interface would be useful during personal interviews.

For the past year, the systems component of the New Collection/Capture Technology team has been evaluating development environments for pen computers. Preliminary evaluations were conducted to look at the potential to meet Statistics Canada's needs. The team examined two operating systems: DOS and PenPoint. Running under DOS, Pen DOS, PenRight! and Windows for Pen Computing were reviewed. Running under PenPoint, PenApps was examined.

Windows for Pen Computing was selected for further evaluation and the first implementation of the new methodology. The SSM was developed using Borland C++ and Application Framework (BC++). This appears to be a promising approach. It offers compatibility with existing investment in DOS applications and potential for gradual phasing in of pen equipment to replace conventional notebooks instead of replacing all equipment at once. The same software will run on a multitude of pen, notebook, and desktop computers with different-sized screens, provided that the computer also runs Windows.

There is also a new breed of computer, the convertible, which can function as a pen computer or a traditional notebook. The application designer can design applications using either the pen or the keyboard interface, whichever is most effective. Application could often be designed with the ability to use both the keyboard *and* the pen. The interviewer can then use the interface with which he or she is most comfortable. Inside a single application, the interviewer might use the pen for most questions, but switch to the keyboard to enter long strings of text. The environment is capable of handling all social survey requirements. It is believed it is also capable of handling all business survey requirements as well, although this has not been tested yet.

The use of Windows as an application development environment has several advantages. It would minimize the development effort and the expertise required to develop CAI applications. There are software development efficiencies in Windows such as object-oriented programming, code sharing, and third party development tools. Windows itself has been adopted by millions of users as the preferred desktop environment and as a result, Windows programming expertise is widely available. Before a final decision is made on the usefulness of Pen for Windows Computing, the BSM will be developed using this package and the results evaluated.

4.2 Testing of DC2 R2.0 The survey models also can be used to test new software packages. DC2 was developing a new release, R2.0, in the summer of 1992. The main new features of this release were navigation functions such as branching and tracking, a go-to function, and a field-select function.

For the first time, the survey models were used for testing. A test vehicle was created using only those elements of the BSM and SSM that specifically tested navigation functions. The existence of the survey models allowed for the creation of a test vehicle quickly, and with a minimum of effort. The use of the survey models in testing was a success, and testing concluded that DC2 R2.0 will be able to handle most of Statistics Canada's navigation requirements.

5. FUTURE ACTIVITIES

Much work has been done on the Evaluation Methodology. However, there is still work to be done. The ultimate aim is to have the methodology adopted by Statistics Canada as a standard evaluation methodology. To do this, more test applications are required, to sell the benefits versus the costs involved. Some ongoing refinements in the tools will be needed to incorporate new or changed requirements, to streamline the survey models and to develop fully the evaluation criteria.

Key factors must be identified for each survey considered for development using new technologies. Then the strengths and weaknesses of each development software can be determined, by survey. Development of a full testing system is not complete. Extension of the methodology to incorporate Automated Sources is desirable. Finally, the team must concentrate on using the methodology as it was designed to be used: for the evaluation of new or improved development environments, including application generator packages such as CASES, PenApps, Blaise, and DC2, where the product offers potential savings for Statistics Canada.

ACKNOWLEDGMENTS

We would like to acknowledge the project manager Lynn Woelfle, the contributions of Gerrit Faber and Jacqueline Pottle for their work on the survey models, and Claude Plouffe and Peng-Kuong Lim for their work on pen computing. Further, we would like to thank the Statistics Canada reviewers, Mark Kinack and Sylvain Perron.

REFERENCES

DC2 Team (1988), "Evaluation Criteria for Proprietary Data Entry Systems", internal Statistics Canada report.

DC2 Team (1991), "Summary of Collection/ Capture Requirements", internal Statistics Canada report.

Egan, Mike, Simard, François, Brule, Mike, and Nadarajah, Raja (1992), "Requirements for CATI; Concentrating on Creating an Application to be Used by Interviewers", internal Statistics Canada report.

National Agricultural Statistics Service, Interactive Survey Software Committee (1992), "Criteria for the Evaluation of Interactive Survey Software", internal NASS report.

Plouffe, Claude (1993), "Pen Computing: Potential Use for Statistics Canada", internal Statistics Canada report.

COMPUTER ASSISTED PERSONAL INTERVIEWING (CAPI) COSTS VERSUS PAPER AND PENCIL COSTS

Bruce Eklund
National Agricultural Statistics Service (NASS)
3251 Old Lee Highway, 3rd Floor, Fairfax, Virginia, 22030

KEY WORDS: Blaise, Computer Assisted Personal Interviewing (CAPI), Interactive Editing (IE)

Computer assisted personal interviewing (CAPI) promises improved data quality and timeliness, data quality through edits invoked *during* data collection and timeliness by telecommunication.

The purpose of this paper is to help an organization ask the right questions. First, is CAPI feasible? The organization must write easy-to-use instruments for its surveys and integrate CAPI into its survey process.

A more comprehensive question is "*should* an organization use CAPI?". Issues include feasibility, whether an organization can realize CAPI's potential benefits, and the topic of this paper: cost. This paper briefly reviews experience that shows NASS can realize CAPI benefits, describes initial cost comparisons, and documents a parallel CAPI and paper comparison.

CAPI Cost Analyses are Specific to the Organization

Britain's Office of Population Censuses and Surveys thought that computer assisted interviewing, including CAPI, would not only save money, but it was imperative to meet new budget constraints (Manners, 1991). How CAPI affects costs, however, depends greatly on an organization's survey program. Cost issues include the degree of centralization, number and frequency of surveys, consistency of surveys over time, length of data collection periods, and the amount and complexity of editing.

NASS's decentralized structure will challenge CAPI management. NASS trains enumerators from 44 state offices. In each state they gather each year for each major survey. Training costs already consume a considerable portion of survey costs. Costs include state office staff and enumerator salaries, lodging, and per diem. If CAPI training increases total training time, then CAPI increases NASS survey costs appreciably.

NASS collects survey data within short periods. Also, most surveys are quarterly or annual, not weekly or monthly. Thus, although NASS needs to write several survey instruments and train for several surveys, there are short, concentrated periods to realize CAPI benefits.

Conversely, during the concentrated periods, NASS spends considerable staff time on office editing after the interviews. Clerks verify arithmetic calculations; data entry clerks key and verify; and professional, subject matter specialists (agricultural statisticians) hand edit questionnaires. Afterward, NASS rents mainframe computer time to batch edit the data. Statisticians usually wait overnight for the output, make corrections, and then run another batch edit overnight. If CAPI reduces office editing time, it will greatly affect survey cost.

NASS Experience

NASS developed CAPI for a cross section of its surveys. The more of the survey program NASS can use CAPI, the more economically viable CAPI is. NASS would show that for a cross section of its surveys:

1) NASS can develop easy to use survey instruments

2) enumerators can learn and use CAPI well

3) telecommunication would speed data retrieval

4) CAPI would help clean data.

NASS first used CAPI operationally for a simple survey, the Livestock Prices Received. CAPI data were compared to paper collected data with a post-collection batch edit. CAPI flagged suspicious data (outside specified ranges) during the CAPI interviews. The enumerator could fix an error or verify valid data. CAPI collected data had 57% fewer suspicious "error" flags in the post-collection batch edit (Eklund, 1991).

More importantly, CAPI reduced "critical" errors by a factor of 16.5. Critical errors are non-sensible entries.

For this survey, batch edit critical errors often resulted from missing items that made data records unusable. CAPI required enumerators to answer these items before proceeding.

Post-Collection Batch Edit "Error" Rates[1]

Livestock Prices Received Survey

For Paper & Pencil

non-critical, suspicious "error"%= 3.74%

critical error%= 0.33%

n ~ 11000

For CAPI

non-critical, suspicious "error"%= 2.15%

critical error% = 0.02%

n ~ 12000

1) Pct. of records (data from livestock sales) with "errors".

In 1989, for the September Agricultural Survey, NASS found what other organizations have found: enumerators can use CAPI effectively and respondents accept CAPI. Two survey software products, CASES and Blaise were tested for CAPI.

NASS also used laptops and electronic calipers to size almonds to forecast production during the growing season. Enumerators entered data both from the key board and the caliper. The caliper, connected to a serial port, put data directly into the survey instrument. Enumerators then sent the interactively edited data to the State office by telecommunication, showing the ability to reduce the time between field assessment and crop forecasting. Furthermore, the technology can eliminate the need for a two-person team. With CAPI, one person can simultaneously measure and record.

Next NASS tested CAPI with a challenging survey, the Farm Cost and Returns Survey (FCRS). The paper questionnaire is over thirty pages, has complex branching, requires plenty of arithmetic, asks detailed financial questions, and averages 90 minutes per interview. It is difficult to convey the specific intent for some questions. Often farmers cannot respond precisely. NASS proved that Blaise CAPI software could handle this complex questionnaire. Enumerators showed that they could learn and use CAPI for one of NASS's most difficult surveys.

Relationship between CAPI & Interactive Editing (IE)

An important part of cost comparison is the relationship between CAPI, Computer Assisted Telephone Interviewing (CATI), and after data collection, Interactive Editing (IE). The Netherlands' Central Bureau of Statistics developed integrated survey software called Blaise. The user can write code and then compile in Pascal into either a CAPI, CATI, or an IE survey instrument. After CAPI or CATI, a subject matter specialist can use an IE instrument to review the data interactively, questionnaire by questionnaire, with automated edit checks. Programmers have only to write one set of code for CAPI, CATI, and IE. Cost estimates hereinafter assume NASS effectively integrates CAPI and IE software functions.

What Does CAPI Cost?

Understandably, NASS did not design its survey cost accounting with categories conducive to contrast paper and CAPI costs. Thus ten state offices estimated costs for several sub-categories of survey preparation, enumerator training, data editing, data entry, batch editing, and mailing costs.

A framework, starting with the FCRS project, was built for estimating and comparing CAPI cost. Spreadsheets can ease updates of cost estimates as costs change and people learn to use CAPI more effectively.

Corresponding costs were estimated or projected for CAPI based on previous CAPI and IE applications. Although survey specific cost estimates were developed for the FCRS, hardware costs were amortized both across surveys within a year and across years for hardware life expectancies. Thus, one can extrapolate

CAPI COSTS
vs.
PAPER COSTS

SURVEY COSTS	*PAPER*	*CAPI*
Enumerator Training		X
Hardware		X
Survey Preparation	x	
Mail Costs	x	
Telephone Costs		x
Data Entry	x	
Data Editing	X	

the cost estimates for the gamut of CAPI targeted surveys. The chart depicts major (large X) and minor (small x) cost increases. An assumption is that people use CAPI/IE effectively, which excludes some initial costs.

Not only hardware but also training cost is likely to increase significantly with CAPI. Effective *and* cost efficient CAPI training is challenging. The organization must balance or integrate *effective* methods such as low student/teacher ratios and field observation with cost *efficient* methods such as home self training and practice. Also, survey trainers should build CAPI practice into the survey schools.

Most savings will come from office editing, which is examined in following sections. CAPI can save some survey preparation cost. Automated distribution of samples is cheaper than the present method of writing a program to print labels in each state office and then manually distributing properly labeled questionnaires to each enumerator. CAPI also saves mail and data entry costs but increases telephone costs somewhat with telecommunication.

Total CAPI cost was estimated close to or slightly less than paper collection in 1992 (Eklund, 1993). The margin of error of some estimates was such that the paper method may or may not have still been slightly less expensive.

The important points are to have reasonable estimates and to understand cost trends. Paper intensive costs such as mail and office labor are increasing. CAPI costs, such as hardware and telecommunication are decreasing (Clayton and Harrell, 1989). An organization should develop, test, and gain CAPI experience on a small scale to prepare for large scale use when cost, data quality, and timeliness, point to CAPI use.

CAPI and the June Area Frame Survey

Although this survey is considerably shorter than the FCRS, enumerators are much more likely to interview farmers outside. They often stand to interview as farmers work. Enumerators also must carry a 2' by 2' aerial photograph of the farm. Smaller computers enabled enumerators to use CAPI successfully in June, 1993.

The Blaise CAPI instrument was also coded for statisticians to use as an IE instrument after data collection. The IE instrument invoked more rigorous edits then CAPI. For example, the Blaise software requires one to "fix" edits deemed critical, but one can override non-critical errors. Programmers coded some critical IE edit checks as non-critical CAPI edit checks. Like the FCRS, enumerators handled this difficult survey well.

CAPI/IE Costs and the June Area Frame Survey

Once enumerators sent data to the State office, the editing flow was divided into four processes.

Process 1: Traditional method

1) paper interview => clerical edit => 1st statistician edit => 2nd statistician edit => key entry => key entry verification => batch edit.

Process 2: IE with one statistician hand edit

2) paper interview => clerical edit => 1st statistician edit => key entry => key entry verification => IE => batch edit.

Process 3: IE with no statistician hand edits

3) paper interview => clerical edit => key entry => key entry verification =>IE => batch edit.

Process 4: CAPI

4) CAPI => IE => batch edit.

The office staff divided into three teams that rotated to edit under each of the first three processes. The office processing time per interview is in the following table. The CAPI/IE combination shows the potential to reduce the office process by a factor on more than three.

Survey Process Times in State Office

(Minutes per "Interview")

Process	1	2	3	4
Clerical Edit	2.5	2.4	3.8	-
1st Statistician Edit	5.2	3.9	-	-
2nd Statistician Edit	3.3	-	-	-
Key and Verify	5.1	5.0	5.0	-
Interactive Edit	-	3.4	4.7	≤ 5
Total Time	**16.1**	**14.7**	**13.5**	**<5**
n - Interviews	757	341	391	86

Statisticians edited using the IE instrument more slowly than they will in the future because of: 1) a learning curve and 2) imperfections in the operationally untested instrument.

The IE part of process 4 was not in a production mode as processes 2 and 3 were. CAPI data were compared item by item to the paper backup to scrutinize enumerators' CAPI. Thus, the IE part of processes 2 and 3 were conservatively extrapolated to IE for process 4.

The project did not measure some survey management costs such as post-collection file handling. Conversely, CAPI/IE should reduce waiting periods for batch edits, reduce time editing those batch edits, and reduce out of pocket costs from the batch mainframe edits. Also, more savings should come as programers, statisticians, enumerators, managers, the entire organization; gains CAPI experience.

Interview Time

Differences in interview times are a negligible part of survey costs. Interviewing takes a small portion of the enumerators' time. Enumerators spend much more time traveling and finding sampled farmers. Interview length is a respondent burden issue, not a cost issue.

The instrument measured CAPI interview lengths automatically for the entire interview and by sections within the questionnaire. Enumerators do not record paper interview times so no comparison was made.

From observation, CAPI slowed the interview within a section that comprises a large table. For other parts of the questionnaire, particularly parts with automated skips, CAPI was faster. CAPI also saved time because it automated calculations.

Enhanced software or faster computers will soon further speed CAPI. Enumerators also suggested specific features that programmers could use to speed CAPI. Relative CAPI and paper interview times depend upon the survey, software, hardware, the intensity of interactive edits, and the imagination of survey instrument designers.

Conclusions

Although CAPI has high initial costs for hardware and training, it shows potential in reducing office work during a peak work period. CAPI can greatly reduce office keying and hand editing. CAPI/IE shows substantial gains over both the paper method and over IE without CAPI.

Results from the 1993 CAPI/IE June Area Frame supported key parts of previous projections of CAPI reductions in office work load. The unpolished instrument was facing its first operational test. Future CAPI/IE development should yield even more savings.

This promising work is fledgling cost analysis, not the final word. Cost estimates should be revaluated on an ongoing basis as the organization refines and expands the CAPI/IE process. The organization should use these cost analyses as a tool to help plan CAPI use.

This work focused on a particular aspect of the CAPI decision: cost. It did not quantitatively include timeliness and data quality. The organization must decide how to weigh all considerations before choosing whether to use CAPI.

Acknowledgments

Mark Pierzchala and Ann Romeo were co-conspirators with the author in coding the June Area Frame survey instrument.

References

Clayton, R. and Harrell L., Bureau of Labor Statistics, "Developing a Cost Model for Alterative Data Collection Methods: Mail, CATI, and TDE". Military and Government Speech Technology Conference. November 1989, Arlington, Virginia.

Eklund, B. Mobile Computerized Data Entry and Edit for the Livestock Prices Received Survey. page 5, SRB Research Report No. SRB-91-10, Washington D.C., USDA, National Agricultural Statistics Service, November, 1991.

Eklund, B. Computer Assisted Personal Interviews: Experience and Strategy. page 11, STB Research Report No. STB-93-01, Washington D.C., USDA, National Agricultural Statistics Service, February, 1993.

Manners, T. "The Development and Implementation of Computer Assisted Interviewing for the British Labour Force Survey". Computer Assisted Survey Information Collection: International Progress, Special Contributed Papers for the American Statistical Association 1991 Joint Statistical Meetings. August 1991, Atlanta, Georgia.

COMPARISON OF TELEPHONE INTERVIEWING METHODS IN THE QUARTERLY APPAREL SURVEY

Richard Sigman, Suzanne Dorinski, and Bob Tinari, U.S. Census Bureau[1]
R. Sigman, Census Bureau, ESMD, Room 3015-4, Washington D.C. 20233

INTRODUCTION

Quarterly Apparel Survey

The U.S. Census Bureau's Quarterly Apparel Survey (QAS) is one of the 75 different Bureau surveys that make up the Current Industrial Reports Program. This program provides timely and accurate intercensal estimates of production and shipments of specific manufactured products. These estimates are used by government agencies for economic policy analysis and by the private sector for market analysis, forecasting, and decision making.

The QAS collects data from a sample of domestic clothing manufacturers. The QAS sample is a cut-off sample of establishments that reported apparel shipments in the previous quinquennial economic census. The Census Bureau mails out the QAS questionnaire one week after the end of the quarter. Three weeks later, firms that have not responded are assigned to telephone follow-up.

This paper describes an experiment that we conducted between October 1991 and September 1992 that compared two different telephone interviewing methods for following up QAS nonrespondents. We compared paper-and-pencil interviewing (PAPI) with computer-assisted telephone interviewing (CATI).

CATI

Nichols (1983) describes the development and testing of CATI at the U.S. Census Bureau. In the early 1980's, the Bureau conducted two experiments that compared CATI and PAPI. The first experiment was part of the National Survey of Natural and Social Scientists and Engineers, and the second experiment was part of the 1982 Census of Agriculture. Both experiments involved the use of telephone interviewing to collect data from non-respondents to a mail survey. Also, both experiments involved separate CATI and PAPI facilities--the CATI interviewing was conducted from a research and development facility located in Suitland, Maryland, whereas PAPI interviewing was conducted by the Bureau's Data Preparation Division located in Jeffersonville, Indiana. Ferrari (1984, 1986) and Ferrari, et. al. (1984) describes these two experiments in detail, and Nichols and Groves (1986) and Groves and Nichols (1986) summarize the major findings.

Other government statistical agencies have also compared CATI and PAPI during the initial development of their CATI systems. House and Morton (1983) describes a comparison of CATI with PAPI conducted by the National Agricultural Statistical Service, and Catlin and Ingram (1988) discuss a study conducted by Statistics Canada.

In 1985, the Bureau established a centralized CATI facility in Hagerstown, Maryland, called the Hagerstown Telephone Center (HTC). Most of the CATI work at HTC was, and still is, associated with household surveys. Therefore, the peak workload at HTC occurs in the evening. To increase the utilization of HTC during the daytime, HTC began in the late 1980's the CATI follow-up calls for QAS. Daytime calling was further increased in early 1992 with the addition of CATI follow-up calls for the Bureau's M3 Survey, which is a monthly survey of manufacture's shipments, orders, and inventories.

The initial computer hardware at HTC was a Digital Equipment Corporation minicomputer system. In late 1991, however, the CATI hardware was changed to a microcomputer-based system. To facilitate this transition, HTC switched the QAS work from minicomputer CATI to PAPI for a number of quarters prior to the migration to microcomputer CATI. We began the CATI-versus-PAPI experiment described in this paper in October 1991. This was the first month that microcomputer CATI was available for QAS follow-up calls.

Objective of the experiment

The objective for this experiment differed from that of the CATI/PAPI experiments conducted in the early 1980's. The early experiments determined if CATI should be upgraded from a research activity to a production operation. As a result of these early experiments, the Bureau made CATI operational by creating HTC in 1985 and later opening a second CATI center in Tucson, Arizona in 1992. The purpose of our experiment, however, was to obtain measurements related to data quality and interviewing costs for the Bureau's operational CATI system as it is used to make QAS follow-up calls. PAPI's role in the experiment was to provide a point of reference for the measurement process. From the start of experiment, it was known that when the experiment was finished HTC would use CATI, and not PAPI, for QAS.

DESIGN OF THE EXPERIMENT

We used a cross-over design involving two teams of interviewers at HTC to compare CATI and PAPI for four quarters of the QAS. In the experiment's first quarter, team 1 used CATI, and team 2 used PAPI. In each of the following quarters, we switched the CATI/PAPI assignments. If a team used CATI one quarter, it used PAPI the next quarter; and vice versa.

At the start of the experiment, HTC had 10 interviewers working on QAS. Seven were experienced, and three were newly hired. We asked the QAS supervisor at HTC to create two teams of five interviewers each. Though we had wanted the two teams to be approximately equal with respect to their distributions of experience, this was not possible. More training resources were available for CATI than for PAPI. Consequently, the newly hired interviewers were assigned to CATI, and not PAPI, for the first quarter of the experiment--that is, assigned to team 1. This caused team 2 to be more experienced than team 1.

There were no changes in team membership for the first two quarters of the experiment. In the third quarter, there were only three CATI interviewers, but there were five PAPI interviewers. Also in the third quarter, one of the five PAPI interviewers was new to the CATI/PAPI experiment, but not new to telephone interviewing nor to apparel surveys. This interviewer had made CATI follow-up calls for the Counterpart Apparel Survey (CAS) the previous quarter. The CAS is an annual survey of the small apparel manufacturers that are not part of the QAS sample. In the fourth quarter of the experiment, there were again five CATI interviewers and five PAPI interviewers, but only nine of these ten were the same as the interviewers in the first two quarters. All together, there were seven interviewers--three on team 1 and four on team 2--that participated all four quarters.

For each quarter of the experiment, the Data Preparation Division (DPD), in Jeffersonville, Indiana, mailed the QAS questionnaire to the establishments in the QAS sample. The QAS questionnaire asks for the value of shipments during the quarter and the associated quantities of goods for 74 different types of clothing. All the different types of clothing do not appear on every questionnaire, however, because the portion of the questionnaire that requests specific product information is custom printed for each respondent. Only those types of clothing the respondent reported in previous quarters appear along witb the respondent's data for the two previous quarters. The questionnaire contains blank lines for the respondent to "write-in" additional types of clothing that are not preprinted. For each type of clothing that the respondent reports, the questionnaire asks that the reported quantity and value be disaggregated into specific size groups and that quantity be also disaggregated into particular fabric types. The questionnaires mailed out by DPD also contained the size and fabric disaggregations reported for the two previous quarters.

DPD mailed the QAS questionnaires approximately one week after the end of the survey period. Then approximately 14 days after mailout, the QAS staff in Suitland, Maryland, created a follow-up file of establishments that had not returned a completed questionnaire. For the four quarters of the experiment, this file contained 1248, 1317, 1387, and 1056 establishments, respectively. The file also contained the phone number of each establishment's contact for QAS data. In a few cases, the phone number was for the establishment's accountant and not the establishment. Because some accountants do the books for more than one apparel manufacturer, it was possible for one phone number to be associated with a group of QAS establishments, which HTC calls a "multiple". Consequently, we randomly assigned phone numbers in the follow-up file to either CATI or PAPI--half the numbers to CATI, and the other half to PAPI. We randomly assigned phone numbers instead of establishments so that "multiples" would not be part CATI and part PAPI.

Though the follow-up file was created approximately two weeks after mailout, the calling from HTC did not start until approximately three weeks after mailout. The reason for the one week lead time was that after creating the follow-up file, the QAS staff provided DPD with a list of establishments assigned to PAPI. DPD printed a second copy of the questionnaire for the listed establishments and then air expressed the questionnaires to HTC.

HTC started the CATI and PAPI calling approximately three weeks after mailout. Prior to and during the calling period, the QAS staff created a daily "no call" file. This file listed the establishments from which mailed questionnaires had been received after the establishments were assigned to follow-up. HTC made no additional calls to establishments that appeared on the no-call file.

In addition to the difference in the method of recording data, there were a number of other differences between CATI and PAPI with respect to the interviewing tasks associated with QAS follow-up. One difference was that the CATI instrument for QAS did not contain as much historical information as the PAPI questionnaire. For reasons of screen design and the size of files transmitted to HTC at the beginning of the calling period, the CATI instrument displayed previous quarters' data only for total value and total quantity by type of clothing and did not display historical information for the disaggregations by size and fabric. The PAPI questionnaire, on the other hand, contained historical data for both totals and disaggregations.

Another difference between CATI and PAPI was the method of call scheduling. For CATI, the computer assigned each call to the first interviewer available to make the call. Thus, an establishment assigned to CATI that required several call backs might be called by several different interviewers. For PAPI, on the other hand, a call was made to a given establishment by the interviewer whose desk the corresponding PAPI questionnaire was on. Thus, PAPI call backs made during the same interviewing shift were usually by the same interviewer.

The CATI system created a log file containing for each call the interviewer making the call, the start and stop times of the calls, and the disposition of the call. The PAPI interviewers, on the other hand, manually recorded this information on provided data sheets.

MEASUREMENTS

We compared CATI and PAPI with respect to the following measures:

- M1. Average number of calls to non-contacts.
- M2. Average number of calls to refusals.
- M3. Average number of calls to cooperators.
- M4. Non-contact rate.
- M5. Time-sheet minutes per completed case.
- M6. Refusal rate.
- M7. Average length of final call.
- M8. Questionnaire-level edit failure rate.
- M9. Questionnaire-level analyst adjustment rate.
- M10. Item-level edit failures per failed questionnaire.
- M11. Item-level analyst adjustments per analyst-adjusted questionnaire

For both CATI and PAPI, several interviewers can be involved at different times in contacting the same establishment. Thus, measures M1 through M5 describe team performance and cannot be associated with individual interviewers. Measures M6 through M11, however, can be associated with individual interviewers. In addition, we calculated measures M8 through M11 for the mail responses.

For PAPI, measures M8 through M11 measure the combined effects of interviewers and keypunchers if they are calculated from the data received from DPD. We were able to separate these effects for the batch-edit measures M8 and M10 because all the batch edits were simple balance tests--that is, tests that checked that disaggregated data added to aggregated data. We examined each edit failure in the PAPI data. By comparing the corresponding keypunched data with the interviewers' hand-written entries on the PAPI questionnaire, we were able to classify each edit failure as either an interviewer error or a keypunch error. Consequently, for PAPI we calculated measures M8 and M10 two different ways: with data received from DPD and with data corrected for keypunch errors.

INFERENCE APPROACHES

Fienberg and Tanur (1988) describe the following three approaches for making statistical inferences about an experiment imbedded within a sample survey:

(a) Model-based approach with interviewers treated as fixed effects. This approach assumes there exists a model describing the data from the experiment. The model's systematic effects include treatment effects, interviewer effects, and blocking factors. Variation arises from model error--that is, the unaccounted for departures of actual data from the modelled systematic effects. Because interviewers are treated as fixed effects, the conclusions based on this approach cannot be generalized to situations involving a different set of interviewers.

(b) Design-based approach. This approach assumes that the only source of variation is the sample-design--that is, the survey's sampling of establishments followed by the random assignment of sample establishments to treatments. The conclusions from this approach can be generalized only to different sample selections and assignments to treatments in which all the other conditions of the experiment are exactly the same.

(c) Model-based approach with interviewers treated as random effects. This approach assumes that there are two sources of variation: model error and interviewer effects. Because the interviewers are thought of as a sample from a population of interviewers, the conclusions based on this approach can be generalized to other interviewers selected from the same population of interviewers. (Obviously, the definition of the assumed population of interviewers is very important.)

Approaches (a) and (c) require interviewer-level measurements. Consequently, we used approach (b) to make inferences about the calculated team measures M1 through M4.

ANALYSIS OF TEAM MEASURES

Because the teams were the same in quarters 1 and 2, we were able to compare CATI and PAPI within teams for the first two quarters. We were able to make this comparison by defining the following quantities:

M(CATI,Qn) = particular measure for CATI used to collect data from the entire population of follow-up establishments in quarter Qn ,

M(PAPI,Qn) = particular measure for PAPI used to collect data from the entire population of follow-up establishments in quarter Qn ,

M(MODE,.) = [M(MODE,Q1)+M(MODE,Q2)]/2, where MODE=CATI,PAPI,

M(.,Qn) = [M(CATI,Qn) + M(PAPI,Qn)]/2, where n=1,2,

E(CATI) = CATI explanatory effect
= [M(CATI,.) - M(PAPI,.)]/2,
E(Q2) = quarter 2 explanatory effect
= [M(.,Q2) - M(.,Q1)]/2, and
E(T2) = team 2 explanatory effect
= [(M(PAPI,Q1) + M(CATI,Q2))
-(M(CATI,Q1) + M(PAPI,Q2))]/4 .

Table 5 of Sigman, et. al. (1993) contains estimates of E(CATI), E(Q2), and E(T2), which we denote e(CATI), e(Q2), and e(T2), respectively, and also contains an estimate of the overall average:

AVG = [M(CATI,.) + M(PAPI,.)]/2
= [M(.,Q1) + M(.,Q2)]/2 .

The explanatory effects are so named because they satisfy the following relationship:

M(MODE,Qn) = AVG + E(CATI)*X1(MODE)
+ E(T2)*X2(Tn)
+ E(Q2)*X3(Qn)

where

X1(MODE) = -1 for PAPI, +1 for CATI;
X2(Tn) = -1 for team 1 and +1 for team 2; and
X3(Qn) = -1 for quarter 1, +1 for quarter 2.

For measures M1 through M4, the following explanatory effects are statistically significant ($p \leq 0.10$) in the sense that the corresponding 90 percent confidence intervals do not include zero:

-CATI increased the average number of calls to refusals. The overall CATI average was 5.8 calls per refusal compared to the overall PAPI average of 3.8 calls per refusal.

-Assignment to team 2 decreased the following measures:

-The average number of calls to non-contacts was 5.8 calls for team 2 compared to 9.6 calls for team 1.

-The average number of calls to refusals was 3.8 calls for team 2 compared to 5.8 calls for team 1.

-The non-contact rate was 3.1% for team 2 compared to 5.9% for team 1.

-The effect of quarter 2 on the average number of calls per cooperator was a small decrease from 4.3 calls to 3.9 calls.

For measure M5, average time-sheet minutes per completed case, we were not able to make any type of statistically valid inference. The reason for this is that the denominator (the number of completed cases) could not be calculated for individual interviewers because several interviewers may contribute toward completing one case. This ruled out inference approaches (a) and (c). Because we could not calculate a sampling error for the numerator (total time-sheet minutes charged by interviewers), inference approach (b) was ruled out. Consequently, there is little we can say about measure M5 other than simply describing the results. In quarters 1 through 3, the within-quarter CATI effect for measure M5 was positive, indicating that the average time-sheet minutes per completed case was greater for CATI than for PAPI. In quarter 4, however, the within-quarter CATI effect was negative, indicating that average time-sheet minutes charged per completed case was less for CATI than for PAPI. The within-quarter CATI effects decreased as the experiment progressed: +6 minutes in quarter 1, +3 minutes in quarter 2, +1.5 minutes in quarter 3, and -5.5 minutes in quarter 2. In the explanatory-effects analysis for quarters 1 and 2, the explanatory effect for interviewing method was the largest in absolute value. The overall CATI average for measure M5 (in quarters 1 and 2) was 40.0 minutes compared to 31.0 minutes for PAPI, a ratio of 1.3 to 1. The explanatory effect for team 2 (in quarters 1 and 2) was -1.5 minutes.

ANALYSIS OF INTERVIEWER MEASURES

We used inference approach (c) (i.e. model-based analysis with interviewers treated as random effects) to analyze interviewer measures M6 through M11. One reason for our selecting this approach was its greater external validity compared to inference approaches (a) or (b). Another reason was that in the experimental design literature treating the subject (i.e, interviewer) as a random effect is the standard procedure for analyzing a cross-over design (Milliken and Johnson, 1984, Chapter 32).

We assumed there existed an explanatory effects model for each interviewer measure. Each model contained fixed effects for interviewing method, team assignment, and quarter of the experiment and contained two error terms for the random interviewer effect and the model error. The models were fitted to two different data sets. One data set (data set A) was the set of the interviewer measures for quarters 1 and 2 for the 10 interviewers assigned to QAS in both of these quarters. The other data set (data set B) was the set of interviewer measures for quarters 1 through 4 for the seven interviewers assigned to QAS in every one of the four quarters.

The models fitted to data set A had the form:

m(MODE,Qn,I) = particular measure for telephone method MODE in quarter Qn associated with interviewer I
= AVG + E(CATI)*X1(MODE)
+ E(T2)*X2(Tn)
+ E(Q2)*X3(Qn)
+ interviewer-effect(I)
+ model-error(MODE,Qn,I),

where the AVG, E, and X terms have the same definitions as in our explanatory-effects analysis for the team measures.

The models fitted to data set B had the following form:

$$\begin{aligned} m(MODE,Qn,I) = AVG &+ E(CATI)*X1(MODE) \\ &+ E(T2)*X2(Tn) \\ &+ E(Q1)*X3(Qn) \\ &+ E(Q2)*X4(Qn) \\ &+ E(Q3)*X5(Qn) \\ &+ E(Q4)*X6(Qn) \\ &+ \text{interviewer-effect}(I) \\ &+ \text{model-error}(MODE,Qn,I)\ . \end{aligned}$$

The first three terms and the last two terms are the same in the two sets of models. The only differences for the data set B models are that X3, X4, X5, and X6 are indicator functions for quarters 1, 2, 3, and 4, respectively, and $E(Q1)+E(Q2)+E(Q3)+E(Q4) = 0$.

Because of the inherent structure of the cross-over design, a non-zero team effect indicates the existence of interaction between the interviewing method and the quarter of the experiment. Consequently, Milliken and Johnson (1984, Chapter 32) recommend that the experimenter first test for a significant team effect (which they call a "sequence effect"). The appropriate denominator mean square for this test is that for the INTERVIEWER*TEAM interaction, which estimates a weighted sum of the among-interviewer variability and the model-error (i.e. within-interviewer) variability. If the team effect is not statistically significant, then the effects for interviewing method and the quarter of the experiment can be tested with the error mean square in the denominator of the F tests. Because the error mean square estimates only the within-interviewer variance and does not involve the among-interviewer variance, the individual interviewers serve as blocking factors for the testing of the non-team effects. (Blocking on interviewers is the method recommended by Fienberg and Tanur (1988) for imbedding an experiment in a sample survey.)

Table 8 of Sigman, et. al. (1993) presents the results from fitting the explanatory-effects models for interviewer measures to data set A and contains explanatory effects for the mail measures for quarters 1 and 2. (We calculated the standard errors for the mail-return measures from the quarter-to-quarter variability.) Table 9 of Sigman, et. al. (1993) analyzes the interviewer measures in data set B and the mail-return results for quarters 1 through 4.

For all of the interviewer measures, we found the team effect not statistically significant in both data sets A and B. The following differences in overall averages are statistically significant ($p \leq 0.10$) in both data sets:

-The questionnaire-level edit failure rate for telephone interviewing (6.8% in data set A, 8.3% in data set B) is less than that for mail returns (54.1% and 50.6%).

-The questionnaire-level analyst adjustment rate for PAPI (8.2% and 10.1%) is less than that for mail returns (27.2% and 27.4%).

-The questionnaire-level analyst adjustment rate for CATI (11.2% and 14.5%) is less than that for mail returns (27.2% and 27.4%).

-The item-level edit failures per failed questionnaire are less for telephone interviewing (1.3) than for mail returns (3.2).

In addition, in data set B the average number of item-level analyst adjustments per adjusted questionnaire for PAPI (2.2) was significantly less than that for mail returns (3.3). This comparison was not statistically significant in data set A. It appears, however, that this result is due to data set B's greater number of degrees of freedom for model error.

The following CATI explanatory effects are statistically significant ($p \leq 0.10$) in both data sets:

-The average length of the final call for CATI (7.5 minutes and 8.0 minutes) is greater than that for PAPI (6.1 minutes and 6.0 minutes).

-When the PAPI data is not corrected for keypunching errors, the questionnaire-level edit failure rate for CATI (4.4% and 5.5%) is less than that for PAPI (9.2% and 11.1%). When the PAPI data is corrected for keypunching errors, however, there is no significant difference between the questionnaire-level edit failure rates for CATI and PAPI.

-The questionnaire-level analyst-adjustment rate for CATI (11.2% and 14.5%) is greater than that for PAPI (8.2% and 10.1%).

-The average number of item-level analyst adjustments per adjusted questionnaire for CATI (3.0 and 3.8) is greater than that for PAPI (2.4 and 2.2).

In addition, in data set A the refusal rate for CATI (1.9%) is significantly less than the refusal rate for PAPI (3.1%). This difference in refusal rates is not statistically significant in data set B, however.

A possible explanation for the increased number of analyst adjustments for CATI compared to PAPI is that the CATI instrument may be preventing the interviewer from making data entry corrections, which the analyst must make at a later time based on information the interviewer types into the "interviewer's notes" section of the CATI instrument. We were unable to obtain the CATI-interviewer notes for the four quarters of the experiment, but we were able to obtain this information for the quarter immediately following the conclusion of the experiment. We found that 21 of the interviewers notes provided information to the QAS analyst about the inaccuracy of the data entered for the current quarter. Only seven of these, however, requested the analyst to correct data entry errors that the interviewer was unable to correct. The other 14 notes provided background information to the analyst that either would assist the analyst in adjusting the data, if necessary, based on the analyst's subject-matter knowledge or would prompt the analyst to call the respondent.

CONCLUSIONS

We were able to statistically test for differences between CATI and PAPI for 10 out of 11 comparison measures and found the following CATI effects to be statistically significant ($p \leq 0.10$):

- -CATI increased the number of calls to refusals (quarters 1 and 2).
- -CATI decreased the refusal rate (quarters 1 and 2).
- -CATI increased the average length of the final call (all quarters).
- -CATI decreased the questionnaire-level edit-failure rate (all-quarters). (We found, however, that the source of this CATI effect is keypunching errors and not interviewing errors in the PAPI data.)
- -CATI increased the number of analyst adjustments (all quarters). An examination of the CATI-interviewer notes suggest that a minority of the analyst's additional adjustments to CATI data are to correct data entry errors. We surmise that the majority of the additional adjustments arise from (1) more abundant interviewer notes from CATI and (2) the easier microcomputer-based retrieval of CATI interviewer notes compared to filing and retrieving the PAPI questionnaires.

We found that assignment to team 2 significantly ($p \leq 0.10$) decreased the average number of calls to non-contacts, the average number of calls to refusals, and the non-contact rate. This may have been caused by the fact that team 2 consisted of all experienced interviewers, but other factors such as within-team dynamics and work schedules may also have caused this.

We were unable to statistically test between CATI and PAPI for the average time-sheet minutes per completed case. This measure, however, was greater for CATI in the first three quarters of the experiment, and was less for CATI in the fourth quarter. The difference in this measure between CATI and PAPI decreased from the beginning of the experiment to the end of the experiment. Perhaps this indicates a CATI learning effect since the first (or second) quarter of the experiment was the first quarter of use of microcomputer CATI by QAS interviewers.

In comparing the data collected by HTC with the data received by mail, we found that HTC decreased the number of edit failures in the data plus decreased the number of questionnaires that received one or more analyst adjustments.

REFERENCES

Catlin, G. and Ingram, S. "The Effects of CATI on Costs and Data Quality: A Comparison of CATI and Paper Methods in Centralized Interviewing," in R.M. Groves, P.P.Biemer, L.L. Lyberg, J.T. Massey, W.L. Nicholls II, and J. Waksberg (eds), Telephone Survey Methods, New York, Wiley, 1988.

Ferrari, P.W. "Preliminary Results from the Evaluation of the CATI Test for the 1982 National Survey of Natural Scientists and Engineers," Unpublished research report, U.S. Bureau of the Census, Washington, D.C., 1984.

Ferrari, P.W. "An evaluation of Computer-Assisted Telephone Interviewing Used During the 1982 Census of Agriculture," Unpublished research report, U.S. Bureau of the Census, Washington, D.C., 1986.

Ferrari, P.W., Storm, R.R., and Tolson, F.D.. "Computer Assisted Telephone Interviewing," Proceedings of the Survey Research Methods Section, American Statistical Association, 1984, pp. 594-599.

Fienberg, S.E. and Tanur, J.M. "From the Inside Out and the Outside In: Combining Experimental and Sampling Structures," The Canadian Journal of Statistics, Vol. 166, 1988, pp. 135-151.

Groves, R.M. and W.L. Nichols II. "The Status of Computer-Assisted Telephone Interviewing: Part II - Data Quality Issues," Journal of Official Statistics, Vol. 2, 1966, pp. 117-134.

House, C.C. and Morton, B.T. "Measuring CATI Effects on Numerical Data," Proceedings of the Survey Research Methods Section, American Statistical Association, 1983, pp. 135-138.

Milliken, G. A. and Johnson, D.E. Analysis of Messy Data. Volume 1: Designed Experiments, New York, Van Nostrand, 1984.

Nichols II, W. L. "Development of CATI at the U.S. Census Bureau," Proceedings of the Survey Research Methods Section, American Statistical Association, 1983, pp. 642-647.

Nichols II, W.L. and Groves, R.M. "The Status of Computer-Assisted Telephone Interviewing: Part I - Introduction and Impact on Cost and Timeliness of Survey Data," Journal of Official Statistics, Vol. 2, 1966, pp. 93-115.

Sigman, R., Dorinski, S., and Tinari, B. "Comparison of Telephone Interviewing Methods in the Quarterly Apparel Survey," ESMD Report Series ESMD-9302, Bureau of the Census, Washington, D.C, May 1993.

---

[1] This paper reports the general results of research undertaken by Census Bureau staff. The views expressed are attributable to the authors and do not reflect those of the Census Bureau.