

RESEARCH AND ANALYSIS TO BETTER UNDERSTAND DATA COLLECTION ACTIVITIES

Judy Lee and Tony LaBillois, Statistics Canada

ABSTRACT

Efficiency of data collection activities has historically been difficult to quantify and management information for improving collection processes was hard to produce in a standardized, meaningful and timely fashion across multiple surveys. With the advent of Computer Assisted Interviewing and the standardization of the software used to collect survey data, a wealth of information on the data collection process is now available. The BLAISE system records the activities of the interviewers as they enter and exit a case during the collection cycle. At Statistics Canada, a Research and Analysis Group is now responsible for coordinating and conducting specific projects on key issues related to data collection and for providing collection management tools for planning purposes. This paper will explain the scope, objectives and challenges of research and analysis on operations at Statistics Canada. It will also provide some examples of the work accomplished to date.

KEY WORDS: Computer-Assisted Interviewing, Data collection, Process Improvement.

RÉSUMÉ

L'efficacité des activités de collecte de données a traditionnellement été difficile à quantifier et l'information de gestion pour améliorer la collecte était difficile à produire de façon normalisée, sensée et actuelle pour un grand nombre d'enquêtes. Avec l'arrivée de l'interview assistée par ordinateur et la normalisation du logiciel de collecte, une montagne d'information sur la collecte est maintenant disponible. BLAISE, le système de collecte, enregistre les activités des intervieweurs lors qu'ils accèdent ou sortent d'un cas pendant le cycle de collecte. À Statistique Canada, un groupe de recherche et d'analyse est présentement responsable de coordonner et de mener certains projets spécifiques portant sur des problèmes importants liés à la collecte des données de même que de fournir des outils de gestion de la collecte utilisables à des fins de planification Cet article donnera un aperçu du mandat, des objectifs et des défis liés à la recherche sur les opérations à Statistique Canada. Il fournira également quelques exemples du travail accompli jusqu'à maintenant.

MOTS CLE: Amélioration des processus; collecte de données; interview assistée par ordinateur.

1.0 INTRODUCTION

A few years ago, Statistics Canada consolidated its resources and activities related to the development of Computer-Assisted Interviewing (CAI) systems. The Agency also decided to adopt one unique software to be used for the development of all new CAI applications and for the conversion of existing applications. BLAISE was chosen for this purpose and an effort of standardization began. BLAISE is a software suite for survey collection and processing developed by Statistics Netherlands and distributed by Westat in North America. For more information, you can visit the websites given in the references. Resources to expand the research and analysis related to data collection activities were also allocated at the same time.

Prior to these events, many management reports specific to a survey or to a type of survey were available for collection managers to assess the ongoing collection efforts. However, only few special studies using quantitative information on collection activities had been produced while a lot of anecdotal information and comments based on the experience of the collection management staff were hard to verify or to generalize with clear facts that would provide a real overview of the collection process.

The allocation of resources for research and analysis on data collection activities, the effort of standardization and the growing interest of managers to get more quantitative information for decision making about collection activities all

Tony Labillois, Statistics Canada, Business Survey Methods Division, Statistics Canada, Ottawa, Ontario, K1A 0T6

contributed to the creation and the success of a Research and Analysis Group which quickly showed that much could be learned by better understanding and using this type of information.

This paper will present a brief description of the scope and objectives of the operations research program, the data sources used for conducting the analysis, and a few examples of the projects that have been explored so far. Finally, the challenges and priorities for the near future will be discussed.

2.0 SCOPE AND OBJECTIVES OF THE OPERATIONS RESEARCH PROGRAM

Quality control and quality assurance initiatives were implemented for many years at Statistics Canada. Interviewer monitoring has been conducted during CATI collection activities in our collection sites using sampling and quality control methods and the results are compiled in the Quality Control Feedback System (QCFS) to help improve the quality of the data and to help the training of interviewer. The scope of the evaluation of our collection operations was considerably extended three years ago to include the ability to assess many aspects of our activities for all our surveys. These aspects are linked to the recent availability of data on data collection.

The primary objective of the group is to collect and compile quantitative information on different components of the collection process. The data can be about many things: editing and follow-up at the data collection stage, call history information, number of accesses to a case, reasons and outcomes for accessing a case, duration of interviews, number of interruptions and of repetitions during a question, etc.

A secondary objective consists in providing tools to collection managers to assist them in the management of data collection activities. In addition to the regular collection progress reports, they would have access to a series of flexible dynamic tools on their desktop to examine various aspects of the process including characteristics of “interviewers” performance.

The Research and Analysis Group also work on specific research projects to show the potential of the use of collection related information and to answer questions posed from operations managers or from subject matter representatives. Additionally, exploration and analysis of the data leads to new questions and lines of research. These activities generate interest towards the program and their results often lead to discussions regarding current practices and changes in the production environment or to the inclusion of the new information review process in other survey collection activities.

Our clients include senior management interested in efficiencies, collection managers who want to know how their staff is doing and subject matter specialist who need to know how their money is spent or if their questionnaire flows smoothly. The program helps to show facts that impact on future collection systems development, on collection procedures or activities and on budgets allocated to these tasks.

In summary, the focus is on optimizing the collection effort in all of our business, agriculture or household surveys and in each of our nine data collection sites. The goal is to produce information that would lead to better ways to conduct the collection activities.

3.0 DATA SOURCES

Reliable, timely, coherent, standardized and easily accessible data files have to be available before being able to do analysis on data collection activities. With hundreds of surveys per year either on businesses, farm operations or households collected in multiple collection sites, it is always a challenge to amass all that information over time for all reference periods. Because Statistics Canada is a very centralized statistical agency and because the development of our computer-assisted applications is centralized in the same division as the research and analysis group, it is possible to coordinate these activities by establishing clear processes, standards and documentation. It is a continuous effort to

ensure that these rules are followed and to assess the quality of the data sources to guarantee the availability of files that are “fit for analysis”.

So far, we are mainly using the BLAISE Transaction History (BTH) files for our projects. During a survey collection period, each time an interviewer accesses a case using the CAI application, a transaction record is written in the BTH file. This record contains a series of variables including dial date, dial time, time duration, outcome code, interviewer ID that can be used later to prepare progress reports or to explore various aspects of how collection was conducted. For example, we can study the amount of times a case was accessed and how much time was spent on a case. BTH files are slightly different depending on the type of survey. Documentation on the structure and content of the BTH files can be found in the documentation mentioned in the references.

When setting up BLAISE applications, an audit trail function can be turned on. With this function on, a short record is generated that records all keystrokes for that session. By combining these records, it is possible to find out things like the average time to complete a section of the questionnaire, how many times the interviewers had to go back to a specific question or to monitor interview patterns to detect possible falsification.

For some projects, we need to combine the collection related data with other sources. These sources may include sample files or survey data files with the BTH files when the purpose is to explore how the respondents’ characteristics might have an impact on data collection. Document control files or files from data capture of cases reintegrated in the CATI application for review and potential follow-up are useful when we want to assess the performance of edits in the collection process.

We are currently working to find the best approach to easily store and access the data from collection activities. The idea is to eventually have a sustainable model for the Collection Activities Repository (CAR). We have also developed prototypes of Reports and Analysis of Collection Events Systems (RACES) for business and household surveys. These prototypes developed in SAS/AF are often used to answer special ad hoc requests and to do some analysis. The ultimate goal is to have a very flexible, easy-to-use and timely tool for the staff in the group and also for other people involved in management of collection activities. There is still a lot of work to do on the CAR and on the RACES before we can reach that stage.

4.0 EXAMPLES OF RESULTS

Despite an evolving collection research infrastructure and a still young research program, we have undertaken many analysis projects and some of them have led to or shown significant changes in our data collection processes. Our goal is always to try to answer current questions or to shed some light on new issues. These projects help to show the potential and the relevance of having a research and analysis program dedicated to the operations areas. Three examples are presented below.

4.1 Review of the collection effort for the Annual Survey of Manufactures (ASM)

As part of a wide review and streamlining initiative started in 2002 at Statistics Canada, a Business Surveys Review Streamlining Committee was formed and was looking for information on costs and processes for each step of two businesses surveys selected as case studies. The committee had to prepare a detailed report and was interested in quantitative information on the Annual Survey of Manufactures (ASM) collection activities as they were accounting for the biggest portion of the total survey budget. More details can be found in the report entitled “Strategic Streamlining Initiative, Business Surveys Review”.

For reference year (RY) 2001, a score function had been implemented with some new procedures to identify critical units. This function attributes a score to the collection units according to their impact on last year’s estimates. Some units are identified with more priority for follow-up while some others with less importance are not.

By using the information in the BTH files, we were able to compare the activities recorded in the collection system for RY2000 and RY2001. It is important to keep in mind that activities like mailing out packages, preparation of questionnaires or scanning are not included in the effort recorded in the systems. Discussion with collection managers to ensure appropriate interpretation of the results and that they are put in the right context were very important.

We found that compared to RY2000, there was a 40% reduction (over 8000) of cases that were followed up in RY2001 despite similar sample sizes between the two years. In RY2001, only 60% of the cases were followed up compared to 97% in RY2000. In terms of effort or accesses to the cases, RY2001 effort was reduced by 36% and translated to a time reduction of 40% or a resource reduction of 4 persons per year.

The total number of access and total time can be broken down to their components: collection or follow-up before a questionnaire was received, edit or follow up after a questionnaire was received, or no data received effort, where there was followed up but no questionnaire was received. In comparing the distribution of these components for the two years, there was an increase of collection and edition effort by 3.5 and 4.9 percentage points respectively, and a decrease in no data received effort by 9 percentage points. This decrease in effort that results in no data translated to a time reduction of 14 percentage points. In short, we are reducing our “wasted effort” and shifting to improving our collection and editing efforts.

4.2 Editing strategy in the Unified Enterprise Survey (UES)

The UES now includes most of Statistics Canada’s annual business surveys all integrated in a unique collection and processing environment. Since RY2000, when questionnaires are received, data are captured in an application called Quick Data Entry (QDE) which has no editing capability. Then, data are loaded in a BLAISE application for review, editing and possible follow-up. By comparing the data file from QDE with the output data file from QDE and by using BLAISE files and indicators, we can monitor edit failures. We want to analyze edit rule performance and follow-up rates as well as the impact of changes due to editing on the survey estimates.

The results of this ongoing project led to the implementation of Quality Control on the Quick Data Entry process. The analysis of the effort involved in editing is very interesting; examination of confirmation rate of existing data; analysis over time from RY2001 onwards; etc. A review of edit rules by subject matter staff for future collection periods was also started in an attempt to reduce the number of rules and to adjust some parameters.

For example, we found that for the surveys in services industries, the number of edits has been reduced from 96 in RY2000 to 59 in RY2001 to 51 in RY2002. Correspondingly, over these three years, the follow-up rates (cases that failed at least one severe or mandatory edit and have to be followed up) have decreased from 89% to 73% to 59% respectively.

Edit failures lead to verification of data that requires much effort and time. We now have the ability to monitor the outcome of these verification efforts. For example, in the case of Repair and Maintenance Services RY2001, 55% of questionnaires failed at least one mandatory edit and 29%, at least one severe edit which required 1240 and 2658 verifications respectively. For the mandatory verifications, 61% were resolved, 7% remained unresolved, and when followed up with respondents, 27% “don’t know” and 5% “refused” to validate the questionable data. For the severe edits, 50% were resolved, and follow up with respondents yielded 34% “don’t know”, 3% ‘refused’, and 13% confirmation of the questionable data. The goal of edits is to reduce the number of confirmations and other unresolved verifications.

This type of analysis can be carried for all questionnaire types and at the edit level, with the aim of identifying those edits and industries where follow up on edits are yielding little change in the data. Furthermore, the next step of analysis would be to identify the magnitude of the change in data where changes did occur as a result of edit follow-up.

4.3 Patterns in completion of cases for the Survey of Labour and Income Dynamics (SLID)

The SLID is a longitudinal survey of households that started in the 90s with two separate annual collection periods for labour variables and for income variables. With a large sample and growing collection costs, the subject matter staff of this survey has a great interest in how collection activities could be more efficient. We looked at different collection related aspects of this survey so far and we will just give a few examples of our findings here.

One aspect that we looked at was the time slot of where cases were most likely to be resolved, or coded to “final” stage. If this pattern holds true throughout the entire survey production cycle and between survey cycles, then this survey could then be scheduled according to this best time scenario. For example, for SLID Income 2003 (RY2002), the average number of cases resolved per interviewer was the highest on Saturdays and the most productive time on Saturdays is between 12:00pm and 3:00pm. However, the time of day that is most productive for resolving cases is in the evenings, between 6:00pm to 11:00pm for all days of the week with the exception of Saturdays. Saturday evenings is the absolute worse time to call in the hope of finalizing and resolving a case.

There has been strong interest in the effort involved in refusal conversions for household surveys. For SLID Income 2003 (RY2002), the average calls per case to convert a refusal to complete ranged from 2.9 calls in Halifax and Sherbrooke to 4.2 in Toronto to 5.5 in Edmonton. The average time it took to convert a case from refusal to complete ranged from 14.8 minutes in Sherbrooke, to 16.4 minutes in Halifax, to 18 minutes in Edmonton, to 23.2 minutes in Toronto. However, when we look at the rate of conversions, the Edmonton was the most effective in that they were able to convert 60% of their refusals to completes, followed by Sherbrooke at 47%, Halifax at 41%, and Toronto at 31%.

Finally, we now have available data over time to compare over survey cycles. For example, in comparing both collection years of SLID Labour and Income by effort, or the number of accesses or calls to acquire usable data (status outcome codes of complete or partially complete), over 80% of the useable data were attained by the 15th call. Call attempts over 15 times resulted in additional 5% to 10% of usable data.

5.0 CHALLENGES

The fruits of establishing a real collection research and analysis team are just beginning to appear. They show a promising but challenging future for this stream of our activities. This section focuses on a few of the upcoming challenges for the program. The document entitled “Quarterly Program Review, Operations Research and Development Division” provides a more detailed description of the upcoming challenges.

The need to consolidate our information infrastructure is at the foundation of future activities. We need to ensure that it can successfully incorporate data from more and more surveys for more and more collection periods in addition to acquiring actual survey data for linkage purposes. We are exploring different ways to build that huge repository of data while we are also improving our data retrieval systems.

We also have to maintain and increase our ability, capacity and expertise to shed light on specific issues when we get special requests. Even with a very small team, we have to ensure continuous learning and renewed relevant outputs shared with our partners. This is what will help to maintain the momentum of interest created around the findings and issues that we have worked on so far. It is a key factor that can guarantee survival and growth of this program.

As we have a wide variety of users and clients, we have a responsibility to coordinate user requirements for the definition of relevant tools, regular reports and standardized indicators. We can help the coordination but we need their input and their buy-in to fully satisfy their analysis requirements. We feel that we are there to light the flame and provide the fuel that we will need together to look closely at the processes and to make joint decisions on tangible actions towards improvements or savings in collection activities.

Results of operations research are to be used for overall improvements of the collection processes. It is not to potentially serve for micro-management from subject matter specialists. Operations have to learn to use the tools to increase efficiency and to communicate their findings and show that they react appropriately by optimizing their use of resources.

REFERENCES

<http://www.cbs.nl/en/service/blaise/introduction.htm> (The BLAISE system homepage)

<http://www.westat.com/blaise/> (BLAISE overview)

<http://www.blaiseusers.org/> (International BLAISE User Group)

Blaise Transaction History Files for Business Surveys RY2001, RY2002, Operations Research and Development Division, Statistics Canada (internal document).

Blaise Transaction History File Requirements for Social Surveys, Operations Research and Development Division, Statistics Canada (internal document).

LaBillois, T. The integration of Electronic Data Reporting (EDR) and other new systems in the ongoing collection process, AMRADS Workshop on Automatic Data Collection, Rome, Italy, May 2002.

<http://amrads.jrc.cec.eu.int/>

Royce, D., Labillois, T., Leduc, J., Lepage J., Marcogliese, M., Zinnicker, D. – Strategic Streamlining Initiative, Business Surveys Review – Interim Report, January 9, 2003, Statistics Canada, internal document.

Quarterly program review of Operations Research and Development Division – April 2000 to November 2003 – November 2003, Statistics Canada (internal document).