Understanding Survey Data Collection through the Analysis of Paradata at Statistics Canada

François Laflamme

Statistics Canada, Business Survey Methods Division, francois.laflamme@statcan.gc.ca

Abstract

The increasing challenge for Statistics Canada is to collect timely and cost-effective data and to maintain a high level of quality despite the evolving population structure and behaviour, the increasing burden on both respondent and Statistics Canada staff as well as the change in the data collection cost structure in the Agency. In order to learn more about the various issues around the data collection process, a series of analysis of call history data for social surveys was performed over the last few months. This paper describes the results of these investigations upon this call history data in an attempt to better understand how data collection progresses through its cycle and to identify strategic opportunities for improvements.

Key Words: Paradata, call transaction history, attempts, contact rate, sequence of calls

1. Introduction

The development and implementation of computer-assisted data collection methods has provided a wide scope of process data recently called "paradata". Scheuren (2005) describes the paradata concept as follows: "Paradata, in short, is part of the computerized metadata that surround many large-scale government and other surveys". In other words, paradata refers to any information that describes the data collection process from beginning to end. For example, sample design, audit trail, call transaction history and cost information could be considered as part of paradata. In particular, the call transaction history which records each attempt made has become an invaluable source of information used by survey researchers to learn more about the data collection survey process.

Much of the research discussed in this paper relies on the call transaction history of many social CATI surveys. The paper begins with an introduction to the research including the objectives, a description of the paradata and a brief overview of the surveys that were used during this study. The next section examines the methodology as well as the main findings of these investigations with regards to the following issues: distribution of attempts, attempts versus time spent on cases, contact rate, sequence of calls, number of attempts to get a first contact, first contact versus interview, call scheduler features and some additional investigations. The last two sections present a summary of the lessons learned so far and describe briefly possible ideas for future research.

2. Past Literature

Previous analysis studies on call history patterns have identified some opportunities for improvements through a better "organization" of the call scheduling rules. Many of these papers present and discuss results related to the best times to call with regards to the time of the day and the day of the week. In particular, weekday evenings and weekends have been systematically found to have the best contact rates (Massey, Wolter, Wan and Liu, 1996; Reedman and Robinson, 1997, Laflamme and Mohl, 2006).

These studies mainly focused on one specific survey or a few surveys making it difficult to generalize the results to different surveys. The study discussed in the following section uses a wide scope of surveys (e.g., cross-sectional, longitudinal, Random Digit Dialling) in order to identify the general and specific data collection patterns for each type of survey.

3. Current Study

The main objectives of the study were to learn more about the data collection survey process; to evaluate this process and to identify strategic opportunities for improvements in the way surveys are conducted and managed in the CATI Regional Office centres of Statistics Canada across the country. In order to achieve these objectives, the current study has taken advantage of the current Statistics Canada data collection infrastructure in which the paradata are stored.

3.1 Paradata Source

The call transaction history records were extracted from the Data Integration and Production Planning (DIPP) database that maintains and updates the ongoing surveys on a daily basis with the most recent information. Since DIPP is a standardized database, it helps to perform timely and coherent comparative analyses of various surveys conducted in Regional Offices.

The call transaction history record that constitutes the basis of the current analysis is created each time a case is opened, either for data collection or other purposes. It contains relevant detailed information about each call or attempt made to contact each sampled unit during the data collection process. It includes information on the amount of time that the case was open, the person who worked on it, the result of the contact attempt and other data such as the method used to access the case.

The sample design information of each survey was also considered as a second source of paradata information. Therefore, the sample design information attached to each survey was used in conjunction with the call transaction history to enhance the analytical value of these investigations.

3.2 Investigated Surveys

The current investigations were performed simultaneously on many cross-sectional, Random Digit Dialling (RDD) and longitudinal surveys in order to identify general and specific data collection patterns of each type of survey.

In particular, most of the investigations and results presented in this paper were performed on the following surveys: General Social Survey (GSS12, GSS18, GSS19, GSS20 - the number besides GSS refers to the survey cycle), Canadian Tobacco Use Monitoring Survey (CTUMS), Travel Activities and Motivation Survey (TAMS), Survey of Labour and Income Dynamics (SLID), Youth In Transition Survey (YITS) and Canadian Community Health Survey (CCHS). Table 1 provides a very brief overview of some characteristics of each survey.

Table 1

Survey	Survey	Type of Frame	Interview	Data Collection
CCHS	Cross-sectional	Area, list & RDD	Personal Telephone	Monthly
SLID	Longitudinal	LFS*	Telephone	3 months
YITS	Longitudinal	LFS	Telephone	4.5 months
TAMS	RDD	RDD	Telephone	4 months
GSS**	RDD	RDD	Telephone	Monthly
CTUMS	RDD	RDD	Telephone	Monthly

* Labour Force Survey rotate-out sample

** GSS has a different cycle and subject matter each year

4. Study Highlights

This section describes the highlights of the initial investigations on the following issues. The graphs essentially refer to one survey but are generally applicable to others, unless otherwise mentioned.

4.1 Number of Contact Attempts

Table 2 provides some descriptive statistics about the distribution of attempts for some major Statistics Canada surveys. For TAMS with a cap on calls, the distribution of the number of attempts per case is comparable to those with no cap (in particular the other RDD surveys) until the highest percentiles at which time the non-capped surveys have some very large values. A cap on the number of calls or attempts permitted per case has recently been implemented for a number of surveys. This initiative has eliminated the possibility of interviewers making an extraordinary number of calls to a single case. However, for other surveys where no cap exists, especially longitudinal surveys or those requiring tracing, the number of attempts for certain cases can still be overwhelming. Table 3 shows the average number of attempts required by final status (respondent, refusal, non-respondent or out-of-scope) of the sampled units. This is another way to represent the amount of effort put on both responding and non-responding cases before finalizing them. As shown in the table below, generally we put twice the effort in terms of attempts to confirm a non-response than to achieve a response.

Table 2

Survey	Sample	Count	Mean	P75	P95	P99	0-25	26+	%	Total
CCHS 2006-CATI	110,787	1,176,189	10.6	13	38	62	61,666	3,879	5.9%	65,545
SLID- 2006	35,045	368,613	10.5	13	37	70	25,952	1,862	6.7%	27,814
YITS	37,379	521,744	13.9	18	45	80	27,775	2,884	9.4%	30,659
GSS 19	58,342	437,999	7.5	10	26	43	18,874	638	3.3%	19,512
TAMS	118,421	640,394	5.4	8	14	20	52,656	0	0	52,656

(Attempts columns: Count, Mean, P75, P95, P99; Respondent Attempts** columns: 0-25, 26+, %, Total)

* GSS and TAMS were Random Digit Dialing (RDD) survey and TAMS was part of the Cap on Calls initiative

** Respondent attempts represent the total number of attempts required to get the interview

Table 3

Average number of attempts required by type of cases				
Survey	Respondent	Refusal	Non-respondent	Out-of-scope
GSS18	5.1	6.6	11.4	2.9
GSS19	7.1	10.1	15.5	4.0
CTUMS 2005	3.2	4.9	9.1	2.3
TAMS	4.2	5.2	10.0	2.9

4.2 Attempts versus Time Spent
In order to better understand the overall data collection process, it is necessary to know where most of the data collection time and effort are spent. Figure 1 provides an example of the distribution of outcome codes and time spent for the whole GSS19 representing about 437,999 attempts. As expected, most of the time is spent to complete an interview (47%) even though it corresponds to only 4% of the attempts (i.e., the interview itself).
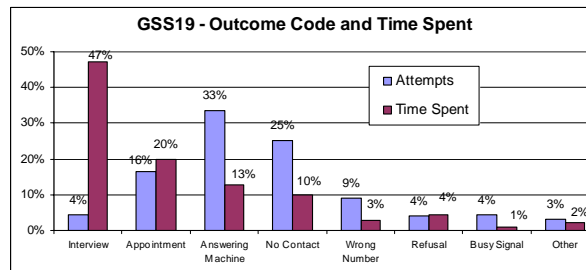

Figure 1

The overall proportion of time spent on depends on the interview length as well as the amount of effort put into reaching both respondents and non-respondents as shown in the previous table. The other items where most of the time is spent are making appointments, answering machines and no contacts. However, the answering machine and no contact outcomes represent 58% of the total number of attempts but only 23% of the total time spent. These types of calls are generally short except when a message is left by the interviewer.

There are other ways to analyze the number of attempts versus the time spent, in particular, by data collection phase. The data collection process can be divided into three steps. The first step is to make an initial contact with the household. Secondly, interviewers need to get the cooperation of the selected respondents to finally conduct an interview as the third and final step.

Figure 2 shows the distribution of attempts and time spent by data collection phase compared to the final status of the sampled units (i.e., response or non-response). About 18% of all the attempts in SLID were made prior to the first contact for responding units but represent only 10% of the total interviewer time spent (most of the attempts prior to the first contact are very short ones). More time (29%) was actually spent after the first contact leading up to the actual interview. A similar situation holds for non-responding cases. In general, interviewers spent most of the interview time and effort after the first contact trying to get cooperation and a completed interview or to confirm a non-response. Historically, the focus has been on the reduction of the number of attempts to get a first contact. But it would be beneficial to pay more attention to the attempts after the first contact to attain more efficiency.

Figure 3 shows the distribution of attempts, system time and response by the total number of attempts made per case for SLID. For example, the cases where more than 21 attempts were made represent 51% of the overall number of attempts, 29% of the time spent and 10% of the total number of respondents (i.e., the last 10% of responses require about one third of the total time spent). This is common in longitudinal surveys or surveys that require extensive tracing.
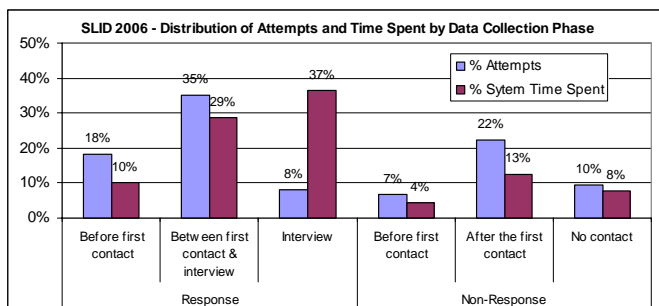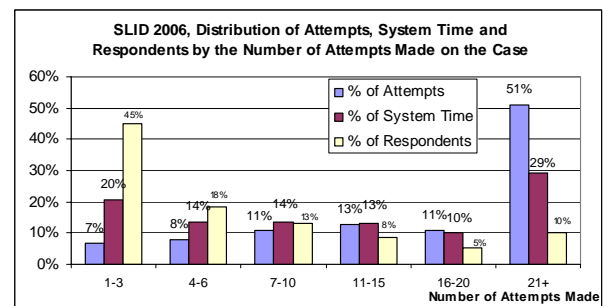

Figure 2

Figure 3

4.3 Contact Rates

For the purpose of this analysis, a contact is defined as any attempt in which an interviewer is able to talk to someone in the household. Figure 4 presents the proportion of attempts and the corresponding contact rates for each of four time slices implemented for TAMS. Making contact with a household is more likely to be successful during the evening than during the

morning or afternoon hours (Figure 4). Surprisingly, even though the contact rate is obviously higher during the evening period, a large proportion of the first and subsequent attempts are still made during the day.
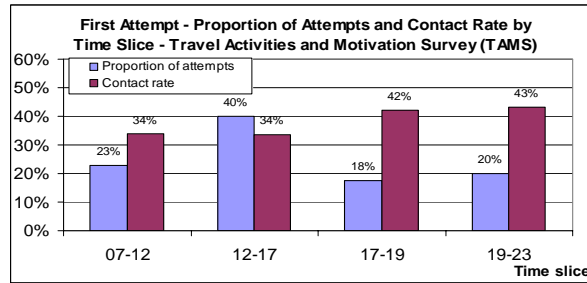

Figure 4

Figure 5 presents the contact rate for each of the four time slices by weekday/weekend for TAMS. Again, making contact is more likely to be successful through the weekend and evening weekdays.

Figure 6 shows the contact rate by day and time of day. The contact rate seems to be better on Monday and Tuesday for the weekdays but decreases slightly on subsequent days. However, it is worth mentioning that Regional Offices generally put more interviewing resources on Monday (which decreases slightly each day until Friday). Also, this daily contact rate pattern varies slightly from survey to survey because it also depends on the starting survey day and the amount of effort put on each day. Finally, due to operational considerations, it should be noted no interviewing is undertaken on Friday in the late evening (19:00-23:00) and Sunday morning (7:00-12:00) in all regions.
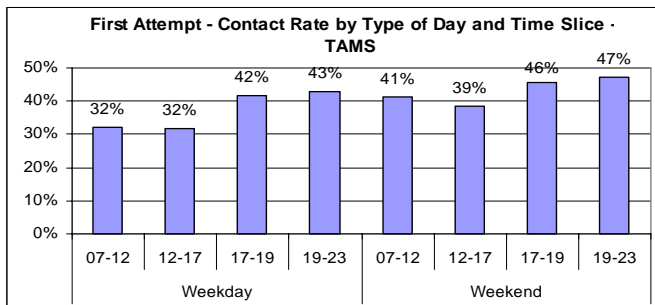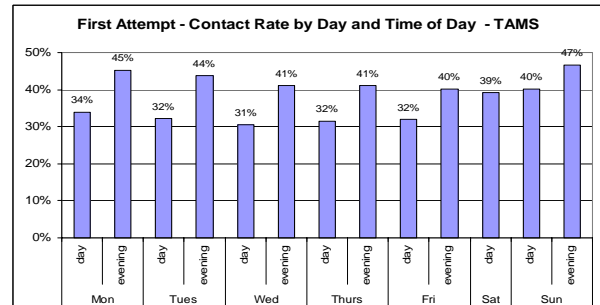

Figure 5

Figure 6

4.3.1 When are individuals at home?
Every five years Statistics Canada conducts the General Social Survey's Time Use Survey which provides a wealth of information about the daily activities of Canadians, in particular for the purpose of data collection, about the specific times they are at home. This survey then provides information to assist in determining the likelihood that an individual will be at home at a specific time prior to the start of data collection. Figure 7 shows the weighted amount of time that Canadians are at home by hour slice. The time spent at home depends on many socio-demographic factors such as the place of residence, the working status as well as the season of the year.
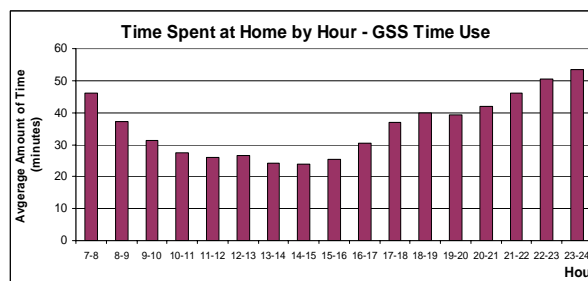

Figure 7

This information is used (when available) at the initial phase of data collection by some surveys (e.g., SLID), in addition to the sample design information, to improve the contact rates for the first few attempts.

4.4 Sequence of attempts

While it is important to make the first attempt at an appropriate time, the majority of calls are subsequent follow-up attempts. For these attempts, the history of attempts or sequence of calls should be taken into account for each sampled unit. Figure 8 shows that the contact rate after the first two calls (i.e., at least one of the first two attempts resulted in a contact) depends on the time slice in which each attempt was made. Not surprisingly, the contact rate is higher when the first two attempts are made early and late in the evening (66%) – regardless the order of the two calls. The contact rate remains relatively high when at least one of the attempts is made in the evening period (17:00-19:00, 19:00-23:00). But when both calls are made during the day, the contact rate decreases by 23 percentage points or more compared to the best scenario when both calls are made in the evening.

In addition, while the sequence of attempts itself is important, another factor to consider is the outcome of the previous attempts. In other words, survey researchers would like to know if the outcome of the previous attempt to a given case is a good predictor of the contact rate of the next call to this case. Table 4 shows the General Social Survey (GSS) contact rate for the second attempt conditional on the result of the previous outcome by the time of day.

The first row of table 4 shows that the contact rate when the second attempt is made during the day is about 24% for the cases that were first called during the day but no contact was made. The results show that when contact was made on the first attempt, the timing of the second call is not as critical as when no contact was made (possibly due to appointments having been made during the first call). It is beneficial for no-contact cases to be re-attempted in the evening.
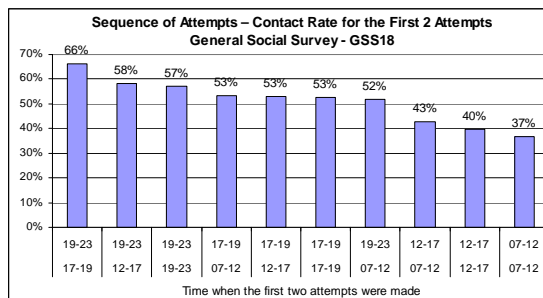

Figure 8

Table 4

General Social Survey - Cycle 18			
First call		**Second call**	**% of**	**Contact**
Time slice	Contact	Time slice	Cases	Rate
Day	No	Day	39%	**24%**
Day	No	Evening	13%	**37%**
Evening	No	Day	13%	22%
Evening	No	Evening	6%	28%

As opposed to the socio-demographic and sample design information that provides general information about the "best time" to call a given household, the sequence of calls provides specific data for individual units about ways to improve the probability of reaching them. This information could be processed, analysed and used during the data collection process to adapt and improve the current data collection strategy. Improvements can be seen when this information could be processed in a timely manner and adjust the collection strategy based upon it.

4.5 First contact versus interview

Generally, it takes less than three attempts to make an initial contact with a household (Table 5). For RDD surveys, the average number of call attempts to make first contact declines slightly with the introduction of time slice for call scheduling.

Table 5

	Survey	Average Number of Attempts Before the First Contact
Before	GSS18	2.7
Time slice	GSS19	3.2
	CTUMS	2.5
After	CTUMS	2.3
Time slice	TAMS	2.3

As discussed previously, the data collection process can be divided into three phases: first contact, cooperation and interview. The next two tables show the relationship between the first contact and the interview. The first table presents the attempts and number of days elapsed between the first contact and the interview. In particular, Lag Call refers to the total number of calls between the first contact and the interview while Lag Date indicates the number of days between the first contact and the interview. Table 6 shows that an interview took place during the first contact attempt for about 47% of respondents (i.e., the Lag Call equals zero). This means that about 47% of respondents are reached on the first contact. On the other hand, about 54% of the respondents are reached on the same day as the first contact was made. For example, the first contact with the household was made in the morning but the interview took place in the evening because the respondent was not available or ready to conduct the interview at the time

of first contact meaning that the Lag Date is 0 even though the Lag Call is greater than 0. Even though about 50% of the respondents are reached on the first contact, a lot of effort in terms of additional attempts and interview time spent is required to reach the second half of respondents for all Statistics Canada surveys (see figure 2).

Table 7 shows the relationship between the first contact and the interview time slice for TAMS. For example, 24% of first contacts took place in the late evening while 28% of the interviews took place during the same time slice period meaning that some of the respondents were not ready or available to conduct the interview in the time slice they were first contacted. Additional investigations could be done on these data to assess the relationship between the first contact and the interview ("best time to call").

Table 6

		SLID 2006	TAMS
Lag Call	0	47%	47%
	1	14%	15%
	2	7%	9%
	3	5%	7%
Lag Date	0	54%	54%
	1	7%	6%
	2	3%	3%
	3	2%	2%
Response		25,127	52,584

Lag Call = ncalls between the first contact and interview

Lag Date = days between the first contact and interview.

Table 7

TAMS First Contact	Interview Time Slice				
	07-12	12-17	17-19	19-23	Total
07-12	**58.5%**	16.0%	10.9%	14.6%	20.6%
12-17	8.9%	**62.1%**	12.2%	16.9%	34.8%
17-19	9.2%	12.9%	**59.4%**	18.5%	21.0%
19-23	10.0%	14.1%	10.9%	**65.0%**	23.6%
Total	**19.4%**	**30.9%**	**21.5%**	**28.1%**	**100%**

4.6 Call scheduler features

This section provides an overview of the impact of some scheduling features such as time slices, cap on calls and browser use. Further analyses will be required to assess the impact of these initiatives on the data collection process.

4.6.1 Cap on calls and time slices

The cap on calls policy aims to limit the number of calls or attempts permitted per case to control respondent burden and to improve the cost benefit of the calls made. Due to the limit on calls, it is important to ensure that all calls are handled in the most effective manner. The time slice feature was introduced to assist in managing the new cap on calls policy and essentially focuses on the first contact. In practice, time slices ensure that a specific number of calls is attempted at different times of the day, and on different days of the week, before a case is finalized. In summary, the introduction of the time slices in social surveys has:

- improved the distribution of attempts during the day (better distribution between day and evening calls);
- reduced the time before a case is called in the evening peak interview period. In the past, cases were called 5 or 6 times during the day with no contact before a first call was made in the evening have been seen;
- reduced the average number of attempts to get a first contact (Table 5).

On the other hand, few surveys have been under the cap on calls initiative since January 2006. Currently, research is ongoing for several surveys to assess the impact of cap on calls on both response rates and quality of survey estimates even though previous internal Statistics Canada research has shown that little gain was achieved after the 20th call. Results to date show that the cap on calls policy could be adapted for some types of surveys, especially longitudinal surveys.

4.6.2 Browser use

With the call scheduler, interviewers have the opportunity to use the browser to access any given case. This means that the interviewer could scroll the list of all cases and manually select a case thus skipping the call scheduler (i.e., the automatic delivery of the next available cases by the system). Initially, the browser was implemented to allow interviewers to access any record primarily in the event a respondent called back. Currently some interviewers use the browser feature quite extensively for various reasons even in the very early stages of the data collection period.

The use of the browser generally increases as the survey progresses and this situation has some impact on the data collection process. Firstly, the call is probably not made at the most appropriate time slice considering the history of the cases. Secondly, there is an impact on the cap on calls. The fact that the cases are likely to be called more often during a given day after a certain point in time (which generally corresponds to a large increase in browser usage) rapidly increases the number of attempts resulting in cases being capped too early in the process and sent back to Head Office.

4.7 Staffing versus in-progress workload

Finally, some investigations were performed to better understand the relationship between the amount of interviewing effort made and the expected in-progress workload through the data collection cycle in particular in the second half of the data collection period.

Figure 9 from the 2006 SLID shows that cases are likely to be called more often during a single day in the second half of the collection period. During the first five days of collection only about 4% of all cases were called three or more times on any one day. However, by the ninth five-day period this number had risen to 25%. This suggests that at that point the interviewer staffing levels are greater than the sample workload. In other words, more interviewing hours are scheduled for this period of time than are practically required. Making multiple futile attempts on a single day can be damaging to surveys with a cap on the number of calls, especially for longitudinal surveys or others that may require tracing. Better tools that would allow the collection centres to more effectively schedule interviewers based on workload would reduce this risk (Cooper et al., 1998).
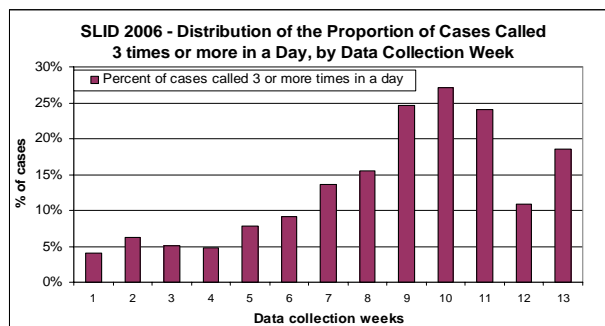

Figure 9

5. Lessons Learned

The objective of this paper was not to produce an exhaustive study of each issue related to the analysis of the data operation files but rather to investigate and identify potential efficiencies in the current data collection process.

Based on the highlights of this research, it is clear that there is a need to develop a more flexible and efficient data collection strategy to make better use of the attempts allowed under the cap on calls policy. The collection approach should evolve through the collection period especially in the second half of the data collection period. In particular, data collection should make better use of the information available prior to the start of collection, but more importantly, future data collection strategies should take advantage of the information that becomes available during data collection (e.g., sequence of calls, audit trail information) to adapt and improve the current data collection strategy for each unit that remains in the sample.

Historically, the focus has been on the reduction of the number of attempts to get a first contact (the time slice initiative is a good example) but this is not where most of the data collection time and effort for both respondents and non-respondents has been spent. It would be more useful to pay more attention after the first contact when interviewers are trying to get cooperation and an interview or to confirm a non-response.

The results also show that a strong and timely active management program needs to be established. In particular, data collection managers need to better monitor the in-progress sample and the call scheduler facilities with regards to the browser use as well as the impact of time slice and cap on call initiatives (Laflamme, Maydan and Miller, 2008).

6. Future Work

This series of investigations has provided a better understanding of the data collection process and it represents an important first step in the initiative to identify strategic opportunities to improve data collection. The next steps could be grouped under four headings:

- Identification of current and new issues that require further investigation;
- Maintenance and improvement of the active management program;
- Assimilating of these results into a common and integrated data collection framework; and
- Better understanding of the relationship between the amount of interviewing effort and the in-progress workload.

The first objective of this report was to investigate and identify potential efficiencies in the current data collection process. Some of the issues identified and discussed require further investigation to determine strategic opportunities that could be used to improve the collection process. For example, a more detailed analysis of the sequence of calls to get a first contact as well as to get respondent cooperation is required. Another example would be the elapsed time between calls to maximize the likelihood of contacting respondents and getting cooperation.

Currently, the active management program (Hunter and Carbonneau, 2005; Laflamme, Maydan and Miller, 2008) is implemented for most of the CATI surveys to   constantly monitor data collection through the entire cycle. However, this program needs to be improved and expanded to other types of surveys. That being said, the implementation of this program for both CAPI and agricultural surveys is planned for the very near future.

The integration of these results into a common data collection methodology framework could take the form of responsive (Groves and Heeringa, 2006) or adaptive collection design (Laflamme and Mohl, 2007). The responsive design could be seen as a multi-phase or phase-in strategy (i.e., collection phases that evolved as collection progressed). Each phase representing a period of time during the collection period in which the same data collection strategy is used. However, some research will have to be conducted to operationalize this new adaptive data collection strategy taking into considerations the scheduling realities for CATI and CAPI interviewers.

Finally, it is clear that some resources will have to be devoted to better understand the relationship between staff levels required and expected workloads at different points in time during the data collection process in order to more effectively schedule interviewers' time and to reduce the risk of reaching the cap on calls too quickly.

Acknowledgements

References

Couper, M. P., Baker, R.P., Bethlehem, J., Clark, C.Z.F., Martin, J. Nicholls W.L. and O'Reilly, J.M. (1998), "Computer Assisted Survey Information Collection", Wiley series in survey methodology section, Chapter 15 by Edwards, Suresh and Weeks, p. 301-306.

Groves, R.M. and Heeringa, S.G. (2006), "Responsive design for household surveys: Tools for actively controlling survey errors and costs". Journal of the Royal Statistical Society Series A. Volume 169, Part 3.

Hunter, L. and Carbonneau, J.-F. (2005), "An Active Management Approach to Survey Collection". Proceedings from the 2005 Statistics Canada International Symposium on Methodological Issues.

Laflamme, F. and Mohl, C. (2006), "Research into Survey Data Collection using Paradata". Advisory Committee on Methods and Standards, internal report, Statistics Canada.

Laflamme, F. and Mohl , C. (2007), "Research and Responsive Design Options for Survey Data Collection at Statistics Canada". 2007 American Statistical Association, Proceedings of the Section on Survey Research Methods.

Laflamme, F., Maydan, M. and Miller, A. (2008), "Using Paradata to Actively Manage Data Collection". 2008 American Statistical Association, Proceedings of the Section on Survey Research Methods.

Massey, J.T, Wolter, C., Wan S.C. and Liu K. (1996), "Optimum Calling Patterns for Random Digit Dialled Telephone surveys". 1996 American Statistical Association, Proceedings of the Section on Survey Research Methods.

Reedman, L. and Robinson, M. (1997), "An Improved Call-Scheduling Method Using Call History and Frame Information". 1997 American Statistical Association, Proceedings of the Section on Survey Research Methods.

Scheuren, F. (2005). "Paradata from Concept to Completion". Proceedings from the 2005 Statistics Canada International Symposium on Methodological Issues.

Stec, J.A., Lavrakas, P.J. and Shuttles, C.W. (2004). "Gain Efficiencies in Scheduling Callbacks in Large RDD National Survey". 2004 American Statistical Association, Proceedings of the Section on Survey Research Methods.