

Research and Responsive Design Options for Survey Data Collection at Statistics Canada

Chris Mohl¹, François Laflamme²

¹Statistics Canada, 17-M RH Coats Building, 100 Tunney's Pasture, Ottawa, Ontario, K1A 0T6, Canada

²Statistics Canada, 6-C2 Jean Talon Building, 170 Tunney's Pasture, Ottawa, Ontario, K1A 0T6, Canada

Abstract

In an effort to continue to collect timely, high quality, but cost-effective survey data, Statistics Canada is reviewing its current data collection practices. The use of paradata is seen as a key component of this analysis. This information could be used to implement a responsive design which would allow the collection strategy to evolve over time. The paradata can be used either on its own or in conjunction with historical information to provide clues on when initial or subsequent contact attempts should be made, assign differing collection priorities to units or sub-sample non-respondents late in the collection cycle. The options being studied as part of this research project are discussed and some initial findings are presented. Possible responsive design steps, their benefits and the information needed to implement them are also discussed.

KEY WORDS: Paradata, Responsive Design, Collection Procedures

1. Introduction

Like many other statistical organizations, Statistics Canada faces increasing challenges in maintaining high-quality data collection within a reasonable budget. Factors such as the unionization of the interviewing staff have increased the costs of collection. Likewise, increasing concerns about privacy, technological advances such as caller ID and increasing burden placed on the respondents from all survey-taking organizations has made the task of getting responses from both households and businesses more challenging.

For these reasons, Statistics Canada has been reviewing its data collection processes with the goal of keeping the survey response rates and quality at a high level, collecting the data more efficiently and keeping the burden on the respondent to a minimum. Information gathered from previous and current collection cycles, in the form of paradata, is the principal data source being used in these analyses and would also play a key role in any subsequent changes to the collection process. The goal is to implement a leaner collection procedure which makes more use of the information that has been gathered both prior to, and during the collection period in order to contact people and businesses more efficiently. A responsive design will provide operations managers with

the information that they need to make more efficient use of their resources. It will also allow the data collection procedure to evolve over time, permitting resources to be allocated to projects that have more need for them and increasing the probability of contacting and interviewing the sampling units with the highest importance.

The focus of this paper is on operational issues related to attaining a response in an efficient manner. While some of the proposals have no impact on the statistical phases such as weighting and estimation, others do. The mathematical impact of modifying collection procedures are mentioned only in passing in this paper, but will have to be thoroughly investigated before such changes are implemented.

In section 2 of this paper we will describe some of the decisions that were made and the changes that were implemented as a result of the initial phase of this analysis. Section 3 will describe some subsequent studies that show the potential areas where the paradata information can be used to further understand and advance the idea of data collection efficiency. Based on this work, Statistics Canada is in the process of studying and putting into place a responsive design approach to data collection. The first component to this approach, referred to as active management, will give the collection and survey managers more information on the progress of collection at any point in time. This is described in section 4. The fifth section describes the second component, called adaptive collection, and discusses different options for such an approach in the future. Finally, some concluding remarks are provided in section 6.

2. Initial Changes to the Current Collection Environment

Recently there have been a few changes proposed and introduced to the Statistics Canada field collection procedures. These decisions were made prior to the current research work, but they have had an impact on the way that collection is done.

The studies leading to these changes relied heavily on the use of paradata. Sheuren (2005) defines paradata as follows "Paradata, in short, are part of the computerized metadata that surround many large-scale government and other surveys". In the case of Statistics Canada surveys, the principal source of these paradata comes from what

are known as call history files. These files contain one data record for each time that a case is opened in the computer collection application, either for data collection or other purposes. 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 if applicable and much more. This information was critical to much of the analysis described in this paper. It is available on a daily basis and thus can be immediately used as part of any decision-making process. In the case of these studies, historical paradata were used to examine and draw conclusions about modifications to the current collection procedures.

2.1 Cap on Calls

Many surveys, especially those on social issues, are conducted using Computer Assisted Telephone Interviewing (CATI). An application built using the Blaise software assigns individual cases in an ordered manner to interviewers working out of a centralized environment. The application is capable of keeping track of the number and time of calls that have been made to an individual household or business.

After performing an analysis on the call patterns for several surveys (both longitudinal and cross-sectional), it was discovered that some cases, or survey units, were being called an extraordinary number of times. This led to a recent initiative to put a cap, or maximum number, on calls for each case. Analysis was conducted on historical collection from different surveys to determine the impact that different cap values would have, by assuming that those cases which took more attempts to complete than allowed by the cap would be non-respondents and would therefore need to be accounted-for in the non-response weight adjustments. Based on this investigation, a maximum of twenty calls was set for Random Digit Dialling (RDD) surveys and twenty-five for longitudinal and targeted respondent surveys. These restrictions have now been implemented but their impact is still being investigated. In the case of longitudinal surveys, there are some initial concerns that the cap may be too low and that it could have a negative effect on the quality of the estimates.

2.2 Time Slices

Making first contact with an individual, especially for social surveys, can be a time-consuming task in the data collection process. Since individuals are only available at certain times of the day, it is important to try to contact them at different periods of the day if the initial contacts are not successful. For this reason, the idea of time slices was implemented into the Blaise application. The day was divided into four slices: 7am-12pm, 12pm-4pm, 4pm-7pm

and 7pm-11pm. The application ensures that the calls are scheduled for different parts of the day to increase the probability of calling the respondents at a time when they are home. An efficient process for making first contact is increasingly important with the introduction of the cap on calls.

3. Highlights of Data Collection Research

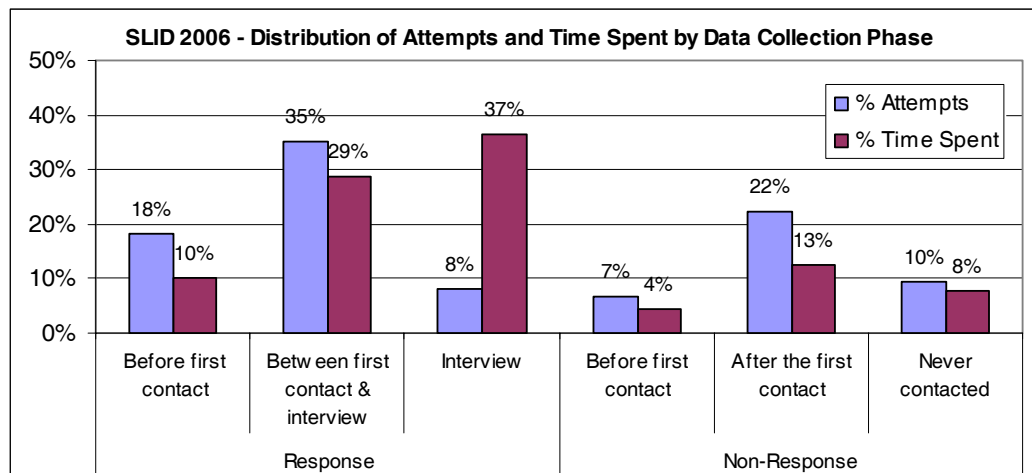
The research described above and other data collection research was often conducted by individual survey statisticians who were responsible for a single survey, or a small number of surveys. Statistics Canada's management has recently promoted the idea of more centralized and concentrated mathematical analysis of data collection procedures in order to study them and draw conclusions that are relevant to a wider range of surveys. These conclusions and the subsequent recommendations can then be applied to groups of surveys. The groupings could be based on the mode of collection, the targeted population (individuals vs. businesses) or the nature of the survey (cross-sectional vs. longitudinal).

As a first step in this process, a survey statistician was assigned to Statistics Canada's Survey Operations Division with the express role of conducting this sort of consistent analysis. In this section, we highlight some of the findings of this work which in turn have led to some of the proposals and recommendations described in the following sections. Note that much of this initial work focused on CATI social surveys. Many of the resulting conclusions may be extrapolated to other types of surveys or collection modes, but some may not. These verifications will be done in the future phases of the project.

The principal goal of this phase of the research was to shed more light on the process used to get respondents to answer a Statistics Canada survey. This information can then be used to identify ways to get high quality responses while at the same time keeping the number of attempts (when the interviewer tries to contact the respondent) and contacts (when the interviewer actually speaks to someone) to a minimum.

3.1 Results

The total number of attempts required to complete a case, resulting in either a response, non-response or out-of-scope designation can be extremely high in many cases. In a typical CATI survey, it takes an average of about ten attempts before a case can be finalized. However there are also a significant number of cases where the number of attempts is much higher. It is not unusual to see more than Figure 1



twenty-five attempts required before a response is finally attained (and extreme cases taking fifty or more attempts). Although the newly implemented cap on calls will eliminate these outlier cases in the future, it is clear that, even with a cap, the potential for savings is still significant.

In order to better understand the data collection process, it is necessary to know where most of the data collection effort is spent. The process can be divided into three distinct parts. The first part is to make an initial contact with the household. The second is to get the cooperation of the selected respondent, while the third is to conduct an interview with the appropriate individual. As shown in Figure 1, representing Statistics Canada's 2006 Survey of Labour and Income Dynamics, most of the interviewers' time is spent after the first contact in trying to get the respondent's cooperation to complete an interview. Historically, the focus of research has been on the reduction of the number of attempts to get a first contact. However it would be beneficial in future research to conduct a more thorough investigation on the data collection process after the first contact if one wants to make a bigger impact on the costs.

As many other studies have shown, weekday evenings and weekends have been systematically found to have the best contact rates for CATI social surveys while early morning, lunch time and early evening are more appropriate for agriculture surveys. These conclusions were reinforced by analysis of the 1998 Statistics Canada Time Use Survey (General Social Survey cycle 12) which collected information about the specific times Canadians are at home. It provided 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, especially when some socio-economic information such as the age, the place of residence, the working status etc. of the individual is known ahead of time. This information could be used at the

initial phase of data collection to improve the contact rates during the first few attempts.

On average, it takes less than three attempts to make an initial contact with a household. The introduction of time slices, which target a specific number of calls at different parts of the day and different days of the week, has shown some encouraging signs of slightly reducing this average. While three attempts may seem reasonable, it is important to note that between 40%-50% of the respondents are reached on the first contact and thus there are many others which require a much higher number of attempts.

While it is important to make the first attempt at an appropriate time, the majority of calls are subsequent follow-up attempts. The results of the analyses 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 an appointment having been scheduled during the first call). However, as shown in Table 1, using data from Statistics Canada's 2004 General Social Survey cycle 18, it is beneficial for no-contact cases to be re-attempted in the evening, even if the first attempt took place during this period and was unsuccessful. Unlike the socio-economic and sample design information that provides general information from which one can determine a "best time" to call a given household, the sequence of calls can provide specific unit-level data to improve the probability of success. This information could be processed, analysed and used during data collection to adapt and improve the current data collection strategy.

Table 1

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

Research was also conducted to better understand the relationship between the number of interviewers assigned to a survey and the amount of interviewing effort required throughout the data collection cycle. In general, cases were likely to be called multiple times (defined as three or more attempts) during a single day in the second half of the data collection period. This suggests that, at that point, the interviewer staffing was larger than the sample workload necessitated. The introduction of the cap on calls could be negatively impacted by this phenomenon if calls are “wasted” by re-attempting to call a respondent only hours after it was previously tried.

This research, along with many others not described here provided a newfound understanding of many aspects of the data collection process. For more information about the results of this research, see Laflamme (2007). It is obvious that there is a need to develop a more flexible and efficient data collection strategy, not only to reduce data collection costs, but also to make better use of the attempts allowed under the new cap on calls policy. The collection approach should evolve throughout the collection period, especially in the second half. This includes making better use of any available socio-economic information prior to the start of data collection, but more importantly, taking advantage of the information that becomes available during data collection such as the sequence of calls to quickly and automatically adapt and improve the collection strategy for each unit in the sample that has yet to respond. The results also show the need to establish a strong and timely active management program. In particular, there is a need to better monitor the in-progress sample and the call scheduler facilities as well as the impact of time slice and cap on call initiatives (Laflamme, Maydan and Miller 2007).

The concepts of a responsive design, achieved by an active management program and an adaptive collection approach are discussed in the next sections.

4. Active Management

Data collection is a key component of the survey process, not only because it often demands a very significant portion of the entire survey budget, but also due to its direct impact on the quality of Statistics Canada’s information. The data collection manager’s challenge is to make the appropriate trade-offs between accuracy, timeliness and cost in order to

meet clients’ needs. At the same time, it is important that information on collection’s progress be made available to all affected parties. Complicating the situation are the varying needs and degree of detail in the information provided to differing levels of management. Higher management does not need data at a detailed level such as that of the interviewer, whereas operations managers do require it. In addition, the data needs often depend on the point of time in the data collection cycle. With the increasing complexity of survey data collection procedures, survey managers need better tools to assess the survey data collection process.

Statistics Canada’s active management program is a set of processes and tools used to manage survey data collection while it is in progress. Active management has two objectives. The first is to identify problems as early as possible and to correct them before collection has finished. The second is to use collection resources effectively to make the most appropriate trade-offs between data quality and survey costs. In other words, active management aims to provide timely, topical and relevant data on survey performance to identify and correct problems based on indicators. Active management activities include monitoring survey progress, performing timely analysis, identifying problems, implementing and communicating corrective action, and evaluating success.

Among these activities, survey monitoring was identified as a key element of the active management initiative to provide timely assessments of the survey progress against the expectations during the data collection period. Three levels of survey monitoring reports have been developed for this use: standard, customized and ad hoc reports.

Standard reports are common to all surveys. These reports are available to head and regional office data collection managers directly through the Statistics Canada paradata warehouse in a timely manner to monitor an agreed-upon set of indicators. For example, the Survey Performance Report brings together production and financial data to monitor collection indicators such as time per unit, cost per unit, completion rate and response rate at various levels of aggregation. The Survey Progress Report assesses collection progress in terms of cases resolved against the prescribed completion targets or response rates. The customized reports provide information for monitoring specific issues and a predefined set of indicators not available in standard reports. In fact, the customized reports often use sample design information to define sub-populations and enhance the analytical relevance of the reports by producing reports at a more detailed level of aggregation not available via the standard reports. Because they use paradata which are available less than twenty-four hours after the interviewer action has taken place, the standard and customized reports are then complementary in providing relevant and timely information at different levels of detail. Finally, the ad hoc

reports are produced when situations arise that are not covered by standard or customized reports. The standard and customized reports do not aim to provide all the answers to all possible issues, but they are built in such a way that they could highlight key elements that might explain the source of the problem and determine if more in-depth investigations are required.

The concept of active management is very similar to that of quality assurance since it attempts to move quality upstream by anticipating problems before they occur and aims to ensure quality via the use of prevention and monitoring strategies.

5. Adaptive Data Collection

In the past, the data collection procedures for a typical Statistics Canada survey cycle have been rather static. A plan was developed ahead of time, which specified the number of interviewers and the budget required for a given survey, the shifts that interviewers would work etc. In general, these procedures would not be significantly modified unless there were unexpected circumstances. As collection progressed, there would be continued attempts to recontact non-respondents in hopes of getting a response.

The data collection process should be more flexible and able to respond and adapt based on not only the success of the data collection for an individual survey, but for all surveys being conducted from the same centralized work area. This would imply that interviewers' shifts would be better scheduled to reflect the time periods in which more (or less) staff was required. Also, cases with a higher probability of successful contact or belonging to a population which has been under-represented among the respondents could be given higher priority in the centralized case scheduler. The idea of a responsive or adaptive approach to data collection has been raised in the past (Groves and Heeringa, 2006). Statistics Canada is examining the different ways in which it might be able to implement such a system (Laflamme and Mohl, 2006). Each potential method uses information from one or more of the paradata sources in an attempt to improve the efficiency of data collection. An adaptive collection approach could be potentially used at four different stages of the data collection cycle – prior to the initial contact attempts, following the first few attempts, midway through the collection cycle, and near the end when only hard to reach units are left in the sample.

5.1 Phase 1: Scheduling of Initial Contact Attempts

This process would take place before the collection period starts and attempts to determine a time of day or week when the probability of contacting the individual or

business would be favourable. In cases of panel or longitudinal surveys, information may be available from previous responses (the time of day they responded last time etc.). Even first-time respondents could be better targeted when information about them is available from the frame. In the case of an RDD survey, a frame of telephone numbers often provides a “household/business” indicator. The initial call attempt for numbers thought to be out-of-scope can be scheduled for the periods when people are less likely to be home, leaving the prime interviewing time for numbers which are thought to be households.

5.2 Phase 2: Adapting the Scheduling of Calls

Following the initial call attempts, information from the call history file should be used in conjunction with the findings from studies on previous data collections to establish a new time to call. If a person has been called three times in the morning because he/she had the characteristics of a person who has a high probability of being available during that time period, but no contact has been made, then there needs to be something in the system to move the call attempts to other times of the day. This is a similar idea to that of time slices, but in a more automated manner. Before such an approach is put into place, additional research needs to take place in order to establish the appropriate reactions to observed call patterns. It will also have to take into account the interviewing resources available.

5.3. Phase 3: Assigning Different Priorities to Units

Once collection has been ongoing for a while and a number of attempts have been made using the methods described in the first two phases, some thought should be given to prioritizing some cases over the others. For several years, business surveys at Statistics Canada have used a score function to give priority to certain cases during data collection (Pursey 2003). This function is based on several factors including the size of the unit and the success in collecting data from other units within its sub-population (e.g. industry) to that point. The business survey score function is currently being re-examined to also take into account paradata from previous collection periods (since there is often a high sample overlap from one period to the next) as well as the contact attempts from the current collection period. This will be used to identify the sub-populations that are most in need of respondents as well as those units which are most likely to actually respond. Some of these ideas could also be transferred over to the collection of social survey data.

In situations where characteristics of the sample units are known prior to collection, either from data collected in previous cycles or from the frame, groupings of important

characteristics can be created and monitored to determine the success in getting responses from these sub-populations. If a certain group or groups start to show poorer response than others, then the remaining individual cases in these groups could be given higher priority, possibly by ensuring that they are attempted more frequently or at times which have been shown to be more successful in getting responses. Decisions about which groups are considered to be under-performing and which units should get higher priority would have to be done in an automated fashion using statistical methodologies so that these decisions could be made on a timely and objective basis.

In other cases, this information on the sample units' characteristics will not be available. However, once again paradata from previous contact attempts could be useful. This information could be used to flag which records are more likely to lead to a completed interview based on observations up to that point. For example, a case which has already partially completed the questionnaire is more likely to be fully completed in future contacts than one which previously had a refusal. One possibility is the use of logistic regression (where the dependent variable indicates if a response has been received or not) to develop a prediction model to determine the probability of future success. See Groves and Heeringa for an example of this type of work.

5.4 Phase 4: Sub-sampling of Non-respondents

The analysis discussed previously has shown that for some non-responding cases, a high number of attempts are made. These interviewing resources might be better served elsewhere (on another survey for instance). It may be fruitful to consider the possibility of sub-sampling non-respondents in order to reduce the amount of time and resources dedicated to cases which have a low probability of responding. This approach has been successfully implemented in other organizations such as the United States Census Bureau's American Community Survey (Tersine and Starsinic, 2003). Once a certain point in the collection has been reached, the remaining non-respondents would be identified and possibly sub-sampled. The focus of the collection would then be only on these sub-sampled units.

The point in time at which this sub-sampling would take place and the extent of it could be determined in several different manners. There could be a specific date in the collection at which time the sub-sampling would be scheduled to take place. The decision on how much to sub-sample (and potentially which records should not be subject to any sub-sampling) would be identified through the use of relevant paradata and by a pre-defined algorithm so as not to delay the data collection process.

Groves and Heeringa also suggest that decisions to sub-sample (or even stop collection completely) can be determined when the estimates for one or more key questions have stabilized. It could be feasible to divide the population into groups – some of which may reach stability for key estimates before others. In the more stable groups, a higher sub-sampling rate could be applied.

However, in order to apply any sub-sampling method, there are additional methodological steps and risks that need to be considered. The weighting and variance estimation under such a multi-phase sampling approach is more complex and may require extra study. In addition, the sub-sampling needs to be done to keep the additional variability due to sub-sampling to an acceptable level. If an approach based on the stability of estimates is considered, one needs to be certain that the stable estimates are also unbiased. Finally, this approach may not be applicable to some specific types of surveys such as longitudinal surveys.

5.5 Other Adaptive Collection Issues

In addition to the processes that could be implemented at the four phases of data collection within a survey described above, there are other ways in which an adaptive collection strategy could be used. These investigations focus more on the overall Statistics Canada collection program and its management rather than an individual survey.

In order to respond to the increasing diversity of ways that Statistics Canada respondents wish to respond to the questionnaire, multiple mode collection is becoming increasingly important. Some respondents to business surveys where the questionnaire was traditionally mailed out and back now want to be able to give their data over the phone or via an electronic data collection instrument. Likewise, household survey respondents, especially those of a younger age, want to fill out the survey at a time that suits them, via a web application rather than over the phone. This preference often comes to light only when initial contact is made. Therefore, there needs to be a better mechanism to quickly switch the collection from one mode to another. This additional flexibility is not easily done in the current environment and its feasibility needs to be investigated.

Collection managers also need to be able to more accurately predict the required number of interviewing staff on a project and when to implement them, both in terms of the time of day and the stage of the collection process. With increasingly limited resources available, there are few excessive resources that can be wasted by having too many people scheduled to work on a particular survey. Tools are needed to analyse the paradata that are available and provide

projections on when and how many staff members are required for a certain project or how to optimize the work across surveys among the available staff members.

5.6 Comments on Adaptive Collection

Statistics Canada's senior management supports the research in data collection as it plays a key role in improving the efficiency of the data collection process. An adaptive collection approach would require a significant amount of both up-front and ongoing resources to reap the benefits. This includes determining how to assess the tradeoff between any savings that may be achieved and the potential impact on the quality of the results. In addition, tools to perform the analysis will have to be built in order to keep the human intervention in the decision-making process to a minimum. Collection systems will also have to be modified and made more flexible to permit changes to the collection strategy at any point in time. This includes the abilities to schedule certain units to be called at specific points in time or to give some units more priority than others as collection evolves.

The human aspect also needs to be accounted for. Interviewers at Statistics Canada have their schedules set in advance. Some of the possible efficiencies noted in this section result in changes being made to the number of cases and the timing of their collection during the middle of the collection period. This alters the number of interviewing resources required. Any analytical tools or processes developed to better plan interviewer workloads and schedules need to take this into account.

6. Summary

Initial research has shown that there is a lot of potential for making the survey data collection process more efficient. Up to now, paradata have played a leading role in this analysis and also form the backbone for many of the changes and improvements to data collection procedures that have been, or could be, put into place. There are two principle areas that show the most potential benefit from this work – the scheduling of interviewing resources and improving the probability of successful contact with the targeted person, household or business.

No matter what modifications are made, it is important that they follow some sort of common data collection framework and that individual projects and surveys are handled in a similar manner. This includes the ideas of both active management and adaptive collection of survey data. The main idea is to constantly assess the progression of the collection and adjust the data collection strategy based on the most recent information in order to make the most efficient use of the available resources. Unlike many

initiatives that have been developed on an individual survey basis, this proposal extends across surveys and issues that the regional collection centres are faced with. The effectiveness of this framework depends on the collective effect of many interdependent and often complex measures.

Acknowledgements

The authors wish to thank Yves Béland, John Kovar, Claude Julien and Caroline Pelletier for their helpful comments on the contents of this paper as well as Gayle Keeley for her editorial review.

References

- 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, pp. 439-457.
- Laflamme, F. and Mohl, C. (2006), "Research into Survey Data Collection using Paradata", Internal report to the Advisory Committee on Statistical Methods, Statistics Canada.
- Laflamme, F. (2007), "Understanding Survey Data Collection Through the Analysis of Paradata at Statistics Canada", internal report, Statistics Canada.
- Laflamme, F., Maydan, M. and Miller, A. (2007), "Active Survey Management using Paradata in Statistics Canada", internal report, Statistics Canada.
- Pursey, S. (2003). "Use of the Score Function to Optimize Data Collection Resources in the Unified Enterprise Survey", *Proceedings of the 2003 Annual Meeting of the Statistical Society of Canada, Survey Methods Section*.
- Scheuren, F. (2005). "Paradata from Concept to Completion", *Proceedings of Statistics Canada Symposium 2005*, Statistics Canada.
- Tersine, A. and Starsinic, M. (2003). "Optimum Non-Response Subsampling Rate for the American Community Survey", *2003 Proceedings of the Joint Statistical Meetings, Washington D.C., American Statistical Association*.