**An Overview of Web Paradata to Understand and Monitor Data Collection Process**

François Laflamme, Statistics Canada, francois.laflamme@statcan.gc.ca

**Abstract**

One of the major challenges for any data collection organization is to collect timely and cost-effective data while maintaining a high level of quality. Over the last few years, respondents have become more difficult to reach and less inclined to participate in surveys, which led to additional pressures on collection costs and on quality. At the same time, a broad range of new communication technologies have been adopted by large segments of the population and these technologies, often used as the main means of communication, continue to evolve rapidly. In response to these challenges, Statistics Canada has been developing a generic Electronic Questionnaire (EQ) platform to allow mixed modes of data collection strategies for both social and business surveys. To learn more about the various issues associated with the Web data collection in the mixed mode context, a research project was initiated in order to conduct a series of analyses using Web paradata.

1. Introduction

The Web is used more and more as a data collection mode by statistical organizations in order to meet respondent expectations and preferences but also in an attempt to reduce collection costs while maintaining high-quality data. Statistics Canada has been providing various Electronic Data Reporting (EDR) options for many years for business surveys (Mayda, 2002; Sunter 2009). In general, business surveys use a combination of collection methods including mail, telephone, EDR and more recently, the Web. The 2006 and 2011 Canadian censuses successfully introduced an internet response option, in addition to the traditional paper questionnaire. As well, governmental surveys conducted by Statistics Canada that targeted Federal employees have also used the Web as the principle collection mode.

Given the constant need to improve efficiency and the success of these mixed mode experiences, Statistics Canada is working towards incorporating the Web option as the primary mode of collection for more than 130 social and business surveys over the next five years, including 2016 Census collection. A generic Electronic Questionnaire (EQ) platform has been developed to allow Web and mixed data collection modes for the vast majority of surveys. In order to better plan, understand, assess and monitor the electronic and mixed mode data strategies and practices, a research plan was developed to conduct a series of studies using paradata recorded through Web self-administered data collection. As with the computer-assisted data collection methods, the development and implementation of Web data collection applications are likely to provide a wide scope of paradata. In that perspective, paradata in the survey research context are very often used to understand, evaluate, monitor and finally improve survey design, process and procedures, not only data collection strategies and processes. Two main paradata sources can be defined in the Web self-administered collection realm (Heerwegh, 2002):

* Server-side paradata provides information about how respondents access the Web application i.e. the interaction between the data collection organization, the respondents and the portal/server. It also includes any activities taken for survey management purposes.
* Client-side paradata provides information about how respondents navigate in the EQ application i.e. the interaction between respondents and the questionnaire application itself. The audit trail (i.e. key strokes) and eye tracking information are good examples of client-side paradata.

Much of the research and the initial investigations discussed in this paper rely on the server-side paradata. The paper begins with an introduction of the research objectives. The next section describes the Web paradata that is (or can be) recorded. The following section assesses the quality and the consistency of available Web paradata including potential analytical limitations. The last section presents the highlights of the initial investigations and discusses the new challenges faced by statistical organizations in the Web and mixed modes data collection context, especially with regards to the evaluation and survey monitoring of the non-response.

1. Research objectives

To learn more about the various issues associated with the Web data collection in the mixed mode context, a paradata research project was initiated to pursue five main objectives:

1. Understand Web and multi mode data collection processes
   1. Assess the quality and limitations of Web paradata
   2. Determine how well the Web collection process is described by the paradata
   3. Identify important paradata information gaps
2. Develop measures to monitor the overall data collection process (including dashboard of key indicators)
3. Determine metrics for planning and costing purposes
4. Evaluate data collection approaches to identify the optimal Web and/or mixed mode strategy
5. Identify requirements for relevant information on how respondents interact with the EQ application.

The preliminary investigations, focussing essentially on the three main research objectives, are based on the server-side paradata. In the long run, the survey data and available client-side paradata that provides information about how respondents interact with the Web application will be integrated into the research.

1. Web paradata description and contents

The two main sources of Web paradata, server-side and client-side paradata, are described below.

**3.1 Server-side paradata: Transaction file**

The transaction file is an “activity” file that describes the interaction between three parties: the data collection organization, the respondent and the portal/server (system). A record is automatically recorded each time an action is taken by any of the three parties or for survey management purposes. The activities are grouped in four categories: 1) activity initiated by the data collection organization (outbound activity: e.g. invitation, reminder(s)), 2) activity initiated by the respondent (inbound activity: e.g. login, save, submit), 3) system activity (response to inbound or outbound activity: e.g. login and submission errors, bounce back email)) and 4) survey management activity (e.g. update respondent information, disable or reactivate case).

Each transaction record contains information about each activity which includes: survey identification, sample unit identification, the category of activity (4), the activity[[1]](#footnote-1), the start date and time of the activity plus additional relevant information about each activity (e.g. questionnaire version and language used). Compared to Computer-Assisted Telephone Interview (CATI) transactions, the contents and the structure of the transaction (activity) file for the Web is different and less detailed. However, contrary to CATI paradata, the contents and the structure of the transaction file across different types of Web surveys (i.e. social and business) remain the same.

**3.2 Client-side paradata: Audit trail**

The audit trail describes the interaction between the respondent and the Web application itself. In practice, a record can be automatically recorded each time a key stroke is hit by the respondent (e.g. answer to a question, escape, back up)[[2]](#footnote-2). This type of paradata can only be recorded for some type of inbound activities i.e. when the respondent directly interacts with the application. A generic approach to record this type of paradata is currently under development at Statistics Canada. However, it is worth mentioning that raw survey data and/or survey data resulting from the editing process can in the meantime be used as “proxy” of an audit trail for decision making during collection (e.g. identification of blank questionnaires or questionnaires with a high level of item non-response or edit failures) and for some types of analysis (e.g. non-response rate to specific question).

**3.3 Investigated Web paradata**

Web paradata is currently available for business surveys and some social surveys that used the first version or one of the two updated versions of the Web database developed at Statistics Canada. With each updated version, many improvements have been made to the quality of paradata. The current analysis focuses on the third and most recent version of the database using the General Social Survey EQ pilot (GSS26EQ) - Cycle 26 (Survey on Caregiving) conducted in 2012 from May 8th to June 17th. The GSSEQ Pilot is a Random Digit Dialing (RDD) cross-sectional survey (n~10,000) which randomly selects households from a list of telephone numbers. Once households are contacted over the phone, the entry part of the survey is conducted (including roster and household demographic composition) and one member aged 15 or older is selected and offered to complete an EQ. An e-mail address is collected for respondents agreeing to participate. The EQ is the only data collection mode available for the pilot. Non-response follow-up consists only of e-mail reminders to participants who did not complete the EQ within a few days. The main objectives of the pilot test are: to assess the EQ collection strategy (including EQ acceptance and EQ submission rates); to compare the EQ key estimates with the CATI estimates (from the GSS Cycle 26 Main survey conducted during the same time period) and to study mode effect between the EQ and CATI collection methods (if possible).

**4. Quality of Web paradata**

Before conducting further investigations, it is important to assess the quality of Web paradata and its potential limitations for reporting, monitoring and analysis purposes. In particular, it is necessary to evaluate the impact of these limitations on the scope of potential research including the resulting conclusions.

**4.1 Quality Assurance Framework**

The evaluation of the quality of Web paradata can be performed according to the six dimensions of the Quality Assurance Framework (Statistics Canada, 2002). ***Relevance*** reflects the degree to which paradata meets the research and analytical objectives. Except for the fifth research objectives, it is fair to assume that Web paradata meets most of researcher needs at least in the short run. ***Accuracy*** is the degree to which paradata correctly describes the phenomena it was designed to measure. Since Web paradata is automatically recorded, the accuracy is expected to be very good in general. ***Timeliness*** refers to the delay between the reference point to which paradata pertains, and the date on which paradata becomes available. Web paradata becomes available after each action or, in the worst case scenario, the day after paradata information is recorded. This is suitable to meet the vast majority of the research objectives. **Accessibility** refers to the ease with which paradata can be obtained. Since Statistics Canada collects its own data, accessibility is not an issue. ***Interpretability*** reflects the degree to which paradata can be easily understood and used by analysts or researchers. For example, is it possible to follow and understand the collection process that took place by analyzing the sequence of activities on each case? ***Coherence*** reflects the degree to which several data sources can be successfully and correctly brought together. In the case of surveys that used multi contact and multi-collection modes, is it possible to integrate transactions (e.g. CATI) and activities (Web) to describe and understand the sequence of all actions made on each case from the beginning to the end of the collection period? The following discussion mainly focuses on the ***accuracy, interpretability and coherence.***

**4.2 Impacting factors**

Factors that affect the quality of the Web paradata can be grouped into 3 categories: technical, design and external.

*Technical factors*

Many technical factors can impact the quality of the paradata recorded. The type of browser used by the respondent, the capacity and speed of the respondent’s hardware and connection, and problems with the respondent’s computer or with the organization’s portal (e.g. freeze session, system down) are probably among the most important factors. However depending on the evolution of the technology, some of these problems can be resolved while new emerging technical problems might also appear.

*Design factors*

Design factors refer essentially to the concepts, definitions and scope of the Web paradata. For example, due to some current design constraints, client-side paradata (Audit trail) is not recorded for the time being which creates a gap in information in particular for analysis and research purposes. The list of activities recorded (e.g. login, save, etc.) is also an important example of factors that can directly impact our ability to make sense of the sequence of actions from the beginning to the end of collection. Design factor also includes the way Web activities and CATI transactions are integrated if multi-mode collection or contact modes are used throughout the collection period. This might impact on the quality of paradata (e.g. processing errors or coherence of the chronological sequence of transactions and activities).

*External factors*

Unlike the CATI environment, the Web data collection is a self-administered collection method (i.e. no interaction with an interviewer during the interview itself[[3]](#footnote-3)). In fact, data collection organizations have limited control over how respondents interact with either the server/portal or on how the Web application is used (e.g. individuals can complete the EQ on behalf of the selected respondent (EQ proxy)). Given that some type of activities initiated by Statistics Canada for the GSS pilot required manual intervention (e.g. sending invitations and reminders), human factors also need to be taken into account. Finally, a trade-off has to be made between the type and amount of recorded paradata to meet short and long term objectives with respect to the development cost and time pressures. It should be noted that the impacts of technical and external factors on the quality of the paradata can be easily confounded.

**4.3 Results and observations about the quality of paradata**

This section presents some preliminary results and observations of the investigations for the GSS26EQ pilot based on the Web paradata and the integrated Blaise Transaction History file (BTH). This includes all CATI transactions and Web activities (i.e. calls made during CATI collection as well as the most important Web activities – i.e. invitation, reminder, login, save, submit, bounce back e-mail – used for reporting and monitoring purposes).

*Missing values*

Missing values for raw Web paradata can occur for either technical or design reasons. However, many of the problems observed with previous versions of the database have already been resolved. Missing values were observed in the integrated BTH (CATI/Web), in particular, for derived variables[[4]](#footnote-4). In the short run, this issue can be easily resolved by processing for many key variables to limit the impact on the scope of our investigations.

*Derived variables*

New derived variables can be defined for reporting, monitoring or analytical purposes. Examples of derived variables include the case’s progress status (not started, in-progress, finalized), the outcome code of some activities (e.g. complete, partial, non-response), the duration of the session or interview time, etc. As mentioned, the derived variables were created as part as of integrated BTH/Web processing to ensure consistency of reporting and ease of analysis. After resolving the issue of missing values for some variables, the accuracy of most derived variables was found to be very good, one exception was the duration variable (see below).

*Duration of session or interview*

It is important to note that the end date and end time of the Web activity are not recorded. Often, the end of the activity corresponds to the beginning of the next one (e.g. login followed by a submit). Therefore, the duration can be estimated by calculating the difference between the last ‘login’ time prior to the ‘submit’ action and the time of the ‘submit’ action (‘Save’ action(s) can also occur in between). Even after making some exclusions, the distribution of the duration still remains very skewed to the right (very long flat tail). Without detailed audit trail information, it is difficult to estimate the duration of the session or the interview to assess respondent burden.

*Coherence in the sequence of activities*

The chronological and logical order of the sequence of transactions and activities for each case is a very important dimension of the quality of the paradata to ensure the interpretability of the overall data collection process. In the GSS26EQ pilot, except for a very few cases that arose due to technical problems, no problems were detected regarding the coherence in the sequence of transactions and activities

Investigations conducted to assess the quality of the process data and its limitations indicate that the quality of the Web and integrated BTH paradata is very good in general. The paradata reflects the overall collection process very well for GSS26EQ and can be used to pursue the first four main research objectives.

5. Highlights of the preliminary investigations

The preliminary investigations focus on the first three research objectives: understanding Web and multi mode data collection processes, providing relevant measures to monitor the data collection process and determining useful metrics for planning and costing purposes. During the GSS26EQ collection process a series of analyses were performed to address these objectives. Given the limited scope and length of the paper, only some highlights are presented. Table 1 shows an EQ acceptance rate of 61.7% (% of respondents cases to CATI that accepted to complete EQ) and a percent of submitted EQ of 38.7% among the 3,487 cases that accepted to complete the EQ for a combined EQ response rate of 23.9%. The EQ acceptance and submission rates represent some of the most important metrics for planning and costing purposes (third objective). The assumptions used to plan the pilot survey were respectively 45% and 45% for a combined rate of 20.3%.

Table1



A series of reports were created to actively monitor key indicators collected throughout the data collection period. The indicators served to monitor the quality of the data collection process as well as to identify potential problems as soon that they occurred (second objective). In addition, many other analyses were conducted during collection to increase our knowledge about the Web data collection process and practices (first objective). For example, the proportion of CATI respondents that login at least once to the EQ application, the percent of EQ submitted after the first invitation and after each reminder, and the average number of days between an invitation (or reminders) were closely monitored during collection.

**5.1 Challenges in assessing and monitoring the impact of non-response**

Given the RDD sample design of the GSS26EQ pilot, information about the characteristics of the sample was not available prior to collection. Information became available for the households that provided the roster and household demographic composition during the CATI phase. Table 2 compares the distribution of the population by age and sex (according to Census 2011) with that of respondents who accepted EQ and also with respondents who effectively submitted an EQ. The younger population (especially males) is under-represented in the group that accepted the EQ. This situation might be partially explained by the fact that this population is also under-represented on the RDD frame. The percentage of submitted questionnaires for the pilot is higher for the older population (45+) which is somewhat surprising.

Table 2



The main challenges with regards to the evaluation and survey monitoring of the non-response for Web surveys are: 1) to closely monitor Nonresponse during collection, especially by domain of interest (when possible); 2) to evaluate the impact of nonresponse on various dimensions of the quality of the survey that includes MSE (variance and potential bias) and mode effect; 3) and finally, to identify the optimal Web and mixed mode strategy (fourth objective) that can minimize the impact of the nonresponse. In the process of determining this strategy, many issues have to be investigated. For example, is it suitable to target specific domains of interest (e.g. those with the lowest EQ acceptance and submission rates) to improve the representativity of the EQ sample? Is it possible to develop a strategy to evaluate the level and impact of EQ proxy response (individual that complete the EQ on behalf of the selected respondent)? The fact that the RDD sampling frame contains no information about the characteristics of sample units prior to collection increases these challenges. Currently the following analyses are performed: comparison of the EQ estimates with the CATI Main estimates (from the GSS Cycle 26 main survey conducted during the same reference period) and comparison of the CATI Main + EQ estimates with the CATI Main estimates. In addition, some interesting questions are also addressed. For example, did the respondents that completed and submitted the EQ (EQ respondents) have different characteristics than those who required one reminder or more? Preliminary results of these comparisons should be available for the workshop.

6. Conclusion

According to the preliminary investigations, Web paradata seems to be as accurate, accessible and timely as paradata available for other collection modes but different in terms of contents and the structure. Results obtained through Web paradata research also indicate a good coherence in the chronological sequences of activities because it is possible to understand the collection process through paradata (and make sense of it (i.e. interpretability)).

In summary, Web paradata meets most of the requirements for reporting, monitoring, analytical and research purposes. However, some improvements have to be made with respect to the integration of Web paradata into CATI BTH and the development of client-side paradata to reduce the gap of information for analysis and research purposes. Finally, the challenge in assessing and monitoring the impact of non-response (and mode effect) in the context of Web and multi-mode surveys still remains and further research will be required.

7. References

Heerwegh D. (2002), *Describing response behavior in Web-surveys using client side paradata*. Paper presented at: International Workshop on Web-surveys; Mannheim, Germany.

Mayda, J. (2002), *Experience with I*mple*mentation of EDR into Existing Survey Programs*. Statistical Journal of the United Nations Economic Commission for Europe, 19(3), 131-140.

Sowan, A.K, Jenkins, L.S. (2010), *Paradata: A New Data Source From Web-Administered Measures,* Continuing Education, CIN: Computers, Informatics, Nursing & Vol. 28, No. 6, 333–342

Sunter, D. (2009), *Internet Data Collection: an Overview of Recent Experiences at Statistics Canada and Internationally*. Internal Statistics Canada report.

Statistics Canada (2002), *Statistics Canada's Quality Assurance Framework*, No. 12-586-XIE.

1. In total about 40 different activities can be potentially recorded but about half of them are currently used. [↑](#footnote-ref-1)
2. Many authors have raised the privacy and technical issues of collecting information about the client’s computer habits. [↑](#footnote-ref-2)
3. But strategies that involve interviewers can be used to contact and try to convince respondents to complete the EQ (e.g. telephone reminder, nonresponse follow-up call). [↑](#footnote-ref-3)
4. The vast majority of the derived variables with missing values on the integrated BTH are missing by design (i.e. few variables are present on both BTH and raw Web paradata) but some are missing for technical reasons. This last problem can be resolved easily. [↑](#footnote-ref-4)