WP.5
ENGLISH ONLY

**UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE**

**CONFERENCE OF EUROPEAN STATISTICIANS**

**Work Session on Statistical Data Editing**
(Vienna, Austria, 21-23 April 2008)

Topic (i): Editing of data acquired through electronic data collection.

# IMPACT OF ONLINE EDITS AND INTERNET FEATURES IN THE 2006 CANADIAN CENSUS

**Supporting Paper**

Prepared by Danielle Laroche and Chantal Grondin, Statistics Canada

## I. INTRODUCTION

1. For the first time in 2006, Canadian households had the option of responding to the Census via the Internet. Almost one in five households (18.3%) chose to do so. This rate is the highest ever achieved for a Census in any country. Introducing this new collection mode in the Census required methodological changes in several areas, such as questionnaire design, security, data quality, mode effect analysis, data analysis and data processing. In the end, the data collected over the Internet had higher item response rates and required less follow-up than those collected using paper.

2. In this paper, we give an overview of the 2006 Census and describe the features of the Internet questionnaire. Some findings about the impact of online edits and some of the Internet features are then presented. Finally, potential opportunities for enhancements to the 2011 Census brought about by these findings are mentioned.

## II. OVERVIEW OF THE 2006 CENSUS

3. Approximately 13 million dwellings and a population of more than 31 million were enumerated in the 2006 Census. The population was counted at their usual place of residence in Canada regardless of their location on Census Day. Self-enumeration was used to count most of the population, while the remainder was counted via interviews or the use of collective dwelling administrative records. Two main types of questionnaires in both official languages were used to collect the majority of the Census data: Form 2A, the short questionnaire with eight questions, was distributed to four out of every five households (80%); Form 2B, the long questionnaire, with 53 person questions and 8 dwelling questions, was distributed to the remaining 20% of households.

4. Major changes were implemented in the 2006 Census methodology in relation to data collection and processing. In order to define and control the enumeration of all dwellings, Statistics Canada created the Master Control System (MCS). This system contained two mutually-exclusive frames that covered all dwellings in Canada, a list frame and an area frame. The **list frame** was derived from updates made to Statistics Canada's Address Register and covered areas for which we had a civic address for each dwelling. Areas in the list frame were called mail-out areas. The **area frame** covered areas for which no dwelling address was available and to which a questionnaire was dropped off or a canvasser methodology was used. These were called list/leave areas and canvasser areas respectively. Each dwelling on the MCS was assigned a unique identifier (Frame ID) and an Internet access code. This access code was linked

either to a physical address (list frame) or to the unique identifier (area frame). All paper questionnaires displayed the identifier, the Internet access code and the Census web site address.

5. Prior to 2006, Census questionnaires were dropped off at households by enumerators. Now, for the first time, questionnaires were mailed out to two-thirds of all dwellings. Households that received a paper questionnaire by mail or drop-off could complete it manually and mail it back, phone the Census Help Line (CHL) and complete it via Computer Assisted Telephone Interview (CATI) or complete it via the Internet. Non-response households were contacted during the Non-Response Follow-Up (NRFU) operation in order to obtain a completed questionnaire. Thus, for the 2006 Census, four response modes were available: 1) Mail-back of paper questionnaire; 2) Computer Assisted Telephone Interview (CATI); 3) Internet; and 4) Canvasser / Field NRFU.

6. All completed questionnaires were returned to a single Data Processing Centre (DPC). Data received from electronic questionnaires were stored on the Census database servers and transmitted on an ongoing basis to the processing servers at the DPC. All questionnaires received were registered using their unique identifier and the MCS was updated on a regular basis. Non-Response Follow-Up (NRFU) lists were created starting ten days after Census day and transmitted to field collection staff regularly to determine the occupancy status of dwellings. Completed questionnaires were obtained for occupied dwellings. Thus, the 2006 MCS was an essential tool for controlling the enumeration of each dwelling and for tracking questionnaires through collection and later processing.

7. All paper questionnaires received were scanned and the data was captured using Optical Mark Recognition (OMR) and Optical Character Recognition (OCR). Keying From Image (KFI) was also used when the OMR or the OCR were of poor quality. Data received from the Internet and from CATI, already in an electronic format, were integrated into the regular flow. Automated edits were performed to identify questions requiring follow-up. A weight reflecting the impact of the missing or invalid data was assigned to each question and a score was compiled for each household. Questionnaires exceeding a predefined value were assigned to Failed Edit Follow-Up (FEFU). Telephone follow-ups were conducted from the three Census Help Line sites using a CATI application adapted from the Internet application.

## III. FEATURES OF THE INTERNET QUESTIONNAIRE

8. In order to minimize the mode effect (i.e. differences in responses due to the collection method used) and to facilitate the integration of data received from different response modes, the electronic versions of the questionnaires corresponded as closely as possible to the paper versions in terms of question wording, instructions and presentation of response choices. Determined efforts were made to adhere to the paper form while incorporating many Internet questionnaire standards and conforming, as much as possible, to the guidelines for presentation of federal government Internet sites.

9. The **short and long online questionnaires** used an interactive multi-page design. With this design, the questionnaire was presented screen-by-screen, each of them displaying a group of questions related to a common theme (questions on labour for example). As with the paper questionnaire, the **matrix format** was mainly used in the online version. A **sequential format** was used in labour market activity and income questions. On both paper and Internet questionnaires, respondents were asked to report their date of birth. An additional screen was added in the Internet version to **confirm the age** of respondents. Confirming the age was particularly important in the long questionnaire as persons younger than 15 were neither presented with the same questions nor subjected to the same online edits as those 15 or over. In fact, no edits were performed for questions related to marital/common-law status, and no questions were asked related to mobility, education, labour market and income for people younger than 15. The Internet questionnaire also had an **additional screen to confirm income** for people 15 years old or over. This additional screen was important because the income part of the questionnaire was very complex, with eleven sub-questions. This confirmation screen provided respondents the opportunity to verify their responses and make corrections if needed.

10. Internet standards were followed, including **check boxes** when multiple responses were possible and **radio buttons** when only one response was allowed. Other electronic features were used, such as the **drop-down menus** for selecting the day and month of birth or for selecting provinces/territories, as well

as **automated skip patterns** to reduce response burden by ensuring respondents were not presented with irrelevant questions.

## IV. ONLINE EDITS IN THE INTERNET APPLICATION

11. When implementing online edits in an application, one has to keep in mind that complex edits can increase the burden on respondents to provide precise responses and therefore increase the time required to complete the questionnaire. The benefits in data quality obtained from complex edits need to be weighed against the increase in respondent burden to achieve an appropriate balance. For the 2006 Census, it was decided to keep the online edits relatively simple; they were performed on one question at a time and no consistency edits were performed between questions. The decision to go this route was based in part on the results of usability tests and qualitative studies done with different versions of the Internet application. These tests were helpful in identifying which edits seemed more appropriate and in line with our goal to stay as close as possible to the paper questionnaire.

12. In the end, the Internet application included four types of **online edits** or **validation messages**. **Non-response** messages appeared when respondents had not answered a question. **Partial response** messages appeared when respondents provided only a partial response to a question, for example, if they omitted the city name from their address. **Invalid response** messages appeared for numerical responses when respondents entered numbers outside the range established for a question. Finally, **amount verification** messages appeared for questions related to money amounts, when the response appeared unusual. This type of message asked respondents to verify that they have entered the correct amount, for example, "*Please verify the amount you entered for part (f), if correct leave as is*". All these messages followed the same approach. When respondents clicked the *Next* button, the information on the current page was validated and, if necessary, the application displayed the same screen again, noting any problems at the top of the page in red text, for example, "*Please answer Question 5 for John Doe*." The question and field requiring attention appeared in red, and a red arrow highlighted the missing response to assist the respondent, who could then either fill in the missing information or continue to the next screen. If the respondent chose to move on without making any changes, the next screen was presented. If the respondent added or changed any information, the responses were validated again. This approach was consistent with the Common Look and Feel guidelines prescribed for Canadian government web sites in that pop-up windows should not be used within pages to convey information to respondents.

## V. IMPACT OF ONLINE EDITS AND INTERNET FEATURES

13. After data capture, questionnaires were transmitted to data processing where completion edits were performed. These edits were based on a score strategy. The questionnaires for households exceeding a predefined threshold value were forwarded to follow-up.

14. The first notable impact when comparing the questionnaires received by the mode of response in the 2006 Census was the difference in rejection rates, as can be seen in the following table.

**Table 1. 2006 Census rejection rates, by mode of response and type of questionnaire**

| Mode of response | Rejection rate (%) by type of questionnaire | |
|---|---|---|
| | Short 2A | Long 2B |
| **Internet** | 2.5 | 5.7 |
| **Mail** | 5.6 | 39.1 |

15. The **much lower rejection rates for Internet questionnaires**, compared to the paper, are a direct result of the overall lower number of partial or invalid answers, as well as lower non-response to each question. These are primarily due to the online edits or validation messages, but also to the other Internet features such as automated skips, radio buttons, online help and explanations of why each question was asked. Another factor that could explain part of this difference in rejection rates may be linked to characteristics of people who chose to report their data on the Internet. Indeed, Internet reporters are more likely to have a higher education level, and this might be associated with a better capacity to respond to the questionnaire.

16. In general, we can say that validation messages were very effective in obtaining answers to questions that respondents might otherwise have overlooked and directing them to correct errors they inadvertently committed. These messages, along with the automated skips for non-applicable questions and the radio buttons that allowed only one response, resulted in a general perception among respondents that the electronic questionnaire was "intelligent". This responded in part to the high expectations the general public had with regard to Internet questionnaires generally and probably encouraged respondents to be more alert when answering their questionnaire. We know however, based on qualitative testing, that some respondents, when receiving validation messages, felt that they had to provide an answer in order to be able to continue. This might have created unwanted behaviours on the part of some respondents, as we will discuss shortly.

17. If we think of reasons why respondents would leave a question blank to begin with, three main reasons come to mind: 1) **inattention** on the respondent's part; 2) the respondent felt that the question was **not applicable**; and 3) the respondent **didn't know the answer** or refused to give it. Unfortunately, it's impossible to know which of these reasons has caused non-response to a specific question for a particular respondent. What we do know based on our experience with Census qualitative studies is which questions are traditionally deemed not applicable by certain respondents, and which ones are more difficult to answer, especially for proxy respondents. Also, non-response due to inattention generally happens at random. Non-response to a question can therefore be attributed to one of these three reasons, but also to a combination of them depending on the circumstances of each respondent.

18. Here are a few examples from the 2006 Census. These initial findings are part of our efforts to evaluate the Internet data quality as well as the mode effect. The rates presented here are unweighted and represent the outcome after all follow-up was done, but before imputation.

19. We first looked at the common-law question. The non-response rate to this question among Internet reporters was 0.48% and for paper reporters, 5.61%. We know from past experience that married people have a tendency to skip the common-law question (no matter which mode they use to report their data) because they feel the question is **not applicable**. Indeed, when looking at the reported marital status of people (15 years or over) who did not answer the common-law question, we see that 7 out of 10 paper reporters were married, while this proportion was 90% among Internet reporters. These proportions are much higher than the proportion of married people in the entire adult population, supporting the observation that the common-law question is often skipped by married people who feel it does not apply. As for the non married people who did not respond to the common law question, the proportion of them is higher among paper reporters (30%) than among Internet reporters (10%). This might indicate that among paper questionnaires, a good part of the non-response was **due to inattention**, while this was much less frequent among Internet reporters, likely due at least in part to the presence of validation messages. Note that there were no validation messages to the common-law question for married people or people less than 15 years of age.

20. Hence, we believe that validation messages have a positive effect in reducing errors due to inattention. But as we said before, they could also become problematic in certain cases, particularly when respondents **don't know the answer to a question**. In these cases, validation messages will prompt respondents to provide an answer which they don't know. Rather than leave it blank, respondents may provide an answer that is valid but incorrect. This is difficult to assess and measure, and would require more study, for example through the use of differently worded validation messages.

21. Here are some results from our initial steps in this research. We looked at the reported day of birth for Internet reporters with a valid date of birth. These reporters were split into two groups: those for whom the first answer provided was either partial or missing, hence for whom a validation message appeared on the screen, and those who entered a valid answer in the first place, and so did not receive any validation message to this question. We looked at the distribution of the day of birth for each group. The distribution showed that for the group who received a validation message, the frequency of the day of birth "1" was much higher (7.3%) than for those who entered a valid answer to begin with (3.6%). It needs further confirmation but this strongly suggests that some respondents may report a valid but incorrect answer when prompted by the validation message. Of course, this behaviour surely happens

among paper reporters as well, but the online validation messages might emphasize it by prompting for an answer. In this case, some small percentage of Internet reporters who received the validation message and did not know the answer tended to chose the first answer in the drop-down menu for day of birth when the actual day of birth was likely not the first answer.

22. Another example where people may have been prompted into providing a valid but incorrect answer relates to the postal code of the previous address for people who moved to a different city (i.e. migrants). For these people, we compared the postal code of the current residence to the one of the residence from one year ago and five years ago. We saw that among Internet reporters, there was a higher proportion of migrants reporting the same postal code for both their current and previous addresses compared to paper reporters, as can be seen in Table 2 below.

**Table 2. Proportion of migrants who reported the same postal code before and after migration**

| Response mode | 1 year mobility | 5 years mobility |
|---|---|---|
| **Internet** | 4.3% | 13.2% |
| **Paper** | 3.6% | 9.5% |

23. These results encourage us to further analyze the possibility that valid but incorrect responses are provided more frequently on the Internet than on paper questionnaires, and the role that validation messages play in this phenomenon.

24. A phenomenon with similarities to what we just described for day of birth and postal code of previous address happened with the number of hours worked. Although the actual numbers were very small, there was a higher proportion of "1 hour worked" noted for Internet reporters compared to paper reporters. We know from our experience with the Census questionnaire that among paper reporters, respondents occasionally enter a "0" in the text box for number of hours worked instead of checking the "None" box. We believe the same has happened with the Internet reporters, but for them, a validation message associated to the text box said that valid values had to be between 1 and 168. Again, further research is needed to confirm the hypothesis but it seems likely that some of the "1 hour" answers from Internet reporters were reported for people who did not work, but were not clear on how to indicate this. This situation differs from the one related to day of birth in that it is not because the respondent didn't know the answer, but rather because he **didn't know how to report it**.

25. Another of the positive impacts of following the Internet standards in the 2006 application was the use of radio buttons for questions where multiple answers were not allowed. The direct consequence of this was to make it impossible to have **invalid** answers (due to multiple answers) for these questions for Internet reporters. Although invalid rates are usually rather low, this still represents a valuable quality improvement for a large number of questions. For example, the largest invalid rate due to multiple answers was seen for the question on activity limitation at work or at school, with 0.38% multiple answers on paper questionnaires. The average invalid rate due to multiple answers on paper was 0.09%. For Internet reporters, the invalid rates due to multiple answers for the same questions was of course zero because radio buttons made it impossible to enter multiple answers.

26. In general, we found that **drop-down menus** worked very well. Respondents found them easy to use and data quality was good. But care is needed as the following example will illustrate there are specific situations where difficulties can arise. Response error can occur for a very small fraction of respondents because once a selection is made in a drop-down menu, one must click outside of the menu (or use the tab key) to move the browser's "focus" away from the drop down list. If the scroll wheel on the mouse is used while the drop-down menu is still "selected"– for example in an attempt to scroll down to the next question after responding – the response will be changed, sometimes without the respondent noticing. The result in such situations is that for this small fraction of respondents, their answer tended to be moved to responses lower – often the last – on the list. Edit and imputation strategies were put in place to detect and remedy the problem. In other situations where the next question was immediately visible without scrolling, we found no evidence of this kind of problem.

27. Despite these observed problems which are easy to fix for the next Census, the data received from the Internet are clearly more complete, require less follow-up and therefore are less costly to process than data received from paper forms. In addition, Internet was and still continues to be a solution to a number of issues such as respondent's privacy concerns regarding local enumerators, the recruitment of a large temporary workforce, the resources needed for keying, etc. In Canada, offering an Internet option is also consistent with Canadian Government policy.

## VI. POTENTIAL ENHANCEMENTS TO CONSIDER FOR THE 2011 CENSUS INTERNET APPLICATION

28. Based on the results of comparing the non-response rates, invalid rates, and distribution of answers by collection mode, a few changes will be considered for the Internet application for the next Census.

29. Other options for drop-down menus for the province/territories answers will be considered for all province/territories questions.

30. As well, a change will be considered to the online edit for the question on number of hours worked to allow a "0" value to be entered in the text box in which case it will be treated as if the "None" check box was selected.

31. Other changes might be done to remove or change some of the validation messages where we feel that respondents could potentially be prompted to enter a valid but incorrect response.

## VII. CONCLUSION

32. From a data reliability standpoint, the Internet response option offered many advantages and is generally considered an enhancement. As demonstrated in this report, questionnaire rejection rates are much lower for Internet questionnaires than for paper ones. With regard to validation messages, the results indicate that in general, they are effective in obtaining answers to questions that respondents might otherwise have overlooked and in having respondents correct errors they inadvertently committed. Along with the automated skips of non-applicable questions, these messages result in a general perception among respondents that the electronic questionnaire is "intelligent". This responds in part to the high expectations the general public has with regard to electronic questionnaires. More research should be conducted to better understand how respondents react to validation messages. For example, would an effective strategy be to ask respondents to provide an answer to the best of their knowledge, or otherwise leave it blank?

33. The data collected in 2006 provides an opportunity to conduct further research into the impact of the Internet response mode. Of course, we must try to minimize mode effects as much as possible when introducing a new collection mode, but realistically, it may not be feasible to completely avoid it.

34. Adding more sophisticated edits might help in ensuring higher quality data, but unfortunately, it could also be at the expense of increasing respondent burden by making the application less user-friendly and more time-consuming. We must achieve the right balance between quality and response burden.

## VIII. BIBLIOGRAPHY

Laroche, D. (2005). United Nations Economic Commission For Europe – Statistical Data Editing, Volume No. 3, Impact on Data Quality, 2006, "Evaluation Report on the Internet Option of the 2004 Census Test: Characteristics of the Electronic Questionnaires, Non-Response Rates, Follow-up Rates and Qualitative Studies", Ottawa, 2005.

Laroche, D. (2007). Advisory Committee on Statistical Methods, "2006 Census Internet Data Collection", Meeting No. 44, April 30th and May 1st, 2007.