Testing the Addition of the Social Insurance Number to the Canadian Census of Population

Cilanne Boulet¹ and Patrice Mathieu¹ ¹Social Survey Methods Division, Statistics Canada, 100 Tunney's Pasture Driveway, Ottawa, ON K1A 0T6, Canada

Abstract

Statistics Canada is involved in a comprehensive review of the potential for administrative and other alternative data sources to replace, complement or supplement the Agency's census and survey programs. As part of this review, the possible addition of a question asking respondents to provide their Social Insurance Number (SIN) on the Canadian Census of Population questionnaire was tested. This article gives an overview of this test and of its principal results.

Key Words: Canadian Census of Population, intercensal test, Social Insurance Number

1. Introduction

Statistics Canada is involved in a comprehensive review of the potential for administrative and other alternative data sources to replace, complement, or supplement the Agency's census and survey programs. Initial research is ongoing to evaluate the potential for a census based principally on tax and benefits data via the creation of an administrative spine. Record linkage of administrative data sources without the availability of a unique identifier or key is one of the many challenges of such a project. Within this context, one option investigated to facilitate the development and evaluation of various ways to use administrative data sources was to ask respondents to provide their Social Insurance Number (SIN) on the Canadian Census of Population.

In Canada, there is no universal identification number assigned to every Canadian citizen or resident. The SIN, used for the administration of income tax and various governmental benefits programs, is the closest that Canada comes to having such an identifier. It is not universal; in particular, children are less likely to have a SIN since it has to be requested and is often only requested when a person enters the work force. That being said, the SIN does have the broadest coverage of any personal identifying number in Canada.

The possibly of asking for the SIN on the Census was tested as part of the 2014 Census Content Test. This test had two principal objectives: to assess the operational feasibility of requesting the SIN on the Census and to determine the degree to which the SINs collected could be used to improve record linkage to tax files. This article begins by describing the key features of the test design in section 2. The second part of the article, sections 3 and 4, presents the results of the test, in particular with respect to its principal objectives.

2. Test Design

Prior to the 2014 SIN Test, Statistics Canada had never asked respondents to provide their Social Insurance Number on a questionnaire. As a matter of practice, the SIN is generally used in a limited number of situations and Canadians are instructed by the government not to give out their SIN except in limited circumstances. As a result, before undertaking this test, very little was known about what to expect when asking for the SIN. How would respondents react to the question? Would respondents be willing to provide their SIN? Would they be able to provide SINs for all household members? Would asking for the SIN influence their decision to respond to the Census overall or to other questions on the Census? The planned test needed to be comprehensive enough to conclusively determine whether Statistics Canada should or should not ask for the SIN on the Census and the planning period available before the next Census in 2016 only allowed for one test rather than a sequence of tests. In this context, with so many unknowns and only one test to address these questions, it was deemed essential to replicate Census methods and conditions as much as possible and for the test sample to cover the Canadian population as broadly as possible.

2.1 Replicating Census Conditions

The Census of Population is mandatory, with every resident of Canada having the legal obligation to respond. To reproduce the conditions under which respondents would encounter the SIN question, it was decided to make the 2014 SIN Test mandatory as well. This is atypical. Most social surveys at Statistics Canada are voluntary, with the Census and Labour Force Surveys being notable exceptions. Previous intercensal tests have always followed this model and have been conducted as voluntary surveys without imposing a legal requirement to respond on the selected households.

There were many reasons for making the test mandatory. Without being mandatory, the test ran the risk of being completed only by more cooperative respondents. The reaction of these respondents to the addition of the SIN question could very well be different than the reaction of the remaining population that would choose not to respond to a voluntary test. This was a particular risk given that one of the unknowns was whether the SIN question would lead to additional non-response. In addition to determining whether the question would have an impact on response rate, the test had to give sufficiently precise estimates of the magnitude of the impact on response rates to be used for planning purposes; for instance, to predict required questionnaire and non-response follow-up volumes.

The reasons for which the test needed to be mandatory also made the inclusion of a nonresponse follow-up operation during which respondents are contacted both by telephone and in person essential. In particular, it was through this operation that less cooperative respondents or respondents who initially hesitated to provide their SIN could be reached. Interviews also provided an opportunity to directly observe attitudes and response issues related to the SIN question.

The Census uses a multi-mode collection strategy, where multiple modes of contact and response are used concurrently and sequentially over a period of about three months. Respondents may self-respond to the Census by completing an electronic questionnaire online, by returning a completed paper questionnaire via Canada Post, or by calling the Census Help Line by telephone. Households that do not self-respond are contacted during

non-response follow-up (NRFU) either by telephone or in person. Since it was suspected that the quantity and quality of responses to the SIN question might depend on the mode used to complete the Census, all response modes had to be covered by the test and the sample had to be sufficiently large to have a sizeable portion of the sample using each mode.

In addition to using multiple modes, the Census uses multiple contact strategies. For some dwellings, the first Census contact is a letter inviting them to go online to complete the Census; for other dwellings, the first Census contact is receiving a questionnaire. Depending on how they are first contacted, respondents may face the SIN question before or after they have initiated a self-response, leading to potentially different reactions. In the test, two different strategies were used to contact respondents. In the first, dwellings were sent a letter inviting households to complete the Census online, followed by a reminder letter sent to non-respondent dwellings, followed by a questionnaire, and finally, a subset of dwellings that remained non-respondent after the first month of contact was included in non-response follow-up. In the second strategy, dwellings were initially sent a questionnaire, followed by a reminder letter, a second questionnaire, and finally a subset of non-respondent dwellings was included in non-response follow-up. With a few exceptions, these are the contact strategies that will be used to contact over 98% of dwellings during the next Census cycle. For more information about the Census collection methodology, see Mathieu et al. (2012).

The Canadian Census programme includes not only the Census questionnaire but also the National Household Survey (NHS), the voluntary survey which replaced the Census long form in 2011. NHS collection is integrated with Census collection operations. The paper questionnaires sent to dwellings selected for the NHS are combined questionnaires that include both the mandatory Census questions and the voluntary NHS questions. By electronic questionnaire, respondents are first asked to complete the Census. Once they have submitted their Census questionnaire, those in dwellings selected for the NHS are presented with an invitation to continue with the NHS questions. The objective of the SIN Test therefore included evaluating the potential impact of the SIN question on the NHS, in addition to its impact on the Census, especially since as a voluntary survey the NHS was thought to be at greater risk of increased non-response. As indicated in the test design discussed below, the SIN Test included both dwellings that were selected for the NHS and not selected for the NHS and took into account the interaction between the NHS and the contact strategies.

Finally, the SIN Test was conducted over the same time period of the year as the Census with collection operations beginning on May 5, 2014 and a reference date of May 13, 2014.

2.2 Sample Selection

The list of occupied private dwellings from the 2011 Census was used as a sampling frame for the SIN Test. For operational reasons, the sample was restricted to mailout areas in the ten provinces; that is, the areas in which Canada Post is used to deliver Census materials. These areas include over 80% of dwellings in Canada but exclude predominantly rural areas.

The SIN Test was conducted using a two-stage probability sample of 30,000 dwellings. Clusters were defined to be geographic areas small enough for one interviewer to cover. Cluster selection was done within 37 geographic strata. This stratification ensured adequate coverage in each of the ten provinces. Within each stratum, clusters were selected proportionally to their size. At the second stage of the sample selection, 75 dwellings were selected by simple random sampling within each of the 400 clusters selected at the first stage.

While concurrent with other content testing done in preparation for the 2016 Census, the SIN Test was conducted using a separate non-overlapping sample to avoid influencing the evaluation of other proposed content revisions.

2.3 Test Panel Design

The SIN Test used a randomized experimental design. The sample was split into three pairs of panels of 5000 dwellings each. Within each pair, one panel received a questionnaire with the SIN question and the other panel received a questionnaire without the SIN question. This was the only difference between the questionnaires. The three pairs of panels were used to cover the two different contact strategies (letter as first contact or questionnaire as first contact) as well as the possibility of being selected or not for the NHS. Table 1 summarizes the panel structure of the SIN Test. It was not necessary to include a panel consisting of dwellings receiving a letter as first contact and not selected for the NHS since the vast majority of respondents contacted by letter respond by electronic questionnaire and, in this mode, they do not see the invitation to continue on to the NHS questions until after the Census questionnaire has been submitted.

		Questio	onnaire
Contact strategy	NHS	With SIN question	Without SIN question
Questionnaire, Reminder letter, Questionnaire, NRFU (if subsampled)	No, Census only	Panel 1 (test panel)	Panel 4 (control panel)
Questionnaire, Reminder letter, Questionnaire, NRFU (if subsampled)	Yes, Census and NHS	Panel 2 (test panel)	Panel 5 (control panel)
Letter, Reminder letter, Questionnaire, NRFU (if subsampled)	Yes, Census and NHS	Panel 3 (test panel)	Panel 6 (control panel)

Fable	1:	SIN	Test	panels
--------------	----	-----	------	--------

After having been selected for the SIN Test sample, dwellings were assigned to each panel as follows. Within each cluster, dwellings were assigned systematically to panels 1 to 6 following a geographic ordering. As a result, each panel contained the same number of units from each cluster (plus or minus one). The principal statistics of interest for this test were the differences arising between the test and the control panels having the same contact strategy and presence or absence of the NHS. Assigning dwellings to panels was done in this manner to reduce the impact of correlation within the clusters on the precision with which these differences could be estimated.

3. Impact of the SIN Question on Respondent Behaviour and Collection Operations

The impact of adding the SIN question to the Census questionnaire on respondent behaviour and on collection operations was evaluated along three main themes: a comparison of return rates, an examination of paradata documenting respondent interaction with the electronic questionnaire, and additional information collected from interviewers about the non-response follow-up phase.

3.1 Return Rates

Return rates were compared between pairs of test and control panels to evaluate whether asking respondents to provide their SIN resulted in some of them deciding not to complete or return their questionnaire right away or at all.

Since non-response follow-up operations for the SIN Test were shorter (one month rather than two) there were more dwellings with unknown occupancy status at the end of collection than would be observed at the end Census collection. As a result, we decided to compare return rates with a constant denominator equal to the size of the sample in the panel rather than response rates whose denominator would have been an estimate of the number of occupied dwellings in the panel. Our sample and test design ensured that occupancy rates were comparable in all panels.

Since non-response follow-up was conducted on a subsample of dwellings that were still non-respondent after about one month of self-response, return rates were weighted following the paradigm of Hansen and Hurwitz (1946).

3.1.1 Census return rates

Final Census weighted self-return rates and weighted return rates including non-response follow-up are shown in Table 2. None of the differences between the test and the control panels are statistically significant at a 95% confidence level. We conclude that neither does the SIN question lead to lower final response rates nor does it lead to lower final self-return rates.

However, a closer examination of self-response reveals an interesting pattern. When we examine self-return rates by date over the self-response period, we notice that self-return rates are consistently lower when the SIN question is asked than when it is not. Figure 1 shows self-return rates by date for the panels whose first contact is a letter, the most commonly used Census contact strategy. Given the final self-return rates, it is not that the SIN question lowers the overall self-return rate; rather, when the SIN question is asked self-returns are received more slowly.

Slower self-response is a concern from a collection operations perspective. It means more reminder letters and questionnaires have to be sent out in follow-up waves and it results in a larger initial non-response follow-up workload. Even the relatively small difference observed in the curves above is enough to have a non-negligible operational impact.

		Final Censu rate (exclud weig	us self-return ing NRFU) - ghted	Final Censu (including weig	s return rate 3 NRFU) - hted
Contact strategy	NHS	With SIN question	Without SIN question	With SIN question	Without SIN question
Questionnaire, Reminder letter, Questionnaire, NRFU (if subsampled)	No, Census only	79.4%	78.1%	88.9%	88.7%
Questionnaire, Reminder letter, Questionnaire, NRFU (if subsampled)	Yes, Census and NHS	80.1%	80.4%	87.7%	88.6%
Letter, Reminder letter, Questionnaire, NRFU (if subsampled)	Yes, Census and NHS	79.8%	79.4%	88.7%	88.8%

Table 2:	Final	Census	return	rates	-	weighted
----------	-------	--------	--------	-------	---	----------



Figure 1: Census weighted self-returns by date

3.1.2 National Household Survey return rates

The National Household Survey return rates were also examined. As a voluntary survey, we wanted to verify whether NHS returns might be influenced by the SIN question. Table 3 shows the final NHS weighted self-return rates and weighted return rates including non-response follow-up. As was the case for the Census, none of the

differences between return rates for comparable panels are statistically significantly different from zero. However, in this case, all rates are lower when the SIN question is included in the Census questionnaire than when it is not.

		Final NHS se (excluding weig	Final NHS self-return rate (excluding NRFU) - (incl weighted		S return rate 1g NRFU) - ighted	
Contact strategy	NHS	With SIN question	Without SIN question	With SIN question	Without SIN question	
Questionnaire, Reminder letter, Questionnaire, NRFU (if subsampled)	Yes, Census and NHS	72.0%	73.0%	79.1%	80.6%	
Letter, Reminder letter, Questionnaire, NRFU (if subsampled)	Yes, Census and NHS	65.3%	66.5%	73.7%	75.5%	

Table 3: Final NHS return rates - weighted

These observed differences can be explained by considering NHS returns by mode. Table 4 shows the NHS transition rate by mode, that is, the percentage of NHS-selected Census respondents using a given mode that continue and also complete the NHS. Here a statistically significant difference emerges. Internet respondents are 2% less likely to complete the NHS when asked to provide their SIN than when the SIN question is not included. Since internet collection is the main mode of collection for the Census, this difference takes on a particular importance. The reasons for this difference are examined in the next section.

Table 4: NHS	transition	rates b	y mode
--------------	------------	---------	--------

	Percentage of Census respondents who also completed the NHS		
Response mode	With SIN question	Without SIN question	
Internet*	78.0%*	80.0%*	
Paper	95.3%	95.4%	
Non-response follow-up	91.8%	92.0%	

* Non-zero difference statistically significant at a 95% confidence level.

3.2 Respondent Behaviour on the Electronic Questionnaire

The internet (or electronic questionnaire) is the most common mode used by respondents to complete the Census and the NHS. For these respondents, we have a wealth of paradata, data about interactions with the electronic questionnaire application such as login time and time spent on various screens, from which we can construct a picture of respondent behaviour.

One of the most striking differences observed was the number of sessions required for a respondent to complete the Census questionnaire with and without the SIN question. Most households complete all Census questions the first time they logged into the electronic questionnaire system. Without the SIN question about 3% of households who eventually submitted the Census by internet did not complete the questionnaire on the first login but returned to do so at a later time. With the SIN question, this percentage increased to 20%. That is, about 20% of households did not complete the Census questionnaire on their first try but had to return and complete it at a later time when the SIN question. This behaviour is highly correlated with household size, with about one third of households with four or more people needing more than one session to complete the Census when asked to provide their SINs.

We also saw differences in the amount of time required to complete the Census questions. Among households who completed the entire questionnaire in one session (that is, those who did not stop and return later) those asked to provide their SINs took an average of 20% more time. Since these are the households who showed the least difficulty at the SIN question, this is a lower bound on the additional time required to complete the SIN question.

The SIN question help screen was the most consulted help screen. The SIN validation edit, a pop-up message asking respondents to provide their SIN if they had not done so or if the number provided was not a valid SIN based on a check digit, was the most triggered edit.

Taken together, these observations are quantitative evidence that the SIN question is not like the other questions on the Census. The SIN question is more difficult to answer and imposes a greater response burden on respondents than the other Census questions.

3.3 Non-response Follow-up

Non-response follow-up was conducted on a subsample of the dwellings that had not responded after the first month of self-response. Four hundred interviewers participated in the test and about 10,000 dwellings were included in non-response follow-up operations. As during usual Census operations, interviewers used both in-person and telephone attempts to contact non-responding households. Each interviewer was assigned a cluster and therefore did non-response follow-up on dwellings that were asked the SIN question and dwellings that were not asked the SIN question.

To collect additional information about respondent reactions to the SIN question, interviewers were asked to complete an exit module after each interview. In this module, they recorded whether they had collected the SIN for some or all household members. If the SIN had not been provided by the respondent, the reason for which this question was not completed was also recorded. Finally, interviewers could also provide additional comments in the exit module and via debriefing questionnaires completed at the end of non-response follow-up operations.

Using these tools, a qualitative evaluation of the impact of the SIN question on nonresponse follow-up could be deduced. We concluded that the SIN question had a small negative impact on the success of follow-up attempts. Additional time was required to collect the SIN question, generally less than 5 minutes but with anecdotal exceptions of longer periods. Interviewers reported having 'some' difficulties with the SIN question; for instance, respondents asking for an explanation of why the question was asked. Some respondents strongly refused to provide their SIN and reluctant respondents were not easily swayed.

In terms of reasons for not providing the SIN, interviewers reported that the most common reason a respondent would not provide their own SIN was refusal and that the most common reason for not providing the SIN of another household member was that they did not know the number.

A difference in respondent behaviour by mode was also observed during non-response follow-up. By looking at response rates to the SIN question by mode in Table 5 below, we observe that respondents were more reluctant to provide the SIN over the phone than in person. Moreover, the SIN question led to more requests for in-person interviews from respondents contacted by telephone. Since in-person interviews are more expensive than telephone interviews, this has the potential to increase collection costs.

4. Utility of the SIN Question

The utility of asking the SIN question depends on two factors: first, how many SINs would be collected by asking the question and, second, by how much those SINs improve our ability to link Census records to records such as tax and benefits files.

4.1 Responses to the SIN Question

Not all Census respondents have a SIN. Therefore, the tested SIN question asked respondents to provide their SIN or to check a box indicating that they do not have a SIN. There is evidence that this box was used by some respondents who have a SIN but did not want to provide it on the questionnaire. As with all questions, some amount of measurement error is expected in the obtained responses. During the SIN Test, some respondents recorded a SIN which was not valid based on a validation rule. Others interchanged SINs between various members of their household. Both of these types of errors will be referred to as providing an invalid SIN.

Overall, in the SIN Test, 78.5% of sample respondents who were asked the SIN question provided a valid SIN, but the percentage of respondents who provided their SIN varied substantially by mode of response, as shown in Table 5.

	Responses to the SIN question by response mode				
Response mode	Valid SIN provided	'Have no SIN' box checked	Blank response or invalid SIN		
Internet	88.6%	9.4%	2.0%		
Paper	75.5%	4.7%	19.8%		
Non-response follow-up – by phone	45.3%	8.9%	45.8%		
Non-response follow-up – in person	53.3%	12.0%	34.7%		

Table 5: Responses to the SIN question by response mode

One reason that may explain the larger proportion of valid SINs collected when the questionnaire was administered over the Internet was the validation edit which was triggered when this question was left blank. Looking at the other modes, it is unfortunate to note that fewer respondents provided their SIN during non-response follow-up since they correspond to a part of the population that is more difficult to link to tax and benefits files.

The percentage of SINs provided also varied substantially by whether respondents were answering questions about themselves or other household members, as shown in Tables 6 and 7. Some of the differences between responses for Person 1, who is generally the person completing the questionnaire, and responses for other household members can be attributed to other household members (for instance, children) being less likely to have a SIN. However, the percentages of blank or invalid responses for Person 1 and for other household members indicate that respondents could provide their own SIN more easily than the SINs of other members of their household.

	Responses to the SIN question by response mode for Person 1				
Response mode	Valid SIN provided	'Have no SIN' box checked	Blank response or invalid SIN		
Internet	96.5%	2.1%	1.4%		
Paper	82.1%	0.4%	17.5%		
Non-response follow-up – by phone	66.7%	0.9%	32.3%		
Non-response follow-up – in person	74.2%	2.4%	23.4%		

Table 6: Responses to the SIN question by response mode for Person 1

 Table 7: Responses to the SIN question by response mode for household members other than Person 1

	Responses to the SIN question by response mode for household members other than Person 1				
Response mode	Valid SIN provided	'Have no SIN' box checked	Blank response or invalid SIN		
Internet	83.8%	13.9%	2.3%		
Paper	70.1%	8.2%	21.7%		
Non-response follow-up – by phone	31.0%	14.2%	54.8%		
Non-response follow-up – in person	40.0%	18.1%	41.9%		

Conditioning on the expected prevalence of the various response modes, we can expect that if the SIN question were asked in 2016, 79.5% of the population would provide a

valid SIN and 8.7% of the population would check the 'have no SIN' box. The remaining 11.9% would be a combination of non-response and invalid responses.

4.2 Impact on Record Linkage to Tax and Benefits Files

The SIN Test was conducted in the context of a push to increase the use of administrative and other alternative data sources at Statistics Canada. In this context, we wanted to verify whether collecting the SIN on the Census could help in the development improved approaches for using administrative data or in the evaluation of these methods. The usefulness of collecting the SIN for either of these goals depends on the degree to which the collected SINs could improve record linkage between the Census and tax and benefits files.

To this end, two linkages of SIN Test respondents to tax and benefits files were performed. The first linkage was done using only geographic, contact and demographic variables, without using the SIN. This linkage was carried through a sequence of 28 deterministic record linkage steps. For the second linkage, the SIN was used in addition to these other variables, thus providing a way to measure the gain that comes from using the SIN.

Respondents were linked to cumulative tax and benefits files dating from 1981 to 2012. Respondents could be linked to two types of records in these files: records for individuals with a SIN on tax and benefits files and records for individuals without a SIN on tax and benefits files. Records for individuals without a SIN on these files are indexed by the Dependent Identification Number (DIN), a number used internally for the administration of these files but unknown to the respondents. Respondents were linked to the cumulative set of files from 1981 to 2012 because, once a respondent was found within this set, the respondent could be tracked to other years via the SIN or DIN available on the file.

The results of the linkage, in terms of the number of links that could be established, are as follows. Without using the SIN, 94.7% of respondents were linked to the cumulative tax files: 89.2% to individuals possessing a SIN on these files and 5.5% to individuals without a SIN but with a DIN. With the SIN, 95.5% of respondents were linked to the cumulative tax files: 90.0% to individuals possessing a SIN on these files and the same 5.5% to individuals without a SIN but with a DIN. In other words, with the linkage strategy used, about 0.8% more respondents were linked to tax and benefits files when using the SIN.

It is worth noting some of the factors that contribute to our ability to link such a high percentage of respondents to tax and benefits files. Much of this success is due to the quality of the geographic information available on both the Census (or in this case the SIN Test) questionnaire and the tax and benefits files. An additional factor helping the linkage is the high quality of name and demographic information collected on the Census electronic questionnaire.

The SIN also provided a way to verify the accuracy of the first linkage carried out using only geographic, contact and demographic variables, without using the SIN. This first linkage established 30,179 links for respondents who provided their SIN on the Census questionnaire to tax and benefits records with a SIN. In only four out of the 30,179 links (or about 0.01%) were the SINs different. Even in those four instances, we were not able to establish whether the link was erroneous. These differences could very well have been a result of incomplete information on tax files about respondents having multiple SINs.

This comparison confirms the excellent quality of the record linkage carried out without the SIN.

Overall, using the SIN resulted in a very small increase in the number of links that could be established between Census data and tax and benefits files. It also showed that the quality of links established without the SIN would barely, if at all, be improved by using the SIN.

5. Conclusions

The SIN Test brought to light the following impacts of adding the SIN question to the Census: an increase in the time required to complete the Census of at least 20% based on the evaluation of responses obtained by electronic questionnaire, with this increase in response burden being especially present for large households; a decrease in the NHS return rate by internet of approximately 2%; and slower self-response resulting in an increased volume of reminder letters and questionnaires to be sent out as well as in a larger initial non-response follow-up workload. Moreover, the record linkage carried out showed that a record linkage of excellent accuracy could be achieved without the SIN and that the number of additional links established by using the SIN would be quite small.

In light of these conclusions and given the additional risks associated with the inclusion of the SIN question on the Census questionnaire, it was decided not to include the SIN question on the 2016 Canadian Census of Population.

Acknowledgements

Implementing the SIN Test was a collective endeavour. Though the number of people who contributed to pulling off this test makes it impossible to name them all, the work of the Census content test team was impressive and cannot go unacknowledged.

Particular thanks go to David Dolson, Marc Hamel and Yulia Romaniuk for their valuable comments which helped shape and improve this manuscript.

References

- Hansen, M.H. and Hurwitz, W.N. (1946). The problem of nonresponse in sample surveys. Jour. Amer. Stat. Assoc., 41, 517-429.
- Mathieu, P., Morin, J.-P., Rodrigue, J.-F. and Taylor, J. (2012). La méthodologie de collecte du recensement et de l'enquête nationale auprès des ménages de 2011 au Canada. Paper presented at the Journées de méthodologie statistiques de l'INSEE, Paris, France, January 24-26, 2012.