

# The Importance of the Long Form Census to Canada

DAVID A. GREEN AND KEVIN MILLIGAN

*Department of Economics*

*University of British Columbia, Vancouver*

In June the federal government published plans to replace the mandatory long form census with a new National Household Survey (NHS) for the 2011 Census cycle. The new NHS is to be circulated to more households but will be voluntary rather than mandatory. The announcement generated a response unique both for its breadth across civil society and its near uniformity. The breadth reflects the plethora of uses for census data, stretching into so many important decisions made by businesses, municipal and provincial governments, and non-profits. The near uniformity reflects the certainty of science on the statistical nature of voluntary versus mandatory sampling techniques. We begin by showing the statistical importance of the distinction between voluntary and mandatory sampling techniques, and then explain how the ubiquity of the census in Canada's national statistics is even greater than many appreciate due to the census's role as the ultimate benchmark for other surveys.

## WHY VOLUNTARY SAMPLING FAILS

To understand the issues in choosing between the mandatory long form census and the NHS, it is useful to consider an example. Suppose that we are interested in knowing the value of a parameter for the whole adult population of Canada—mean income, for example. The most accurate way to proceed would be to get the incomes for every adult and then average them. Mandatory censuses, which have been used in Canada and around the world, aim to get responses for everyone in the country. But if it is too costly to survey everyone in the country, then we can get an estimate of adult mean income by surveying a sample of the

population. The key issue is the conditions under which a given sample will provide an accurate, or unbiased, estimate of the population mean.

To continue with our example, suppose we selected two adults at random from the whole population. By basic laws of statistics, the average of their incomes would be an unbiased estimate of the true population mean. But for any specific sample, we would not expect the sample mean to be the same as the population mean. The difference between the sample and population mean, however, would not be systematically too high or too low. If we drew many such samples and then took the average of the mean incomes for each of the two-person samples, that average would equal the mean income for the entire adult population.<sup>1</sup> This, in fact, is the definition of an unbiased sample: one that delivers estimates that are not systematically off in one direction or the other.

Of course, in most surveys we sample many more than two people. The larger we make the sample size, the closer we expect the sample mean in any specific sample to be to the population mean. This is called the Law of Large Numbers, and it is what underlies the size of the long form census. With a sample of 20 percent of the Canadian population chosen at random, we can expect that the sample mean income will be very close to the true population mean.

What happens with a voluntary survey? As we discuss below, a sizeable proportion of people do not answer such surveys. If people refuse to answer at random, then non-response would not cause a problem; we just would not get as large a sample.

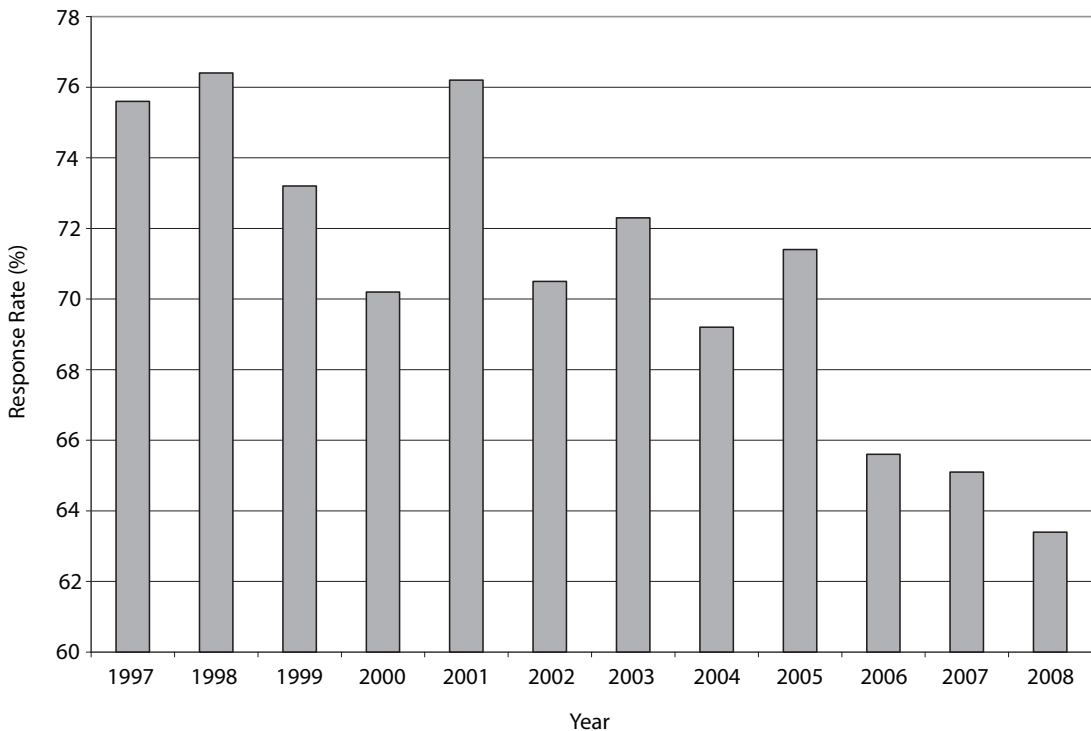
We would expect more variation in the mean incomes we calculate, but they would not be systematically different from the population mean. But if there is something systematic in non-response (if, say, poor people respond less than rich people), then there will be biases—the sample produced from the survey will not be representative of the Canadian population. It is as if we are sampling from a different population; rather than sampling from the population of all adults in Canada, we are sampling from the population of adults who chose to respond to a survey. Increasing the sample size will not fix the fact that we are sampling from a systematically different population.

If there is systematic bias in the response rates across groups in Canadian society, then Industry

Minister Tony Clement's claim that the accuracy of a voluntary survey is preserved by sampling more households is wrong. It is an incorrect application of the Law of Large Numbers. Former Chief Statistician Munir Sheikh made this point famously in his resignation letter with his definitive statement that a voluntary survey cannot substitute for a mandatory census.

How pervasive is non-response in voluntary surveys? Response rates in voluntary surveys conducted by Statistics Canada are in the range of 60 to 70 percent. To take one example, the Survey of Household Spending, which is used as part of the construction of the Consumer Price Index, had a response rate of 63.4 percent in 2008. As shown in Figure 1, this response rate has fallen since the

FIGURE 1  
Survey of Household Spending Response Rates, 1997–2008



Source: Authors' compilation from Income Statistics Division, Statistics Canada, "User Guide for the Public-use Microdata File Survey of Household Spending."

1990s, and the trend shows no sign of abating. This example is representative of the broad patterns in response rate levels and trends in Canada and abroad.

Crucially, there is strong evidence that survey non-response is non-random. That is, certain groups systematically are less likely to respond. There are many examples of this fact, but we focus here on one example coming out of our own research. In three studies (Frenette, Green, and Picot 2004; 2006; Frenette, Green, and Milligan 2007), we examined trends in family income inequality over the last three decades in Canada. In the 2006 study we compared measures of the income distribution based on three different sources. The first was a combination of the Survey of Consumer Finances (SCF), which ran annually up to 1997, and the Survey of Labour and Income Dynamics (SLID), which took over as the flagship income and labour survey after 1997. These surveys were both special voluntary surveys sent to a subset of those surveyed for the Labour Force Survey. Their response rates changed over time but were near 80 percent for both. We also examined incomes using census data and tax data. We found very substantial differences between the SCF/SLID and the other two data sources, with a tendency for the SCF/SLID to overstate incomes at the bottom of the distribution and to some extent to understate incomes at the top. Table 1 recreates part of Table 3.5 from that study, showing the ratio of either the mean income from the SCF/SLID or from tax data to the mean income from census data for a variety of vintiles for 1995 (a year in which we have data from all three sources).<sup>2</sup>

The column showing the relative means obtained from census and tax data shows a relatively close agreement across the income distribution.<sup>3</sup> In contrast, for the poorest 5 percent of families, the mean income reported in the SCF is over double that in the census and the tax data. This problem decreases at higher incomes. The SCF and census data provide nearly identical mean incomes for those in the middle of the distribution. A less severe problem emerges at the top, where average income

TABLE 1  
Ratios of Mean Income by Vintile from Various Data Sources, 1995

| <i>Vintile</i> | <i>SCF/Census</i> | <i>Tax Data/Census</i> |
|----------------|-------------------|------------------------|
| Bottom         | 2.33              | 0.87                   |
| 2nd            | 1.26              | 0.94                   |
| 3rd            | 1.13              | 0.89                   |
| 4th            | 1.08              | 0.87                   |
| 8th            | 1.01              | 0.90                   |
| 10th           | 1.00              | 0.91                   |
| 12th           | 1.00              | 0.93                   |
| 17th           | 0.98              | 0.97                   |
| 18th           | 0.98              | 0.97                   |
| 19th           | 0.97              | 0.98                   |
| Top            | 0.93              | 1.02                   |

Source: Table 3.5, Frenette, Green, and Picot (2006).

in the SCF is about 7 percent lower than that found in the census. We tried to investigate the sources of these discrepancies and came to the conclusion that “a likely explanation for the discrepancy between the SCF and the other data sources is relative under-coverage at the very bottom of the income distribution in SCF (and SLID)” (Frenette, Green, and Picot 2006, 89).

#### THE HIDDEN UBIQUITY OF THE CENSUS

The public debate on the census has made clear the multitude of direct uses for data coming from the census long form. For example, housing information is used by the Canadian Mortgage and Housing Corporation to fulfill its legislative mandate, and also by local governments and private sector actors to learn about trends in housing. Information on languages is used to determine local linguistic service levels—again both by governments and others in Canadian society. Less attention has been focused on the more indirect—yet in ways even

more vital—role the census plays in the Canadian statistical system.

To make this discussion more concrete, we take the example of the Labour Force Survey (LFS). The LFS is a monthly survey used to obtain key information for constructing unemployment, employment, and participation rates. It is the basis for all the unemployment rate statistics reported by Statistics Canada, making it the basis for many policy considerations. For example, the unemployment rates constructed from it are also used at the regional level as part of calculating eligibility for Employment Insurance. Beyond its direct uses, the LFS is made even more important because it is used as a starting point for a large proportion of voluntary surveys conducted by Statistics Canada. For example, the SLID, mentioned earlier as the primary source for labour and income data, is based on a sample drawn from the LFS.

The creation of the LFS sample starts with choosing a sampling frame.<sup>4</sup> A stratified sampling design is used, which means that the survey designers first cut up Canada into strata or geographically and demographically defined groups. Clusters within each stratum are then chosen, and random dwellings within these chosen clusters are contacted. Both the definition of the strata and clusters and the sampling rates within clusters are determined using data from the most recent census. Importantly, some “special strata” are used to target specific populations of interest. These populations include aboriginal peoples, immigrants, and those with high income. As our discussion in the previous sections indicated, to obtain estimates for these groups that have low variance (i.e., to take advantage of the power of the Law of Large Numbers), we need sample sizes that are as large as possible. If the group of interest is small within the population, then simple random sampling will imply small numbers of sample members for that group. In response, Statistics Canada uses long form census counts to determine which geographic areas have large concentrations of the

targeted group (e.g., immigrants), and those areas are then over-sampled.

Beyond the stratification, once survey responses are collected, a further correction for non-response bias is made. Households are effectively sorted into groups by certain demographic characteristics. The proportion of responders in these groups is compared to a benchmark, and corrections are calculated so that the responses from the survey can reflect the whole population. This approach is called “weighting” and can be understood with a simple example. Imagine that all native born Canadians responded to a survey, but only half of the immigrants contacted did so. We could recreate a picture of the overall population using this data if we counted each immigrant as if he or she were two people (giving them a weight of two) while counting each native born person as one person (giving them a weight of one).

In the LFS, an initial set of weights is created for groups who are known (based on comparisons to census data) to respond to the survey at low rates. These groups include people in remote rural locations, aboriginals, and high-income households. There is also a final exercise to generate weights that ensure that weighted counts in the LFS for groups defined by age, gender, and province match those in the most recent census data. The latter weights can be generated using just short form census information, but the other weights (relating, for example, to aboriginals) are based on the long form.

All voluntary Statistics Canada surveys come with a set of weights of this type that researchers need to use to obtain accurate statistics. But constructing those weights requires having a “true” population benchmark, and the census is that benchmark. Thus, without the census, both the stratification and weighting stages of all other surveys would be affected. For the LFS this would mean inferior statistics on unemployment and employment. Beyond the set of surveys collected by Statistics Canada, privately collected (e.g., by polling firms)

surveys must also be compared to some standard to ensure they are providing unbiased statistics. Comparing them to some other voluntary survey (such as the NHS) which has its own, unknown, response biases is obviously of limited usefulness. Thus, to ensure the quality of these surveys, the mandatory census short and long forms are important.

Finally, as pointed out in Veall (2010), the extent to which reweighting can provide a useful solution is based on whether the census contains information on the full set of variables that affected the response decision. With reweighting, we assume that we can make up for any non-response biases by, for example, counting each immigrant as if he or she is two people. But if non-responding immigrants are fundamentally different from immigrants who respond to a survey in ways that relate to measures of interest (e.g., if non-respondent immigrants are out of the house working when survey-takers come around, and so are likely higher income than respondents), then even a carefully reweighted survey will not give a completely accurate estimate of the population. Only a census, which generates a random sample of the Canadian population, will provide the needed accuracy.

## CONCLUSION

The census is a vital, even pivotal component of our statistical infrastructure. If the government announced ill-advised technical changes to the power grid or road system, few Canadians might notice or care initially. It is only when their electricity blacks out, or a bridge fails, that the folly of the changes becomes clear. Similarly, the degradation of the Canadian census has impacts that, while perhaps not immediately clear to Canadians, will eventually have large influences on the quality of Canadian society.

We have argued in particular that tampering with the census has impacts on a long list of other

data in Canada. It will have negative impacts on the many economic decisions based on statistics (e.g., whether and where to establish a new store). It also has potentially important effects on what we might call coordination efficiencies. For example, economic agents making contracts will base their arguments in their bargaining, ultimately, on statistical data. With trusted sources of information, both parties to a bargain can rely on the same data. Once biases become evident, however, we would expect agents to challenge the validity of numbers that are not advantageous to their position and, possibly, to shop around for numbers they like better. Without an ultimate benchmark for the statistical data generated either by private firms or by Statistics Canada itself, there would be no way to arbitrate such disputes, and basic contract-making in our economy would suffer. It is probably the recognition that impartially provided statistics are so important for the functioning of modern economies that has led business groups to weigh into the census debate on the side of preserving the long form.

There can be tensions at times in our democracy between expert technical advice and populist instincts. We feel that an important task of experts is to inform those populist instincts with reasoned arguments and facts. Ultimately, we have faith in the good judgment of Canadians. We can only hope their elected representatives are listening.

## NOTES

<sup>1</sup> More precisely, the mean of sample means would equal the population mean in a circumstance with a finite sample and sampling without replacement.

<sup>2</sup> If we sort the sample by income from lowest to highest and divide it into 20 equally sized groups, each such group is called a *vingtile*.

<sup>3</sup> As a side point, our investigations in these papers focus on household income, and there is some evidence that attempts to impute families in the tax data have been problematic at times. For that reason, we favour census over

the tax data. The tax data also have issues related to incentives not to report income that are not present in the census.

<sup>4</sup>This discussion follows the description in Statistics Canada (2008).

## REFERENCES

- Frenette, M., D.A. Green, and G. Picot 2004. "Rising Income Inequality in the 1990s: An Exploration of Three Data Sources." Statistics Canada, Analytical Studies Research Paper Series, No. 219.
- Frenette, M., D.A. Green, and G. Picot. 2006. "Rising Income Inequality in the 1990s: An Exploration of Three Data Sources." In *Dimensions of Inequality in Canada*, ed. D.A. Green and J.R. Kesselman, 65-100. Vancouver: University of British Columbia Press.
- Frenette, M., D.A. Green and K. Milligan. 2007. "The Tale of the Tails: Canadian Income Inequality in the 1980s and 1990s." *Canadian Journal of Economics* 40(3):734-64.
- Income Statistics Division, Statistics Canada. 1997–2008. "User Guide for the Public-Use Microdata File Survey of Household Spending." Cat. No. 62M0004XCB. Ottawa: Statistics Canada.
- Statistics Canada. 2008. "Methodology of the Canadian Labour Force Survey." Cat. No. 71-526-GIE. Accessed at <http://www.statcan.gc.ca/pub/71-526-x/71-526-x2007001-eng.pdf>
- Veall, M. 2010. "2B or Not 2B? What Should Have Happened with the Canadian Long Form Census? What Should Happen Now?" *Canadian Public Policy/Analyse de politiques* 36(3):395-9.