

# Designing Coverage Studies for the 2001 Canadian Census

Colleen Clark and Jocelyn Tourigny  
Statistics Canada

## ABSTRACT

The 1996 Canadian Census missed 3.18% of the persons that it should have enumerated while 0.74% of the persons actually enumerated were enumerated in error. The methodology for measuring population undercoverage has varied little since the first study was conducted for the 1961 Census. A sample of persons who should have been enumerated is taken from a frame constructed from the last Census and administrative data. These persons are contacted to obtain all possible addresses where they could have been enumerated and the Census documents for these addresses are verified to determine what enumeration occurred. The 1996 coverage studies introduced a more comprehensive methodology for estimating overcoverage and a more efficient Reverse Record Check design that used demographic stratification and a one-stage design. The success of the 1996 approach sets the stage for no major changes for 2001. As analysis of the 1996 coverage studies completes and development of the 2001 studies commences, some themes are emerging. First, the RRC can provide reliable estimates of components of intercensal population growth. Second, better information is required to differentiate temporary from permanent international migration. Last, the methodology for measuring overcoverage can be improved by initiatives such as exploiting automated matching methods.

KEY WORDS: Census coverage error, Undercoverage, Overcoverage, Net undercoverage, Census non-sampling error

## 1. Introduction

The 1996 Census required the participation of the entire population of Canada, some 29 million people distributed over a territory of 9.2 million square kilometers. Although there are high quality standards governing the gathering and processing of the data, it is not possible to eliminate all errors. Undercoverage occurs when persons, households, or dwellings are missed by the Census while overcoverage occurs when persons, households, or dwellings are enumerated in error. The 1996 Census missed 3.18% of the persons that it should have enumerated while 0.74% of the persons actually enumerated by the Census were enumerated in error. On a net basis, the Census missed 2.45% of the persons (723,486 persons) that it should have enumerated. The 1996 Census household net undercoverage rate was 1.89% (209,143 households).

The 1996 undercoverage rate follows historical levels. The national undercoverage rate was close to 2% for 1971, 1976 and 1981, but then rose to 3.21% in 1986. The increase is thought to reflect both an increase in the construction of dwellings that are difficult to enumerate such as renovated inner-city homes, and a change in the public mood towards government. As a result of the increase in 1986, coverage improvement initiatives were introduced for the 1991 Census. In particular, the use of the Address Register to provide a separate list of dwellings which should be enumerated helped to keep undercoverage, at 3.43%, near the 1986 level. For the 1996 Census, the addition of enumeration by a Census enumerator in some Enumeration Areas in large cities, rather than self-enumeration, served to control undercoverage.

The error resulting from the use of the published Census population count,  $C$ , instead of the true number of persons,  $T$ , is  $N$ , the net coverage error. That is,  $N = T - C$ . If  $U$  is the number of persons missed in the Census and  $O$  is the number of persons enumerated more than once or in error, then  $N = U - O$ . Population coverage error is usually expressed as a proportion of the total number of persons who should have been enumerated in the Census: Undercoverage Rate  $R_U = U/T$ , Overcoverage Rate  $R_O = O/T$ , and Net Undercoverage Rate  $R_N = R_U - R_O$ .

An extensive survey program measures Census coverage error. The «reverse record check» methodology for measuring undercoverage has not varied fundamentally since its inception for the 1961 Census. Essentially, a sample of persons who should have been enumerated is taken from a frame constructed from the last Census and administrative data. These persons are contacted to obtain all possible addresses where they could have been enumerated and the Census documents for these addresses are verified to determine if enumeration occurred.

The 1996 Reverse Record Check (RRC) design was an improvement over the 1991 design when a two-stage design with largely geographic stratification was replaced by a more efficient one-stage design with demographic stratification. Also, for the first time, the measurement of overcoverage was integrated into the RRC so that the number of times that a sampled person was enumerated at the address(es) he/she provided was also determined. Last, new automated matching techniques applied to the Census database allowed the identification of a substantial part of overcoverage.

The success of the 1996 Census coverage studies has set the stage for very little change to the design of the coverage studies for the 2001 Census. Besides even further exploitation of automated matching for detecting overcoverage, the design of the RRC will be stretched to improve the RRC estimate of overcoverage and to provide better estimates of the demographic components of intercensal population change. These goals will be accomplished without any compromise to the quality of the provincial estimates of net undercoverage.

The following section presents the methodology of the 1996 coverage studies. Section 3 gives the estimated 1996 Census coverage error while Section 4 describes the evaluation of the RRC estimates. Section 5 lists the changes that are under consideration for the 2001 coverage study.

## 2. Design of the 1996 Coverage Studies

The 1996 coverage studies measured population and household undercoverage and overcoverage as well as classification errors involving unoccupied private dwellings.. The Vacancy Check (VC) estimated undercoverage from occupied dwellings misclassified as vacant. VC estimates were included in the final Census counts to account for this type of undercoverage. Undercoverage from all sources was measured by the largest study, the Reverse Record Check (RRC). Overcoverage was measured by the RRC, the Automated Match Study (AMS), and the Collective Dwelling Study (CDS). The AMS focused on persons counted more than once within the same region (Atlantic, Ontario, Quebec, rest of Canada) while the CDS estimated overcoverage resulting from persons enumerated as usual residents in a collective dwelling who were also enumerated at a private dwelling. Although the RRC provides a measure of total overcoverage, only the overcoverage exclusive of the AMS and CDS was used for estimating coverage error since the latter studies provided more reliable estimates.

**1996 Census Coverage Studies**

Study	Sample Size	Measures
Vacancy Check (VC)	1,396 Enumeration Areas	Undercoverage from occupied dwellings misclassified as vacant.
Reverse Record Check (RRC)	57,016 persons	Undercoverage from all sources. Overcoverage not included in the AMS or CDS.
Automated Match Study (AMS)	7,688 pairs of households	Overcoverage from persons enumerated in two households.
Collective Dwelling Study (CDS)	12,561 persons	Overcoverage from persons enumerated in both a collective and a private dwelling.

Estimates of coverage error are of critical importance in setting the level of federal transfer payments to the provincial governments. This has led to a high degree of user interest and scrutiny. Periodic workshops were held with representatives from the provincial and territorial governments to share methodological details and solicit input on methodological options. The financial importance of the net undercoverage estimates meant that small changes in the estimates, relative to standard error, were of concern to users and required thorough investigation.

### 2.1 Reverse Record Check

*Frame*

Several data sources were used to create a list of all persons who should have been enumerated by the 1996 Census. The sample for the ten provinces was drawn from a multiple frame constructed from five sources. The 1991 Census Frame is a list of all persons who were enumerated in the 1991 Census. The Missed Frame consisted of persons classified as missed by the 1991 RRC. The remaining frames, based on administrative data, accounted for change in the target population since 1991. The Birth Frame listed all persons born between June 4, 1991 and May 13, 1996. The Immigrant Frame was a list of all intercensal immigrants while the Non-permanent Residents (NPR) Frame contained all persons with a valid permit to reside temporarily in Canada on May 14, 1996. The sample for the two territories was drawn from the administrative files for each territory's health care plan. Sampled persons who were found to be covered by another frame were considered as out of scope at estimation.

### *Stratification, Allocation, and Sample Selection*

Each of the frames was stratified by province/territory and then, with the exception of the Missed Frame, further stratified by mainly demographic variables specific to the frame such as age, sex, and year of birth. The sample of 50,141 persons was distributed among the provinces so as to ensure that the standard error of the estimate of net undercoverage in each province would not exceed a specified limit (0.30) with the remainder allocated proportional to population. Last, the change in sample size from 1991 was limited to 10%. Within the provinces, the sample was allocated to the strata so as to maximize the precision of the provincial estimate of net undercoverage. This means that more sample was allocated to the strata where survey non-response and/or Census coverage error were expected to be high. Within each of the two territories, the remaining sample (6,875 persons) was allocated proportional to population. Within all strata, simple random sampling was used to select the sample.

**Provincial Characteristics of the 1996 RRC Sample**

<b>Province/ territory</b>	<b>Sample Size <i>persons</i></b>	<b>Sampling Fraction <i>percent</i></b>	<b>Actual Standard Error Estimated Net Undercoverage Rate <i>percent</i></b>	<b>Not Traced <i>percent</i></b>	<b>Total Non- response <i>percent</i></b>
Newfoundland	3,343	0.54	0.31	0.69	1.62
PEI	3,113	2.19	0.32	0.35	1.00
Nova Scotia	3,403	0.34	0.27	1.00	2.50
New Brunswick	3,408	0.43	0.31	0.70	2.11
Quebec	8,236	0.11	0.20	1.57	3.15
Ontario	11,378	0.09	0.19	3.86	5.81
Manitoba	3,405	0.28	0.34	1.67	2.79
Saskatchewan	3,379	0.31	0.34	1.15	1.95
Alberta	4,569	0.16	0.27	3.39	4.57
British Columbia	7,443	0.19	0.25	5.58	8.04
Yukon	2,292	6.58	0.51	3.27	5.15
NWT	3,047	4.04	0.51	1.05	1.21
Canada	57,016	0.18	0.10	2.51	4.12

### *Collection, Processing, and Classification*

The first stage of collection was tracing. Frame information (name, sex, birth date, address) on the Selected Person (SP) and his/her household members was sent to the field with more recent address information from taxation records when available. Interviewers may have consulted additional sources (motor vehicle registrations, directory assistance, Canada Post, etc.) to obtain a current address. Once the SP or an acceptable proxy respondent was contacted, the interviewer collected the exact address of the SP on Census Day, the names and demographic characteristics of all persons who

were living with the SP on Census Day, all the addresses where the SP could have been enumerated, and some basic Census characteristics. Census documents for each address were located and manually verified to determine what enumeration had occurred. Finally, each SP was classified as enumerated (at least once), missed, deceased, emigrated, abroad, or out of scope (not in the Census target population, or covered by another frame). Non-response occurred when there was not enough information to derive a classification. Selection information was insufficient to commence tracing for the «Not Identified». Tracing was attempted, but the «Not Traced» SPs could not be located. The information collected for the «Not Classified» SPs did not permit a final classification, usually because of vague addresses.

**Final Classification of RRC Sample by Frame**

Final				Provinces		Territories	Total
	Census	Births	Immigrants	91	NPR	HCF	
Enumerated	37,608	3,117	1,726	1,510	716	4,521	49,198
Missed	1,347	86	184	212	209	254	2,292
Deceased	1,348	22	7	60	1	11	1,449
Emigrated	271	18	72	33	0	6	400
Abroad	132	18	50	20	0	14	234
Out of Scope	59	6	351	349	251	135	1,151
Not Identified	400	13	0	0	27	0	440
Not Traced	686	97	193	125	224	107	1,432
Not Classified	214	13	22	32	37	102	420
Total	42,065	3,390	2,605	2,341	1,465	5,150	57,016

### Estimation

The design weights were subject to two adjustments to arrive at a final weight for each respondent. That is,  $w_f = w_0 * f_{NI} * f_{NT} * f_{NC} * f_{PS}$  where  $w_0$  = inverse of probability of selection,  $f_{NI}$ ,  $f_{NT}$ , and  $f_{NC}$  are adjustments for each type of non-response, and  $f_{PS}$  is an adjustment for controlling to external data. The weight of the non-respondents was shared among the respondents in five independent stages. «Probable mobility» played an important role in determining which respondents would have their weight augmented. It was assumed that an SP who had moved had a higher chance of being missed than someone who had not moved. The remaining estimation stage was a series of post-stratification steps. First, an adjustment was made to compensate for incomplete coverage of the Health Care Files. Second, an adjustment was made to incorporate better frame counts for non-permanent residents. Last, a post-stratification of the 1991 Census sample by age and sex to the 1996 counts of enumerated persons was done to reduce the impact of persons selected with incorrect age and/or gender.

### 2.2 Automated Match Study

The Automated Match Study consisted of searching the Census database for pairs of households containing at least two persons with the same gender and date of birth characteristics in the same geographic region (Atlantic, Quebec, Ontario, rest of Canada). All pairs of households were then grouped according to the number of persons in common and the geographical proximity of the two households. A sample of household pairs was then selected from each group and the Census questionnaire for each of the two households was verified to determine the exact number of overcovered persons. This process allowed a very accurate estimation of a large part of the overcoverage contained in the Census database.

### 2.3 Collective Dwelling Study

The Collective Dwelling Study consisted of two parts. The first part covered institutional collective dwellings such as hospitals, retirement homes and prisons, while the second part covered non-institutional collective dwellings such as

hotels and student residences. In the 1996 Census, all the usual residents of collective dwellings were asked to provide an alternate address. A sample of residents was drawn for each of the two categories of collective dwellings and the alternate address was verified to see if the person was enumerated there. From this verification, an estimate of overcoverage from persons enumerated at both a collective and a private dwelling was produced.

### 3. 1996 Census Coverage Error

Apart from the northern territories, undercoverage is largest on the western coast in British Columbia, reflecting recent population increase. A rate slightly larger than the national rate in central Ontario could reflect the draw of a healthy economy while lower rates in the east reflect less residence mobility. There is greater variation in undercoverage among the gender and age groups with generally higher undercoverage for men than for women and the highest rates for young adults. The rates are strikingly high for those aged 20-24; 9.48% for males and 6.45% for females. Nationally, slightly more persons were missed in a missed household (53%) than in an enumerated household (43%), with the remainder missed in a collective dwelling (4%).

**Estimated 1996 Census Coverage Error by Age and Gender**

Age	Undercoverage				Overcoverage			
	Males		Females		Males		Females	
	Rate	Standard Error	Rate	Standard Error	Rate	Standard Error	Rate	Standard Error
0-4	2.56	0.47	3.24	0.55	0.52	0.09	0.69	0.18
5-14	1.46	0.24	1.45	0.22	0.99	0.15	0.92	0.14
15-19	3.68	0.43	3.28	0.55	1.12	0.24	1.36	0.29
20-24	9.48	0.50	6.45	0.48	2.34	0.34	2.55	0.46
25-34	7.74	0.42	3.84	0.40	0.65	0.11	0.66	0.11
35-44	3.94	0.39	1.62	0.28	0.38	0.06	0.37	0.10
45-54	2.12	0.27	1.68	0.33	0.35	0.07	0.61	0.20
55-64	2.50	0.54	1.97	0.40	0.37	0.12	0.66	0.19
65+	1.64	0.45	1.43	0.32	0.33	0.02	0.38	0.11
Total	3.89	0.14	2.49	0.12	0.70	0.04	0.77	0.06

The Automated Match Study measured overcoverage from persons enumerated in two households in the same region (93,688 persons). The Collective Dwelling Study measured overcoverage from persons enumerated in a collective dwelling and a private dwelling (8,467 persons). The RRC measured single persons enumerated in different dwellings; persons and households enumerated in different dwellings but who reported different gender and date of birth; and persons and households enumerated in different regions (122,406 persons).

As for undercoverage, overcoverage is highest for young adults aged 20-24, again reflecting the higher degree of residence mobility. Overcoverage, however, is generally higher for females than for males. Apart from those aged 20-24, overcoverage is concentrated in children and youths aged 5-19 for both sexes. This phenomena reflects children and youths with parents who do not reside in the same household, as well as those who were enumerated more than once because their families moved around Census Day.

### 4. Evaluation of the 1996 Coverage Studies

In spite of some conceptual differences between the RRC and the 1996 Census, three comparisons are instructive; enumerated, total of immigrants and non-permanent residents, and interprovincial migration. Since the RRC one-stage stratified design results in unbiased estimators, differences between RRC estimates and estimates from the Census are due to sampling error on the part of the RRC estimates, conceptual differences between the two sources, and/or biases in the two sources which result in a systematic underestimation or overestimation.

In order to compare the Census count of enumerated persons and the RRC estimate, reasonable assumptions about the magnitude of the conceptual differences between the two sources can be applied. Nationally, the RRC estimate of persons enumerated in the 1996 Census falls marginally short, 0.08%, of the comparable 1996 Census figure. The chi-squared test statistic to test if at least one province has a difference is 13.3, which is not significant at the 90% level (critical value of 14.9). The gaps for some provinces are of some concern since they may indicate a bias in the RRC classification. Apart from sampling error, the gap is explained by biases in the adjustments for conceptual equivalence with the Census, RRC non-response bias as the adjustment methodology is chosen for estimating missed persons, and undetected frame overlap between the Immigrant Frame, the NPR Frame and the 1991 Census Frame.

Immigrants and non-permanent residents (NPRs) are of particular interest since they have considerably higher rates of undercoverage than the general population. Nationally, the RRC overestimates the Census count by 0.77%. Among the provinces with the highest concentration of immigrants and non-permanent residents, the RRC underestimates for British Columbia, by 3.69%, and overestimates for both Quebec (5.23%) and Ontario (2.15%). The chi-squared test statistic to test if at least one province has a difference is 9.5 indicating that none of the provincial differences are statistically significant.

In general, the RRC overestimates both intercensal interprovincial 5-year «in migration» and «out migration». The difference is striking for the total number of migrants where the RRC overestimates the Census count by almost three times the standard error. This result likely reflects the weakness of the Census recall approach whereas the RRC uses the actual province of residence in 1991 as recorded in the 1991 Census database. The chi-squared test statistic for testing if at least one of the provincial differences is significant, is 16.3 for in migration, close to significance at the 95% level (16.9). For out migration, the test statistics is 18.7, indicating significance at the 95% level. On a net basis, the RRC tends to slightly underestimate net migration. The chi-squared test statistic indicates that none of the net migration provincial differences are statistically significant.

RRC estimates of the number of intercensal deaths are consistently larger than counts from provincial vital statistics (VS) counts. At the national level, the RRC overestimates the comparable VS count by 59,831 (5.9%). In order to render the two sources comparable, two adjustments were made. First, the RRC estimate of persons who died outside the country and could not be found on the vital statistics files was added to the VS count. Second, the RRC estimate was adjusted for overcoverage in the 1991 Census Frame. The chi-squared test statistic for testing if at least one of the provincial differences is significant is 11.1 indicating no statistical significance at the 90% level (14.7). The RRC overestimation of deaths is of some concern because the VS counts are considered to be extremely accurate. Although some bias hypotheses have been investigated, the research does not show conclusive evidence of bias.

#### Comparison of 1996 RRC Estimates With Census and Vital Statistics

Province/ territory	Enumerated		Deceased	
	RRC less Comparable Census	RRC Standard Error	RRC <sup>1</sup> less Comparable Vital Statistics	RRC <sup>2</sup> Standard Error
Newfoundland	-898	5,170	-902	1,766
PEI	-2,657	2,462	1,191	1,155
Nova Scotia	873	9,452	5,882	3,309
New Brunswick	4,499	7,919	307	3,779
Quebec	51,486	29,160	17,487	18,262
Ontario	18,929	51,275	19,345	26,940
Manitoba	-18,209	10,404	9,283	4,753
Saskatchewan	-3,933	10,194	-1,383	3,303
Alberta	-32,752	21,613	-1,754	8,298

British Columbia	-40,158	22,994	10,374	9,620
Yukon	-183	0	-	-
NWT	-718	0	-	-
Canada	-23,723	58,647	59,831	35,830

<sup>1</sup> Adjusted to account for overcoverage in 1991 Census Frame and exclude.

<sup>2</sup> Standard error is for estimates that are not adjusted to exclude overcoverage in 1991 Census Frame.

## 5. Design of the 2001 Coverage Studies

The main goal of the Reverse Record Check continues to be to provide a measure of undercoverage. Any secondary objectives will be attained without compromise to the quality of the provincial estimates of net undercoverage. The RRC design will not change in any substantial fashion for 2001. The allocation will be updated according to the 1996 results and population growth. The following changes are under study:

### *Overcoverage*

First, the small samples for the RRC contribution to overcoverage mean that individual cases have a large impact on provincial estimates of net undercoverage. The use of the national rate of the RRC contribution to overcoverage for all provinces is being considered. Second, comparison of AMS and RRC results demonstrate that the RRC does not measure all sources of overcoverage since it is restricted to those households identified by the respondent. Research will be conducted to exploit the AMS approach to detect even more overcoverage. Third, during the course of processing the address for an SP, additional cases of overcoverage (e.g.: other members of the collection household) are found, but these are not used in estimation. Research is required to develop an estimation methodology that allows such cases to contribute to the RRC estimate of overcoverage. Last, a new survey similar to the Vacancy Check is planned to investigate overcoverage resulting from the Census field procedure used to impute household and person characteristics for households that are absent for the duration of the collection period. Overcoverage is suspected as the imputation does not take account of the possibility that the household members could be temporary residents or that the dwelling is vacant.

### *Undercoverage*

First, an additional question asking why the SP is away from Canada would give better information for distinguishing temporary from permanent international migration. Second, the focus in the RRC has been to determine whether or not the SP has been enumerated and no attention is given to establishing where the SP should have been enumerated. The difference between where a person should have been enumerated and where they were actually enumerated is relevant when the difference crosses provincial boundaries. It is proposed to establish the province of usual residence by doing a follow-up for those SPs where the province of their usual residence according to the RRC questionnaire differs from the province of enumeration. For 1996, this would have meant following up about 200 more cases. Third, the lower sampling fractions in Quebec and Ontario mean that migrant SPs can have a substantial impact on provincial estimates. Adding about 5,000 SPs to the Ontario and Quebec samples will decrease the impact of such migrant SPs.

### *Reverse Record Check*

First, the number of strata in the 1991 design may be too high for the 1996 design. An examination of design effects and net undercoverage rates at the stratum level is required. Second, with numerous longitudinal household surveys now underway, Statistics Canada has gained a lot of experience in tracing since the 1996 RRC. This experience should result in higher tracing rates, reduced effort, and quicker tracing. Other tracing issues include improving tracing across provincial boundaries. Third, the weight adjustment methodology should recognize that the probability of tracing an SP varies across provinces according to the quality and accessibility of administrative data. Also, research is required to ascertain if «probable mobility» as used for the 1996 RRC is the best indicator of mobility. Last, the design weights for the sample from the 1991 Census Frame are marginally too large in that frame overcoverage is not removed before the

sample is selected. A weight adjustment procedure is required to deflate the design weights accordingly.

### *RRC As A Population Survey*

Kerr and Lachapelle (1999) documents an extensive study comparing RRC estimates of the intercensal components of population growth and estimates from demographic sources. Their work set the stage for recent research into a composite estimator for provincial population estimates. See Dick (1999). The composite estimator uses the components of population growth as measured by the Reverse Record Check. Under this scenario, the RRC estimates of deceased, interprovincial and international migration become an important objective of the RRC. Design changes which improve the RRC estimate of deceased persons with nil or marginal impact on the estimation of coverage error include further stratifying the collective dwellings by type and ensuring that the sample in the elderly stratum reflects the frame age distribution. RRC estimates of international migration are addressed by the additional question already listed to better distinguish temporary and permanent international migration. The determination of the province of residence as already listed would improve the RRC estimates of interprovincial migration. Last, the RRC may provide a mechanism for a sample-based estimate of «returning Canadians», persons who emigrated from Canada before the last Census but then returned before Census Day.

## **6. Conclusion**

Overall, the degree of change expected for the 2001 coverage studies is far less than the changes implemented for the 1996 coverage studies. As development commences, some themes are emerging. First, the RRC can be a useful vehicle for estimating the components of intercensal growth. This is especially important if composite estimators for official population estimates are adopted. Then, the RRC component estimates are part of the population estimate. Second, better information is required for those who were outside Canada on Census Day to distinguish between temporary and permanent absence. Third, there is scope for improving the measurement of overcoverage.

## **Acknowledgments**

The authors acknowledge the contribution of a large number of Statistics Canada staff. Past and current project teams developed and carried out the coverage studies with a high degree of attention to quality. Internal and external users provided invaluable evaluation and input.

## **Bibliography**

Dick, P. (1999). *The Composite Estimator: A Preliminary Report*. Statistics Canada, Social Survey Methods Division, Ottawa.

Lachapelle, R. and Kerr, D. (1999). *Census Coverage Studies: A Demographic Evaluation*. Statistics Canada, Demography Division, Ottawa.

Statistics Canada. *Coverage*. 1991 Census Technical Reports; Reference Products Series. Ottawa: Minister of Industry, Science and Technology, 1994. Catalogue number 92-341E

Tourigny, J., Clark, C., and Provost, M. (1998). *Evaluation of the March 1998 Preliminary Results of the 1996 Census Coverage Studies*. Statistics Canada, Social Survey Methods Division, Ottawa.