Sampling and Weighting Technical Report

Census of Population, 2021

Release date: November 15, 2023

Statistics Canada Statistique Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

Email at infostats@statcan.gc.ca

Telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following numbers:

• Statistical Information Service	1-800-263-1136
• National telecommunications device for the hearing impaired	1-800-363-7629
• Fax line	1-514-283-9350

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under "Contact us" > "Standards of service to the public."

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

An HTML version is also available.
Cette publication est aussi disponible en français.

Release date: November 15, 2023
Catalogue number 98-306-X, issue 2021001
ISBN 978-0-660-47835-7

Table of contents

Introduction

The 2021 Census Program enumerates Canadian households using two main types of questionnaires: the short-form questionnaire and the long-form questionnaire. In 2021, a sample of 25% of Canadian households received a long-form questionnaire, which included the questions from the short-form questionnaire. The other households received the short-form questionnaire.

In addition to the short-form questions, the long-form questionnaire includes a series of questions to paint a full portrait of the Canadian population and households, according to their demographic, social and economic characteristics.

The estimates produced from responses to the questions on both questionnaires are obtained from the entire population via a census. All households that respond to both types of questionnaires contribute to a specific number, such as the population figure for a specific age group.

The estimates produced from responses to at least one question found only on the long-form questionnaire are obtained from the sample. In those cases, only respondent households in the long-form sample contribute to the estimate. The highest level of educational attainment is an example of this type of estimate.

The long-form sample is evenly distributed geographically to ensure a high degree of reliability of the estimates for all areas of the country and to grant the same degree of importance to all geographic units of a given size.

This technical report presents the methodology used to produce the estimates based on the 2021 Census of Population long-form sample. Chapter 1 details the collection methods used for the census and for the long-form sample. Chapter 2 describes how the sampling was applied for the long-form questionnaire. Chapter 3 explains the data processing procedures. Chapter 4 gives an overview of the procedures used to assign weights to the respondent units in the long-form sample to obtain estimates for the population. Chapter 5 covers different evaluations of the weighting procedures, while Chapter 6 provides an overview of the variance estimation methodology used for the 2021 Census long-form sample. Chapter 7 introduces statistical inference using confidence intervals. A conclusion follows.

1. Census data collection

During the data collection phase, the objective was to ensure that responses were obtained from all households in Canada. Field operations included listing dwellings, delivering invitation letters, determining the occupancy status of a dwelling and conducting interviews with non-respondents.

1.1 Census delivery methods

For most private dwellings, respondents were asked to complete the questionnaire for themselves and all members of the household.

On May 3, 2021, all private dwellings in the mail-out (MO) areas (approximately 86% of private dwellings in Canada) received by mail a bilingual invitation letter to complete the questionnaire online. As in 2016, this letter contained a secure access code (SAC), the web address of the 2021 Census website, and a telephone number to allow the respondent to request a paper questionnaire if they preferred.

In list/leave (L/L) areas, which represent 7% of dwellings, census employees dropped off an invitation letter. L/L door-to-door delivery took place from May 3 to May 10, 2021. During the L/L operation, census employees listed all private dwellings in specific areas in their Visitation Record. The invitation letter had a SAC so that respondents could fill out the questionnaire online. Paper questionnaires were available upon request, using a toll-free number. In the L/L areas, it was necessary for the respondent to provide a mailing address to an operator in order for the paper questionnaire to be mailed.

In 2021, the mail-out with drop-off (MODO) methodology was introduced. MODO areas are those where all dwellings have addresses, the majority of which are mailable. In these mixed areas, those dwellings with a valid mailing address were mailed the regular MO material (just like the MO areas), while those that did not have a valid mailing address (that correspond to the civic address) received an invitation letter dropped at their door by a census employee. The MODO areas were introduced to maximize the number of census MO dwellings. MODO areas represent more than 6% of the dwellings, and allowed an increase in the use of the MO methodology to extend to about 90% of dwellings (82% in 2016).

Traditionally, the remaining dwellings, located in First Nations communities, Métis settlements, Inuit regions and other remote areas, are enumerated in person using canvasser methods. However, for the first time, in 2021, all First Nations communities, Métis settlements, Inuit regions and other remote areas were offered the opportunity to self-respond, provided it was operationally feasible (i.e., Internet was accessible in the community). Depending on the situation, the invitation letter of the MO, L/L or MODO methodology was used (with minor changes, e.g., the paper questionnaire option was not offered), followed by non-response follow-up (NRFU). Households in areas where it was not operationally feasible to offer self-response completed their census questionnaire with a census employee (in person or over the phone). In 2021, dwellings in remote, northern and Indigenous communities represent about 1% of dwellings in Canada.

1.2 Census wave approach

Statistics Canada implemented a wave approach for the 2021 Census, which consisted of reminding Canadians to fill out their questionnaire by various contact methods at specific times throughout the collection period. It also encouraged respondents to complete their questionnaire online, while mitigating the risk of a decline in overall response by also offering other response options such as ordering a paper questionnaire. The following table outlines the key dates for the different waves.

Table 1.2.1
Census collection phases and schedule

Collection phase	Main activity	Coverage	Start date
Wave 1—Invitation letter	Dwellings in MO areas received an invitation letter with a SAC.	All MO dwellings, including those in MODO areas (90% of all dwellings)	May 3, 2021
	Dwellings in L/L areas and drop-off dwellings in MODO areas received an invitation letter with a SAC.	All L/L dwellings and drop-off dwellings in MODO areas (9% of all dwellings)	May 3, 2021
Wave 2—Reminder letter or card	Dwellings in MO areas received a reminder letter with a SAC.	All non-responding MO dwellings, including those in MODO areas	May 12, 2021
	Dwellings in L/L areas received a reminder card.	All L/L dwellings	May 12, 2021
Wave 3—Second reminder letter	Dwellings in MO areas received a second reminder letter with a SAC.	All non-responding MO dwellings, including those in MODO areas	May 21, 2021
Reminder message	Dwellings in MO areas received either a text reminder (if a cellphone number was available), a voice broadcast message (if a landline phone number was available) or an email reminder (if an email address was available).	All non-responding MO dwellings, including those in MODO areas	May 30, 2021
Non-response follow-up	NRFU began in L/L areas with telephone calls or in-person visits.	All non-responding L/L dwellings	May 21, 2021
	NRFU began in MO and MODO areas with telephone calls or in-person visits.	All non-responding MO and MODO dwellings	June 2, 2021
Final notice letter	Dwellings in MO areas received a final notice reminder letter with a SAC.	All non-responding MO dwellings, including those in MODO areas	July 13, 2021

L/L = List/leave
MO = Mail-out
MODO = Mail-out with drop-off
NRFU = Non-response follow-up
SAC = Secure access code
Source: Statistics Canada, Census of Population, 2021.

In First Nations communities, Métis settlements, Inuit regions and other remote areas, depending on the situation, an invitation letter was delivered, by mail or in person, followed by non-response follow-up, which started on May 14, 2021. Starting on August 3rd, a reminder letter was also delivered to non-responding households in mail-out (MO) areas. If Internet was not available, questionnaires were completed in-person with a census employee from Statistics Canada starting on May 3, 2021.

Census Help Line

The Census Help Line, a free, nationwide, multilingual service, was available to all respondents. The toll-free number was advertised in all census communications materials.

1.3 Occupancy verification and follow-up activities for the 2021 Census

Apartment occupancy verification—The purpose of apartment occupancy verification (AOV) was to verify the occupancy status of all units in an apartment building through one management contact. The information was collected through a telephone interview with the contact person. This contact person could be the owner of the building, or the superintendent or the building manager, for instance. AOV is an important activity, because it helped produce a more accurate status of occupancy for these types of dwellings, and it reduced the workload of the census non-response follow-up (NRFU) activity. AOV was conducted by Collection Support Unit operators from May 10 to 18, 2021.

Dwelling occupancy verification—For a sample of dwellings in mail-out (MO) areas, the status of occupancy was verified immediately before NRFU. Dwelling occupancy verification was conducted from May 21 to 28, 2021, to identify as many unoccupied or cancelled dwellings as possible close to Census Day, May 11, 2021, to remove these dwellings from the NRFU workload. The accuracy of the occupancy status is higher if identified closer to Census Day. This operation is independent from the AOV described above.

Non-response follow-up—The purpose of NRFU was to obtain a completed questionnaire from all households that did not return a questionnaire. Follow-up was done via telephone or in-person visits. In list/leave areas, follow-up was conducted from May 21 to August 13, 2021, and in the mail-out and mail-out with drop-off areas from June 2 to August 13, 2021. In canvasser and reserve areas, NRFU was conducted from May 14 to September 24, 2021. If Internet was not available, questionnaires were completed in-person with a census employee starting on May 3, 2021.

1.4 Census of Population questionnaires

The majority of Canada's population resides in private dwellings. For residents of private dwellings, census data are collected primarily by having one adult member of the household respond on behalf of the entire household through self-enumeration using an online questionnaire.

The census is the primary source of exhaustive demographic data in Canada. In 2021, the census questionnaires collected the following information:

Information collected from both short-form and long-form questionnaires:

- address
- names of usual residents
- date of birth, age
- sex at birth, gender
- relationships of household members (including marital or common-law status)
- knowledge of official languages
- languages spoken regularly at home and language spoken most often at home
- first language learned at home in childhood
- instruction in the minority official language
- Canadian military experience

Information collected from the long-form questionnaire only:

- activities of daily living
- place of birth of person/parents
- citizenship
- knowledge of non-official languages
- ethnic or cultural origins

- First Nations, Métis or Inuk (Inuit) identity
- population groups
- Registered or Treaty Indian status
- membership in a First Nation or Indian band
- membership in a Métis organization or Settlement
- enrolment under, or beneficiary of, an Inuit land claims agreement
- religion
- mobility (one year and five years)
- education
- labour market activities
- language of work
- place of work and commuting
- expenditures (child care, child and spousal support)
- housing.

Most census data were collected using either the short-form or long-form questionnaires. In 2021, a sample of 25% of Canadian households received a long-form questionnaire.

1.4.1 Short-form questionnaire (forms 2A, 3A and 2C)

Form 2A:

This is the short-form questionnaire that is used to enumerate all usual residents of all private dwellings.

Form 3A:

This is the short-form questionnaire for individuals (similar to Form 2A), which is used to enumerate one person. It is delivered to usual residents in private dwellings who wish to be enumerated separately from other members of the household (e.g., roommates, boarders). It is also used to enumerate residents in some collective dwellings such as lodging and rooming houses for example.

Form 2C:

This is the short-form questionnaire for people living abroad (similar to Form 2A), which is used to enumerate residents who are temporarily overseas at the time of the census. For 2021, this includes Canadian government employees (federal and provincial) and their families, and members of the Canadian Armed Forces and their families.

1.4.2 Long-form questionnaire (forms 2A-L and 2A-R)

The long-form questionnaire complements the short-form questionnaire and is designed to provide more detailed information on people in Canada according to their demographic, social and economic characteristics.

Form 2A-L:

This is the most commonly used long-form questionnaire.

Form 2A-R:

This questionnaire is similar to Form 2A-L but is used in remote, northern and Indigenous communities only. It contains the questions from the long-form questionnaire with examples adapted for First Nations communities, Métis settlements, Inuit regions and other remote areas, as well as two additional questions on band housing. For 2021, there is a new question on band housing fees.

1.5 Collection response rate

The overall collection response rate for the 2021 Census of Population was 98.0%. This rate was calculated directly from the collection results, i.e., before data processing and data quality verification were completed. It represents the number of private dwellings for which a completed questionnaire was returned, divided by the number of private dwellings that enumerators coded as being occupied. The collection response rate for the long-form sample was 97.4% (for more information, see the 2021 Census of Population collection response rates).

2. Sampling[1]

When a sample survey is conducted, the sample selection must be planned properly. In sampling, a subset of the survey's target population is selected to receive the questionnaire. The responses of the subset are used to draw inferences for the entire population. Two types of sampling exist: probability sampling and non-probability sampling. Probability sampling is preferable when producing statistical inferences for the entire population is important, since the probability of unit selection can be calculated and the sampling error can be estimated. This chapter discusses the selection of the sample that received the 2021 Census long-form questionnaire.

2.1 Long-form sample universe

The census household universe was broken down into three parts: private households, collective households and households outside Canada. The long-form sample universe consists only of private households, including those living in private dwellings attached to collective dwellings in Canada. This universe excludes incompletely enumerated reserves and settlements. Unless otherwise specified, the term "in scope" indicates that a household is part of the long-form sample universe (i.e., private households that are not living in incompletely enumerated reserves and settlements). "Out of scope" refers to households not in the universe (i.e., households living in collective dwellings, outside Canada, or in incompletely enumerated reserves and settlements).

2.2 Long-form sampling design

In most cases, the long-form questionnaire was distributed to one-quarter of the households in the long-form universe to gather demographic and socioeconomic data on the Canadian population. The sample was selected from the list of dwellings for the 2021 Census of Population. At the time the sample was selected, the addresses of out-of-scope dwellings were unknown. This meant that some dwellings erroneously received a long-form questionnaire. Once a dwelling was determined to be out of scope, no further collection or processing activities were carried out.

Dwellings were selected to receive the long-form questionnaire according to a stratified systematic sampling design. To define the sampling design and to improve efficiency of field operations, Canada is partitioned in smaller geographical units called collection units (CUs). Each CU is assigned one of the delivery methods described in Chapter 1. The sampling design strata were the CUs. For mail-out, list-leave, and mail-out with drop-off CUs, the sampling was systematic, with a one-quarter sampling fraction. The selection starting point was random. In CUs in First Nations communities, Métis settlements, Inuit regions and other remote areas, all households were selected. These CUs were take-all strata.

The sampling design had one exception. Private dwellings attached to collective dwellings were added to the sample with certainty. However, they completed only the short-form questionnaire. Long-form questionnaire responses were later imputed for these households.

Except private households attached to collective dwellings, all households selected for the sample were asked to complete the long-form census questionnaire. Households in private dwellings that were not part of the long-form sample were asked to fill out the short-form questionnaire.

[1] For more information on the history of sampling in Canadian censuses, see Appendix B.

3. Census data processing

3.1 Introduction

This chapter discusses the processing of all the completed questionnaires (all questionnaire types), which encompasses everything from the receipt of the questionnaires through to the creation of an accurate and complete census database. It describes the steps of questionnaire registration, questionnaire imaging and data capture, editing, error correction, failed edit follow-up, coding, dwelling classification and non-response adjustments, linkage of administrative data, imputation, weighting, and final response rates.

Automated processes, implemented for the 2021 Census, had to be monitored to ensure that all Canadian residences were enumerated once and only once. The Master Control System (MCS) was built to control and monitor the process flow, from collection to data processing. The MCS held a master list of all the dwellings in Canada, where each dwelling was identified with a unique identifier. This system was updated on an ongoing basis with information about each dwelling's status in the census process flow (e.g., delivered, received or processed). Reports were generated daily by the system and made accessible online to managers to ensure that census operations were efficient and effective.

3.2 Receipt and registration

Responses received through the Internet or help-line telephone interviews were received directly at the Data Operations Centre (DOC), where the receipt of the responses was registered automatically.

Respondents completing paper questionnaires mailed them back to the DOC. Canada Post registered their receipt automatically in multiple locations in Canada (as part of the normal mail flow process) by scanning the barcode on the front of the questionnaire through the transparent portion of the return envelope. The envelopes were then delivered to the DOC throughout each business day. Canada Post would also send files daily listing all census questionnaires received at each regional processing plant, by date of receipt.

The registration of each returned questionnaire was flagged on the MCS at Statistics Canada. A list of all the dwellings for which a questionnaire had not been received was generated daily by the MCS and transmitted to field operations to prevent follow-up on households that had already completed their questionnaire during non-response follow-up.

3.3 Scanning and keying from images

In 2021, all paper census forms (2A, 2C, 2A-L, 2A-R, 3A) were imaged. The following steps were part of the imaging process:

- Document preparation: Mailed-back questionnaires were removed from envelopes and foreign objects (i.e., clips and staples) were detached in preparation for scanning. The questionnaires were batched by form type. Their spine was cut off to separate them into single sheets.
- Scanning: The questionnaires were converted to digital images.
- Automated image quality assessment: An automated system analyzed the images for errors or anomalies. Images failing this process were sent to be reviewed by a document analysis operator.
- Document analysis: At this step, images containing anomalies were presented to an operator for review. The operator could accept the image as is and send it directly to key entry (bypassing automated recognition), or the operator could send the entire questionnaire to be pulled at the check-out step. See below for more details on the key entry and check-out steps.
- Automated recognition: This step attempted to automatically recognize all handwritten responses and marks on the questionnaires.
- Key entry: Operators entered responses that automated recognition could not determine with sufficient accuracy. About 12% of all responses were sent to keying.

- Check-out: Once the questionnaires were processed through all of the above steps, the paper questionnaires were checked out of the system. Check-out is a quality assurance process that ensures that the images and captured data are of sufficient quality that the paper questionnaires require no subsequent processing. Questionnaires that had been flagged as containing errors were pulled at check-out and reprocessed.

3.4 Coverage edits, completion edits and failed edit follow-up

At this stage, a number of automated edits were performed on respondent data. These edits were designed to detect cases where the number of persons counted in the household was incorrect because of an error in collection, a respondent error or a data capture error. Most of these errors occurred on paper questionnaires, including:

- data erroneously entered in the wrong person column
- crossed off data that are captured in error
- data not being provided for every household member listed in the roster at the beginning of the questionnaire.

Errors that can occur both on paper and online include:

- data provided for the same person on more than one questionnaire (e.g., a person completes their own 3A questionnaire and is also included on the household 2A questionnaire)
- the receipt of duplicate questionnaires (e.g., a person completes the Internet version and their spouse completes the paper version and mails it back).

For about 54% of edit failures, the system resolved the case automatically. This was done when the error was such that the solution was obvious. The solutions included deleting false person data that were created because of respondent or capture error and deleting duplicate responses. The remainder of the edit failure cases were forwarded to processing clerks for resolution. An interactive system enabled the clerks to compare data across questionnaires and examine the images of paper questionnaires to detect data capture or respondent errors. Edit failures were resolved by deleting invalid or duplicate persons or by adding missing persons (i.e., creating blank person records), as necessary and appropriate.

Following the coverage edits, another set of automated edits was run. These edits detected cases where too many questions had missing responses or where data had not been provided for all the usual residents in the household, including cases where missing persons were added by coverage edit clerks. Households that failed these edits were followed up with. An interviewer called the respondent to resolve coverage issues and obtain missing responses, using a computer-assisted telephone interviewing application. For households that responded to the long-form questionnaire, only data missing for the short-form questions were followed up on. The data obtained through this follow-up activity were introduced into the system for subsequent processing steps. If the follow-up was unsuccessful, the data were imputed in the edit and imputation step (see Section 3.8).

3.5 Coding

The census questionnaires contained questions for which answers could be selected from a list, as well as questions requiring a written response. Where possible, written responses were automatically assigned a numerical code according to Statistics Canada reference files, code sets and standard classifications. Reference files for the automated match process were built using actual responses from past censuses or other surveys measuring the same concepts, as well as administrative files. For cases where a code could not be automatically assigned, codes were assigned using machine learning models that were developed with the natural language processing algorithm “fastText.”[2] Finally, records that were not assigned a code automatically through either a reference file or machine learning were coded by specially trained coders and subject-matter specialists.

The following questions required coding on both the long- and short-form questionnaires:

- gender
- relationship to Person 1
- home language
- mother tongue
- instruction in the minority official language.

The following questions required coding for the long-form sample only:

- place of birth of person
- place of birth of parents
- citizenship
- knowledge of non-official languages
- ethnic or cultural origins
- population group
- religion
- First Nation/Indian band
- place of residence one year ago
- place of residence five years ago
- major field of study
- location of study
- industry
- occupation
- place of work
- Inuit land claim
- main reason for working part-time
- main reason for not working full-year
- Métis organization
- language of work.

A total of about 85 million write-ins were coded from the 2021 Census questionnaires. Overall, about 88% were coded automatically, and about 9% were coded using machine learning, although these rates varied considerably from one question to the next.

[2] Joulin, A., Grave, E., Bojanowski P., and Mikolov T. (2016), “Bag of Tricks for Efficient Text Classification”, *arXiv preprint arXiv:1607.01759v3.*

3.6 Classification and non-response adjustments for unoccupied and non-response dwellings

The Dwelling Classification Survey (DCS) was used to estimate the rate of enumerator error in classifying private dwellings, excluding those in collection units (CUs) in First Nations communities, Métis settlements, Inuit regions and other remote areas, and all private dwellings attached to a collective dwelling, as occupied or unoccupied. This information was used to make adjustments to the census database. The DCS selected a random sample of 1,903 mail-out, list/leave, and mail-out with drop-off CUs. Enumerators revisited these CUs in June, July and August 2021 to reassess the occupancy status as of Census Day of each private dwelling for which no response was received. The DCS estimated that 17.3% of the 1,259,149 private dwellings classified as unoccupied were actually occupied and that 38.5% of the 342,162 private dwellings with no response that were classified as occupied or that had an unknown occupancy status were actually unoccupied. Estimates based on the DCS sample were used to adjust the occupancy status for individual dwellings. This resulted in an increase of 3.0% in the number of occupied private dwellings and a decrease of 6.8% in the number of unoccupied dwellings at the Canada level.

The final non-response status is determined after this adjustment of the occupancy status by the DCS. Occupied private dwellings with non-response had their household size imputed based on the estimated distribution resulting from the DCS and then had the rest of their data imputed. The imputed responses came from another census-responding household or administrative data and were generally the geographically nearest neighbour with the same household size. This process is called whole household imputation (WHI). This imputation process is explained in Sections 3.7 and 3.8.

The WHI process has another component that is separate from the use of the DCS estimates to adjust the census database. The non-DCS areas—CUs in First Nations communities, Métis settlements, Inuit regions and other remote areas, and all private dwellings attached to a collective dwelling—require a different imputation strategy. In these areas only, all unoccupied private dwellings are assumed to be truly unoccupied. This implies that unoccupied dwellings are assumed to be classified correctly and no imputations are done. All private dwellings with no response that were classified by enumerators as being occupied were assumed to be occupied and were imputed as occupied. As in DCS areas, dwellings imputed as occupied had their household size and responses imputed, and the imputed response came from another census-responding household or administrative data. No restrictions were placed on the household size for these imputations, as was done in the DCS area.

The WHI process results in all private dwellings being classified as either occupied or unoccupied (i.e., there is no longer any total non-responding dwellings). At the Canada level (for DCS and non-DCS areas), 3.1% of occupied private dwellings were imputed through the WHI process.

More details on the DCS and the WHI process will be available in the *Coverage Technical Report, Census of Population, 2021*, Statistics Canada Catalogue no. 98-303-X.

3.7 Use of administrative data

The use of administrative data increased for the 2021 Census compared with 2016. In addition to the administrative data used for the Income process, they were used for Immigration, as well as in the context of the WHI process. All these uses benefited from the linkage of administrative data.

Income

As was the case in 2016, administrative data were the only source of information on income for the Census Program. This not only reduced response burden, but also increased the quality and quantity of the income data available. The information on individuals' income was compiled from administrative data for the entire population aged 15 and older. The T1 Income Tax and Benefit Return; the T3, T4, T4A, T4RIF, T4RSP, T5, T4A(P), T4A(OAS), T4E and T5007 tax slips; Canada Child Benefit data; and goods and services tax/harmonized sales tax credit data are examples of the sources of administrative data used. Regular, recurring taxable and non-taxable income received during the 2020 calendar year[3] was included. One-time receipts, such as lump-sum withdrawals from registered retirement savings plans and other savings plans, lump-sum insurance settlements, lump-sum pension benefits, capital gains or losses, inheritances, and lottery winnings, were excluded.

Immigration

The Immigration process is the successor to the 2016 Admission Category[4] process, which also incorporates elements that were in the 2016 Ethnocultural[5] process. For the first time, in 2021, administrative data from Immigration, Refugees and Citizenship Canada (IRCC) were the main source of information for most variables processed in the Immigration process for the census long-form sample. In 2016, respondents were asked their place of birth, citizenship, immigrant status, and year of immigration (if applicable). For 2021, the immigration status and year of immigration questions were replaced by administrative data. In addition to the variables processed in 2016, the IRCC administrative data provided new variables with information on non-permanent residents, year of arrival, province or territory of intended destination and more.

Whole household imputation

During the WHI process, administrative data at the household and person levels were used to impute some non-responding households to improve the data quality of the population and the dwelling counts. Administrative data were used to impute for the household size, date of birth and sex at birth when the administrative data were of sufficient quality.

3.8 Edit and imputation

The data collected in any survey or census contain some omissions or inconsistencies. For example, a respondent may be unwilling to answer a question, answer something that contradicts a previous answer or enter a meaningless answer. Other errors, such as incorrect coding, can also occur.

The final clean-up of data, done in the edit and imputation process, was fully automated using the Canadian Census Edit and Imputation System (CANCEIS) (Statistics Canada 2020) for all census topics. Two imputation methods were applied. The first method, called “deterministic imputation,” involved assigning specific values under certain conditions when problems were clear and unambiguous to resolve. Detailed edit rules were applied to identify these conditions, and the variables involved in the rules were assigned predetermined values. The second method, called “minimum-change nearest-neighbour donor imputation,” applied a series of detailed edit rules that identified any missing or inconsistent responses. When a record with missing or inconsistent responses was identified, another record that met the edit rules and was the most similar to it with respect to a set of defined characteristics was selected as a donor. Data from this donor record were borrowed and used to make the minimum number of changes to the variables to resolve all cases of missing or inconsistent responses.

The edit and imputation process starts with the WHI applied to census non-respondents in CUs with a response rate lower than 90%. For those with good quality administrative data records, these non-respondents have their household size, date of birth and sex at birth imputed from their administrative data for all members of the household as a first step. The remainder of the missing variables are imputed in subsequent steps. The remainder of the census non-respondents are imputed by the geographically nearest neighbour among the set of full or partial respondents, or the set of non-respondents now imputed by administrative data. In the DCS areas, the donor must have the same household size.

[3] The reference year for Income is the calendar year 2020, unless otherwise stated. The 2021 Census is the first census to have linked more than one year of income data. Note that the Weighting process relies on only 2020 income data.

[4] In 2016, Statistics Canada was asked to add Admission Category variables to the census. The data were obtained as a result of a record linkage between the Immigrant Landing File provided by Immigration, Refugees and Citizenship Canada IRCC and 2016 Census data.

[5] The Ethnocultural process comprised five subtopics in 2016: Place of Birth, Citizenship and Immigration, Place of Birth of Parents, Aboriginal, Ethnic Origin, and Visible Minority.

Once WHI is completed, the remainder of the missing or invalid information is imputed deterministically or by nearest neighbour donor imputation, module by module. These modules are built to process all variables of a common topic together.

3.9 Non-response

A non-response status may differ during the collection and processing phases. The main differences arise because the occupancy status can change between collection and processing, and because the household must answer a minimum number of questions to be considered a respondent in the processing phase. Unless otherwise specified, the term “non-response” refers to non-response in the data processing phase. The same applies when response is referred to rather than non-response.

For the 2021 Census long-form questionnaire, two types of households were considered non-respondents:

- households from the sample that answered only the questions common to both types of questionnaires, i.e., only the short-form questions
- households that did not answer any questions.

This refers to total non-response, which is processed differently depending on the collection method and the type of household.

3.10 Weighting

The 2021 Canadian Census Program consisted of a Census of Population and a sample survey for which one-quarter of Canadian private households were selected. Households not sampled for the survey received a short-form questionnaire, while sampled households received a long-form questionnaire. In addition to the short-form questions, the long-form questionnaire gathered sociocultural information, as well as information on daily activities, mobility, place of birth, education, labour market activity, etc. Weighting was used to represent the entire population based on the information gathered from the sample.

The first step in the weighting process was to assign a design weight to each household that reflected its probability of being sampled. In most CUs, the sampling fraction was one-quarter, and therefore, households in these CUs were assigned a design weight of 4. The design weights in these CUs then underwent an initial adjustment for coverage and total non-response. This adjustment was applied to the weights of respondent households. Finally, a second adjustment, referred to as final calibration, was made to establish closer agreement between the estimates obtained from respondent households in the sample and the census counts for a number of characteristics from the short-form questionnaire or from administrative data sources. The weighting methodology is described in detail in Chapter 4. All private households attached to collective dwellings and all private households in CUs in First Nations communities, Métis settlements, Inuit regions and other remote areas were selected for the long-form sample and received a design weight of 1. They were then excluded from the coverage and non-response adjustment processes, as well as from the final calibration process.

Long-form sample households with a non-zero weight at the end of the weighting process were the respondent households, along with the households who were assigned a design weight of 1, i.e., private households attached to collective dwellings and all private households in CUs in First Nations communities, Métis settlements, Inuit regions and other remote areas. These households made up the set of households that contributed to the long-form estimates.

3.11 Final response rates

Table 3.11.1 presents the final response rates for private households in the 2021 Census of Population, for Canada and for each province and territory, followed by non-weighted and weighted response rates for the long-form sample based on the definition of non-response given in Section 3.9.

The final response rate is the ratio of the numerator to the denominator, where:

- the numerator is the number of private dwellings for which a questionnaire was completed[6]
- the denominator is the number of private dwellings classified as occupied, according to the census database.

The final classification of a dwelling's occupancy status is based on an analysis of the data gathered by field staff, data provided by respondents and the results of a study into the quality of occupancy status in the DCS (see Section 3.6). The response rates indicated in Table 3.11.1 differ from the collection response rates, which were previously published and were mentioned in Section 1.5, in that they take data processing and dwelling occupancy verification into account in identifying non-respondent households. These response rates are therefore considered final.

Weighted response rates were produced for the long-form sample. They are defined as the ratio of the numerator to the denominator, where:

- the numerator is the design-weighted count of private dwellings for which a questionnaire was completed
- the denominator is the design-weighted count of private dwellings classified as occupied, according to the census database.

Table 3.11.1
Final response rates for private households from the 2021 Census of Population and the long-form sample

	Response rate—short-form questionnaire	Non-weighted response rate—long-form questionnaire only	Weighted response rate—long-form questionnaire only
Region		percent	
Canada	96.9	94.9	95.7
Newfoundland and Labrador	97.0	95.0	95.6
Prince Edward Island	97.6	96.5	96.8
Nova Scotia	97.1	95.6	96.1
New Brunswick	96.8	94.8	95.7
Quebec	97.1	95.7	96.3
Ontario	97.2	95.8	96.2
Manitoba	96.5	93.1	94.4
Saskatchewan	95.5	91.8	93.5
Alberta	96.5	93.4	94.4
British Columbia	96.5	94.0	95.1
Yukon	95.7	85.5	89.5
Northwest Territories	91.8	86.2	89.2
Nunavut	79.7	78.1	78.1

Note: All private households and occupied dwellings are included in the calculation of these response rates, without exception.
Sources: Statistics Canada, 2021 Census of Population and 2021 Census long-form sample.

[6] Private dwellings attached to collective dwellings, which were included in the long-form sample but received only a short-form questionnaire, were considered as non-respondents for the purposes of calculating long-form questionnaire response rates.

4. Estimation from the census long-form sample

Any sampling process requires an associated estimation procedure for scaling sample data up to the population level and for ensuring that survey estimates are representative of the population. The choice of an estimation procedure is generally governed by both operational and theoretical constraints. From the operational viewpoint, the procedure must be feasible within the processing system of which it is a part, and from the theoretical viewpoint, the procedure should minimize the statistical error of the estimates it produces.

The estimation procedure produces a set of weights, and the weight for each sample unit corresponds to the number of units in the population that the sample unit represents. These weights are applied to the sample data to produce millions of estimates from the census long-form sample. Estimates are summary measures such as totals, averages, proportions and medians calculated from the sample for various characteristics of interest.

4.1 Considerations in the choice of an estimation procedure

4.1.1 Operational considerations

Mathematically, an estimation procedure can be described by an algebraic formula, or estimator, that shows how the estimate for the population is calculated as a function of the observed sample values and other information from the sample design or external data sources. Most of the time, this estimator is a simple function of weights and of the variable of interest for the responding units. Using a unique set of weights to produce all estimates guarantees a certain level of consistency among the different estimates of the survey.

Therefore, the approach taken for the census long-form sample (and in most sample surveys) was to split the estimation procedure into two steps: (a) the calculation of weights (known as the weighting procedure) and (b) the use of weights to produce estimates, such as the estimation of a particular population count by summing the weights of those persons or households with the characteristic of interest. Most of the mathematical complexity is contained in step (a), which is performed just once. Meanwhile, step (b) is reduced to a simple process, such as summing weights whenever tabulation is required. Since the weight attached to each sample unit is the same for any tabulation involving that unit, consistency between different estimates based on sample data is assured.

4.1.2 Theoretical considerations

For a given sample design and a given estimation procedure, one can, from sampling theory, make a statement about the chances that a certain interval will contain the unknown population value being estimated. A primary criterion in the choice of an estimation procedure is the minimization of the width of such intervals for a given level of confidence so that these statements about the unknown population values are as precise as possible. A common measure of precision for comparing estimation procedures is known as the standard error. Provided that certain conditions are met, intervals of plus or minus two standard errors from the estimate will contain the true population value for approximately 95% of all possible samples. Chapter 7 details the conditions and methods to compute confidence intervals for the census long-form.

As well as minimizing standard error, a second objective in the choice of an estimation procedure for the long-form sample is to ensure, as far as possible, that sample estimates for census characteristics are consistent with the corresponding known census values. Fortunately, these two objectives are usually complementary in the sense that sampling error tends to be reduced by ensuring that sample estimates for certain basic characteristics are consistent with the corresponding population figures. However, while this is true in general, forcing long-form sample estimates for census characteristics to be consistent with corresponding census figures for very small subgroups can have a detrimental effect on the standard error of estimates for the sample characteristics themselves. For example, if in several dissemination areas only a few subjects have a given characteristic, such as birth in a certain country, ensuring consistency between the sample estimates and the census counts for that place of birth would unduly increase the standard error for the rest of the characteristics.

In cases where no information about the population being sampled is available other than that collected for sample units and unit non-response has not occurred, the estimation procedure would be restricted to weighting the sample units inversely to their probability of selection. For example, if a unit had a one-in-four chance of selection, then that selected unit would receive a weight of 4. When unit non-response is observed, the weight must be further adjusted according to the estimated probability of response of the unit, for example. In practice, some supplementary knowledge about the population (e.g., its total size and possibly its breakdown by a certain variable—perhaps by province and territory) is often available. Such information can be used to improve the estimation formula so as to produce estimates with a greater chance of being close to the unknown population value. In the case of the census long-form sample, a large amount of very detailed information about the population being sampled is available from the census short-form data at every geographic level. This wealth of population information is used in the coverage, non-response and calibration adjustments to improve the estimates made from the long-form sample.

Nevertheless, the long-form sample estimates for census characteristics cannot be made consistent with all the census counts at every geographic level. Differences between sample estimates and census counts become visible when a cross-tabulation of a sample variable and the corresponding census variable is produced. The tabulation of sample-based estimates of totals for particular characteristics will not necessarily agree with the equivalent census count tabulations for those characteristics.

Adjusting the weights by the most minimal amounts possible to achieve perfect agreement between long-form estimates and census counts for certain characteristics and subgroups is known as “calibration.”

4.2 Weighting areas

The various adjustments to design weights were made independently by weighting area. The geographic areas used for this purpose were aggregate dissemination areas (ADAs) and super aggregate dissemination areas (SADAs). ADAs were first introduced with the 2016 Census. SADAs were created specifically for the weighting procedures by ADA aggregation.

4.2.1 Aggregate dissemination areas

In total, for the 2021 Census, Canada was divided into 5,433 ADAs. Households were selected for the long-form sample in 5,191 ADAs. Of the 242 ADAs without sampled households, 237 consisted solely of out-of-scope households. The other five ADAs had only a handful of in-scope households, and none of them were selected.

The 2021 ADAs were constructed by making minimal changes to the 2016 ADAs to accommodate for changes at the dissemination area (DA) level. The goal was to allow for historical comparability in ADAs. Because criteria related to size are most relevant to the weighting process, the 2016 ADA delineation criteria are presented below.

ADAs satisfy the following delineation criteria:

1. ADAs cover the entire country and, where possible, have a population count of 5,000 to 15,000 (based on the population counts from the previous census).
2. ADAs respect provincial and territorial borders, as well as the boundaries of census divisions (CDs), census metropolitan areas (CMAs) and census agglomerations (CAs) subdivided into census tracts (CTs) in effect for the 2016 Census.
3. ADAs are based on one of three 2016 Census dissemination geographic areas: DAs, census subdivisions (CSDs) or census tracts (CTs):
    - Within CMAs and CAs with CTs, adjacent CTs are combined to meet the ADA population criterion.
    - In areas without CTs (areas outside CMAs and the largest CAs) where CSDs have a population of fewer than 15,000, adjacent CSDs are combined to meet the ADA population criterion.
    - In areas without CTs where CSDs have a population of over 15,000, adjacent DAs are combined within these CSDs to meet the ADA population criterion.
4. Every CSD that consists of an Indian reserve and a small number of other areas where the canvasser method is required constitute distinct ADAs.

“For more information about aggregate dissemination areas, refer to the *Dictionary, Census of Population, 2021,* Catalogue no. 98-301-X.”

Table 4.2.1.1 shows the degree to which ADAs with households in the long-form sample were properly adjusted to CSDs. The first scenario occurred in most cases, since ADAs were designed above all to respect the boundaries of CTs and CSDs. Scenario 4 is the only one where CSD boundaries were not respected. CTs were not included in the table because they were all in the first scenario except one, which was in scenario 3.

Table 4.2.1.1
Number of census subdivisions within the boundaries of aggregate dissemination areas with households in the long-form sample, 2021 Census

		Census subdivision	
Scenario	**Description**	number	percent
1	The CSD was small enough to be fully contained in an ADA, and this ADA only had complete CSDs. No CSDs in the ADA were part of another ADA.	4,526	93.26
2	The CSD was small enough to be fully contained in an ADA, but another CSD in the same ADA was part of a different ADA.	39	0.80
3	The CSD was large enough to contain full ADAs. No ADAs were part of another CSD.	262	5.40
4	The CSD was part of two or more ADAs.	26	0.54
Total		**4,853**	**100.00**

CSD = Census subdivision
ADA = Aggregate dissemination area
Source: Statistics Canada, 2021 Census long-form sample.

Table 4.2.1.2 shows the distribution of ADAs with households in the long-form sample by province or territory.

Table 4.2.1.2
Number of aggregate dissemination areas with households in the long-form sample, by province or territory

Region	**Number of ADAs**
Newfoundland and Labrador	83
Prince Edward Island	23
Nova Scotia	148
New Brunswick	129
Quebec	1,144
Ontario	1,659
Manitoba	222
Saskatchewan	263
Alberta	515
British Columbia	912
Yukon	29
Northwest Territories	38
Nunavut	26
Canada	**5,191**

ADA = Aggregate dissemination area
Source: Statistics Canada, 2021 Census long-form sample.

Table 4.2.1.3 shows the number of ADAs by the number of in-scope households in the census. The majority of ADAs with households in the long-form sample had from 2,000 to 4,999 households. A considerable number of ADAs had small populations.

Table 4.2.1.3
Distribution of aggregate dissemination areas with households in the long-form sample, by number of in-scope households

In-scope households	Number of ADAs	Percent
0 to 499	996	19.19
500 to 999	118	2.27
1,000 to 1,999	359	6.92
2,000 to 2,999	1,190	22.92
3,000 to 3,999	1,189	22.91
4,000 to 4,999	733	14.12
5,000 to 5,999	356	6.86
6,000 to 6,999	143	2.75
7,000 to 7,999	46	0.89
8,000 to 8,999	29	0.56
9,000 to 9,999	13	0.25
10,000 and over	19	0.37
Total	**5,191**	**100.00**

ADA = Aggregate dissemination area
Source: Statistics Canada, 2021 Census of Population.

Table 4.2.1.4 presents the number of ADAs by range of numbers of households that responded to the 2021 Census long-form questionnaire. For the ADAs with the fewest respondents, a specific type of processing was applied to have enough households for weighting purposes (see Section 4.5).

Table 4.2.1.4
Distribution of aggregate dissemination areas with households in the long-form sample, by number of respondent households for the long-form questionnaire

Number of respondents	Number of ADAs	Percent
0 to 99	690	13.29
100 to 199	276	5.32
200 to 299	132	2.54
300 to 399	128	2.47
400 to 499	272	5.24
500 to 599	478	9.21
600 to 699	559	10.77
700 to 799	583	11.23
800 to 899	499	9.61
900 to 999	411	7.92
1,000 to 1,099	322	6.20
1,100 to 1,199	246	4.74
1,200 to 1,299	189	3.64
1,300 to 1,399	128	2.47
1,400 to 1,499	98	1.89
1,500 and over	180	3.47
Total	**5,191**	**100.00**

ADA = Aggregate dissemination area
Source: Statistics Canada, 2021 Census long-form sample.

4.2.2 Super aggregate dissemination areas

SADAs were created specifically for weighting 2016 Census data, so that certain weighting procedures for which a large number of observations is desirable could be conducted.

The 2021 SADAs were constructed by making minimal changes to the 2016 SADAs to accommodate for changes at the ADA level. Since criteria on size are of particular interest for the weighting process, the 2016 SADA delineation criteria are presented below.

SADAs were created according to the following rules (in order of priority):

1. SADAs are created by combining ADAs (mandatory).
2. SADAs respect provincial and territorial borders (mandatory).
3. SADAs have a population of 50,000 to 150,000 persons (except for CDs with a population of 40,000 to 50,000 persons that constitute their own SADA) excluding persons living in canvasser collection units (CUs).
4. SADAs respect the boundaries of CDs.
5. SADAs respect the boundaries of CMAs and CAs.
6. SADAs respect the boundaries of CSDs.
7. SADAs are single contiguous entities.
8. SADA are as compact as possible.

The first two rules were mandatory, and rules 3 to 9 were followed where possible. A total of 409 SADAs were created.

Table 4.2.2.1 shows the distribution of SADAs by province or territory.

Table 4.2.2.1
Number of super aggregate dissemination areas, by province or territory

Region	Number of SADAs
Newfoundland and Labrador	8
Prince Edward Island	2
Nova Scotia	13
New Brunswick	8
Quebec	97
Ontario	150
Manitoba	15
Saskatchewan	14
Alberta	44
British Columbia	55
Yukon	1
Northwest Territories	1
Nunavut	1
Total	**409**

SADA = Super aggregate dissemination area
Note: In the case of the three territories, the SADA corresponds to the territory.
Source: Statistics Canada, 2021 Census long-form sample.

Table 4.2.2.2 shows the degree to which SADAs were properly adjusted to CDs and CMAs. SADAs respected the boundaries of the majority of CDs (scenarios 1 and 3) and the boundaries of three-quarters of CMAs. The other CMAs were part of at least two SADAs (scenario 4).

Table 4.2.2.2
Number of census divisions and census metropolitan areas within super aggregate dissemination area boundaries, 2021 Census

		Census divisions		Census metropolitan areas	
Scenario	**Description**	number	percent	number	percent
1	The CD or CMA was small enough to be fully contained within a SADA, and the SADA included only complete CDs or CMAs. No CDs or CMAs in the SADA were part of another SADA.	249	84.98	6	14.63
2	The CD or CMA was small enough to be fully contained within a SADA, but another CD or CMA in the same SADA was also part of another SADA.	2	0.68	0	0.00
3	The CD or CMA was large enough to contain complete SADAs. No SADAs were also part of another CD or CMA.	40	13.65	26	63.41
4	The CD or CMA was part of two or more SADAs.	2	0.68	9	21.95
Total		**293**	**100.00**	**41**	**100.00**

CD = Census division
CMA = Census metropolitan area
SADA = Super aggregate dissemination area
Source: Statistics Canada, 2021 Census of Population.

Table 4.2.2.3 shows the number of SADAs by the number of in-scope persons.

Table 4.2.2.3
Distribution of super aggregate dissemination areas with households in the long-form sample, by number of in-scope individuals

In-scope individuals	Number of SADAs	Percent
30,000 to 39,999	3	0.73
40,000 to 49,999	20	4.89
50,000 to 59,999	23	5.62
60,000 to 69,999	29	7.09
70,000 to 79,999	101	24.69
80,000 to 89,999	66	16.14
90,000 to 99,999	46	11.25
100,000 to 149,999	114	27.87
150,000 and over	7	1.71
Total	**409**	**100.00**

SADA = Super aggregate dissemination area
Source: Statistics Canada, 2021 Census of Population.

4.3 Design weights

The design weight for each household in the long-form sample was calculated differently, depending on the census delivery method of the area where the corresponding dwelling was located.

- If the delivery method was mail-out (MO), list/ leave (L/L), or mail-out with drop-off (MODO), the design weight was equal to the inverse of the survey fraction, giving a weight of 4.
- Households located in First Nations communities, Métis settlements, Inuit regions and other remote areas were assigned a design weight of 1.

Households living in private dwellings attached to collective dwellings were an exception to the rule. As mentioned in Section 2.2, all of these households were included in the sample. They were considered take-all, so their design weight was 1.

4.3.1 Weights for households counted in the sample

Sampled households with a design weight of 1 did not have their weight adjusted. These households kept their weight of 1 after the weighting procedures were completed (coverage and non-response, as well as calibration to census counts). They either were located in canvasser CUs or were private households that were attached to a collective dwelling.

Total non-response and partial non-response for these households were addressed by imputation. Once the missing data were imputed, these households were considered to be respondents for estimation purposes (although they were considered to be non-respondents for the calculation of response rates in Section 3.11).

4.4 Coverage and total non-response adjustment

While there are several ways of treating non-response in surveys, they can be divided into two main categories: imputation and reweighting. The former is usually applied for the treatment of items missing values and the latter for the treatment of total non-response. A household was considered to be a respondent to the long-form questionnaire when it answered at least one of the long-form questions. With the high response rate to the long-form questionnaire, any non-response adjustment method would have had, for the most part, only a modest impact on the final survey weights and estimates. Coverage and total non-response for households in CUs in First Nations communities, Métis settlements, Inuit regions and other remote areas were compensated for with imputation procedures and, for the most part, with whole household imputation (WHI) as described in Section 3.6. In the rest of the country, reweighting procedures were used. The rest of this chapter describes those weighting procedures.

The main purpose of coverage and non-response adjustments is to minimize the impact of any potential biases from lack of complete coverage (or from duplicates) and from unit non-response. For the adjustment to actually reduce the potential bias, a rich set of information about the non-respondents is very useful. Otherwise, the non-response adjustment that can be applied is limited, and the potential bias will not be greatly lessened. Only geographical information was known for every non-responding household. The information on non-respondents was therefore somewhat limited. Fortunately, before the coverage and non-response adjustments, the process of WHI occurred. An important part of WHI is to impute the short-form characteristics for all non-respondents to the short form. This included long-form sample non-respondents who did not answer any short-form questions. This additional information served as the basis for the long-form sample non-response adjustment.

The method used to adjust for coverage and total non-response in the long-form sample was a reweighting calibration-based procedure applied to the design weights. The procedure can be divided into the following main steps:

1. selection of calibration constraints for steps 2 and 3
2. non-linear calibration coverage adjustment
3. estimation of a non-response propensity based on non-linear calibration for non-response
4. application of a score method based on the propensity of step 3.

Steps 1 to 4 were applied independently in each SADA. In other words, the non-response adjustment was applied by SADA. See Section 4.2 for the definition and information about ADAs and SADAs.

The first step consisted of a forward selection of calibration constraints in the SADA. It was performed as follows:

- The set of potential constraints was derived from the variables common to both the short-form and the long-form, as well as from some administrative data obtained with record linkage strategies (where all units of the long-form population undergo the linkage procedures). The requirements of the non-linear calibration method used in the second and third steps meant that only constraints at the SADA level, and the number of households and persons in each ADA of the SADA, were considered.
- In each SADA, two mandatory constraints were selected first: the number of households in the SADA (TOTHHLD) and the number of persons in the SADA (TOTPERS).
- The ADA-specific constraints—number of households (HHADA) and number of persons (PPADA)—were evaluated for selection.
- All other potential SADA constraints were evaluated; priority was given to the ones that split the SADA population as closely as possible into halves.

The selection process excluded constraints that occurred in fewer than 250 households in the SADA and constraints that were redundant or almost redundant in terms of collinearity with those constraints or with constraints already selected. Constraints that were redundant with constraints already selected were excluded since they did not add any new information. Given those filters, the order of priority used in the evaluation of constraints ensured that the constraints selected complemented each other and corrected for any potential coverage differential between the long-form and the short-form, as well as for census total non-response.

The second step applied a coverage non-linear calibration adjustment to the whole sample in the SADA (i.e., respondents and non-respondents). The long-form sample weighted counts, for the constraints selected in the first step, were made to coincide with the corresponding population counts. The purpose of this step was to correct for any potential coverage differential between the long-form sample and its complement (i.e., the set of households receiving only the short form). One way in which overcoverage can occur is if some individuals are counted in two different households. The coverage for the two populations could also be different if, for example, occupied dwellings were more likely to be incorrectly treated as unoccupied dwellings for the long-form than for the short-form. Another objective of this step was to isolate as much as possible the sampling error. Without this step, the non-response calibration carried out in the next step would confound the non-response error with the sampling error. This step makes the sample estimates coincide with the population estimates. In addition, the same control totals are used in both calibration procedures. As a result, the non-response propensity estimation done next does not have to correct (directly or indirectly) for the sampling error. Combining a correction for the sampling error and for the non-response error in the next step would have been inappropriate. The calibration procedure would have failed if the weight of any respondent was required to decrease to match the census counts, because the estimated propensity would have been greater than 1. Moreover, the score method applied in the last step required an estimate of the response propensity alone. To the extent that the variable of interest was related to the selected constraints, the sampling variance was also reduced by this step.

After these two steps, the main non-response adjustment took place. The weights, adjusted in the previous step, of non-respondents were set to 0 and the weights of respondents were increased so that the weighted sums in the SADA coincided with the corresponding population counts for the selected constraints. A logistic link function between the response propensity and the characteristics used in calibration enabled the implicit estimation of the response propensity. Folsom and Singh (2000) proposed this non-linear calibration method as a way of adjusting for non-response while ensuring both that the estimates coincided with selected population counts and that the estimated response probabilities were between 0 and 1. This last condition does not necessarily hold when linear calibration is used for non-response adjustment. To the extent that the response propensity was related to the selected constraints, this step reduced the potential non-response bias without increasing the variance.

The inverse of the estimated response probabilities obtained in the previous step could be directly used to adjust the weights for non-response. However, the score method was used for the last step of the non-response adjustment to smooth the estimated probabilities from the previous step. This further ensured the quality of the non-response adjustment and avoided overly large adjustments. For each ADA, homogeneous weighting classes were formed according to the estimated response probabilities. In each class, the weighted harmonic mean of the response probabilities was calculated. The harmonic mean was used because it is less affected by outliers in the estimated response probabilities. The inverse of this mean was applied to the weights of respondents in the class as the non-response adjustment. This is equivalent to applying the weighted arithmetic mean of the weight adjustment factors in each homogeneous weighting class, where the adjustment factors would be the inverse of the estimated response propensities.

In summary, the coverage and total non-response adjustment was a product of two quantities: the coverage adjustment and the inverse of the score-method harmonic mean.

4.5 Final calibration

Final calibration is a linear calibration that was done to minimize the sampling variability of estimates derived from long-form questionnaire responses, while ensuring consistency between estimated totals and Census of Population totals. This weighting step was necessary, since ensuring consistency between estimated totals and Census of Population totals was important for a large number of variables and geographic areas, i.e., satisfying calibration constraints.

Only the weights for households in MO, L/L or MODO areas were calibrated, since these households were sampled. Exceptions to this rule were households in these areas that lived in a private dwelling attached to a collective dwelling. Since all these households were included in the long-form sample and all the long-form questionnaire responses for these households were imputed, no calibration was done. The final weights for these households were therefore equal to 1. The weights produced by the calibration process were the final weights used to calculate the long-form estimates, and these weights applied to households as well as families and persons. In other words, all families and persons from the same household received the household weight. For this final adjustment, the variability of the calibrated weights needed to be limited to avoid having an excessive portion of the weight applied to a single household or person. Therefore, weights were constrained to range from 1 to 20.

Calibration constraints were defined at the person, household and census family levels. Additionally, constraints can be selected at two different geographical levels, at the ADA or at the SADA level. These two levels maximize the overall consistency between estimated totals and Census of Population totals, while minimizing the number of calibration constraints. This helps to reduce the variability of estimates. Appendix C lists all the ADA and SADA constraints that were taken into consideration during the calibration process. Characteristics available from the census, administrative sources and the long-form questionnaire and for which consistency was attempted included age, gender, marital status, common-law status, household size, dwelling type, official language spoken, year of immigration and place of birth.

The constraints selection process is applied simultaneously to a SADA and its ADAs, but independently across SADAs. Calibration was then performed using all of the selected constraints. The 2021 calibration process saw the addition of three new constraints. These were the number of persons who live in an apartment in a building that has five or more storeys (APT5PLUS) and two constraints related to the number of persons who immigrated from 2016 to 2021 (YRIMD_2016 and YRIMG1_2016). Additionally, constraints previously based on the 2016 sex concept are now based on the 2021 two categories gender variable.[7] In total, 203 constraints were defined for SADAs and 271 for ADAs. Various factors drove the choice of geographic level for calibration constraints. This choice was made in collaboration with subject-matter experts. For example, some constraints were defined only for SADAs, since they would not have been populated enough at the ADA level. Other constraints, such as age groups, were chosen in a way that ensured they were not only populated enough but also not too similar when assessed by the selection process.

[7] For more information on the two category gender variable, please refer to *Age, Sex at Birth and Gender Reference Guide, Census of Population, 2021*, Catalogue no. 98-500-X2021014.

To facilitate their calibration, small ADAs were combined before the selection of calibration constraints to ensure a minimum of 60 long-form respondent households per ADA. Small ADAs that fell entirely within a CSD were initially combined with other ADAs in the same SADA. Next, small ADAs in CDs were combined with other ADAs in the same SADA. Finally, the remaining small ADAs were combined with an ADA from an adjacent SADA. The ADA grouping procedure produced 4,207 groups of ADAs with 60 or more respondent households.

The first step in the process to select calibration constraints was to categorize each of the constraints into one of three groups:

Mandatory constraints: These constraints had to be used in the calibration because the census counts had to agree with the long-form estimates at the geographic levels that are usual aggregates of ADAs and SADAs (e.g., Canada, provinces and territories). The number of persons and the number of households in the ADAs and SADAs were the two mandatory constraints.

Low-response constraints: Constraints evaluated for a population of 200 or fewer households were not used in the calibration because they can make survey estimates unstable.

All other constraints: These constraints were examined further to see whether they should be used in the calibration.

The second step was to determine which constraints from the third group should be used in the calibration process, in addition to the mandatory constraints. The constraints from the third group were added one by one, by repeatedly choosing the constraint that divided the population of the SADA or ADA in two as evenly as possible. Constraints that were too linearly dependent were excluded. To avoid introducing a bias in the point estimates and to avoid increasing their variance, the number of selected constraints was limited. Evaluations determined that this number had to be smaller than the square root of the number of respondent households involved in the constraint.

After the calibration constraints were selected, a final edit was done to check whether the set of constraints chosen at the ADA and SADA levels was free of collinearity.

The calibration itself was then carried out for the final set of constraints from the second step. The weights adjusted for coverage and non-response were modified as little as possible, so that the weighted estimates would be equal to census totals for these constraints. Statistics Canada's Generalized Estimation System (GES) was used to carry out the calibration.

Sample estimates can differ from census counts for a few reasons, particularly for small areas, even after the calibration step. A few of these reasons are given below.

- Constraints excluded during the constraint selection process: As described above, possible constraints could be excluded for having low counts, for being linearly dependent (or overly dependent) on other chosen constraints or for being linearly dependent (or overly dependent) on low-response constraints. This led to some differences between census counts and long-form estimates for these variables when a perfect linear dependency with the chosen constraints was not present.
- Sub-weighting area: The ADA was the smallest weighting area for which agreement was attempted between the census counts and the long-form estimates. Any entity smaller than an ADA, such as the majority of DAs, is referred to as a sub-weighting area. These sub-weighting areas could have discrepancies between the census counts and the long-form estimates.

4.6 Details on the selection of constraints

Constraints were selected twice during the weighting process: first during the coverage and non-response adjustment discussed in Section 4.4, and again during the final calibration discussed in Section 4.5. The variables making up the constraints were essentially the same, but the inclusion or exclusion of constraints varied slightly between the two weighting steps to better align with the objective of each step. This section explains how the constraints were selected during these weighting steps.

The constraint selection process, for both adjustments, started from a set of mandatory constraints detailed in the previous sections and then evaluated the addition of every other candidate constraint one by one. The order in which candidate constraints were evaluated was identical for all SADAs. When a constraint was introduced, the **no population** and **small population** criteria were evaluated and the constraint would be rejected if either criterion failed. If a constraint passed both criteria, the augmented set of constraints including it was then evaluated for **linearly dependent, high collinearity** and **explanatory redundancy** criteria. If it failed any of the criteria, the constraint was rejected. Otherwise, the constraint was added to the pool of constraints included and the selection process iterated to the next candidate constraint from the list. Table 4.6.1 summarizes those five criteria, whether they were applied for each of the two processes and differences in parameterization of the criteria between the two weight adjustment processes.

For each weight adjustment process, the constraint selection was carried out independently in each of the 408 SADAs that had sampled households with an adjusted weight.

See Appendix C for the list of constraints and a frequency distribution of their respective inclusion or exclusion for each of the two weighting process.

Table 4.6.1
Criteria applied in selecting coverage, non-response and final calibration adjustment constraints

Criteria	Adjustment for coverage and non-response	Final calibration
No population according to the census counts: If the constraint had no population in the weighting area, then the estimate after adjustment must also be 0 for that constraint. These constraints are not classified as excluded but rather as ineligible to the adjustment process.	Applied at the SADA/ADA level.	Applied at the SADA/ADA level.
Small population according to the census counts: If a constraint involves less than a certain number of households in the population of the weighting area, then it is considered small and the constraint is excluded. Including such a constraint would unduly increase the variance. However, constraints with small population can be implicitly calibrated and in this case are included in the total number of calibrated constraints.	Applied at the SADA/ADA level. The number of households in the population is larger than 0 but less than 250 in the weighted area.	Applied at the SADA/ADA level. The number of households in the population is more than 0 but less than 200 in the weighted area.

Table 4.6.1
Criteria applied in selecting coverage, non-response and final calibration adjustment constraints

Criteria	Adjustment for coverage and non-response	Final calibration
Linearly dependent: If the value of a constraint can be calculated by combining the values of other constraints, one of these constraints is not necessary and must be deleted during the adjustment process because of its linear dependency. However, constraints that are excluded because of their linear dependency are implicitly calibrated. They are therefore included in the total number of calibrated constraints.	Applied at the SADA level. The selection of constraints can be compared to the selection of explanatory variables in a linear regression. The VIF[1] and the condition number[2] are thus used to detect high collinearity.	Applied at the SADA/ADA level. Two dependency checks are conducted to identify linearly dependent constraints. The first check is done when the constraints at the SADA/ADA level are selected, and the second check includes all the constraints chosen at both levels of the geographic hierarchy (SADAs and ADAs).
High collinearity: If a constraint value can be almost calculated by the combination of other constraint values, then at least one of those constraints must be avoided in the adjustment process. Such a constraint is not perfectly calibrated.	Applied at the SADA level. The selection of constraints can be compared to the selection of explanatory variables in a linear regression. The VIF[1] and the condition number[2] are thus used to detect high collinearity.	Applied at the SADA/ADA level. Two linear dependency checks are conducted to identify constraints that are close to being linearly dependent. The first check is done when the constraints at the SADA level and the ADA level are selected, and the second check includes all the constraints chosen at both levels of the hierarchy simultaneously (SADAs and ADAs).
Explanatory redundancy: If a constraint explains the non-response (almost) as well as other constraints already selected, then the non-response calibration procedure would fail. This is equivalent to saying that if a constraint does not add any information about the non-response mechanism, beyond what is explained by the already-selected constraints, then it should not be included.	Applied at the SADA level. A sequential procedure is applied (a form of logistic regression) to test the convergence of the logistic regression.	N/A

SADA = Super aggregate dissemination area
ADA = Aggregate dissemination area
VIF = Variance inflation factor
N/A = Not available

[1] The VIF quantifies the increase in variance of regression coefficients attributable to collinearity.

[2] The condition number quantifies the degree to which a matrix is close to singularity.

Source: Statistics Canada, 2021 Census long-form sample.

5. Evaluation of the weighting procedures

As described in Chapter 4, the first step in weighting the long-form sample was to assign design weights to households. Weights were assigned differently depending on the collection method of the area where the household was located. Private households attached to a collective dwelling or part of a First Nations community, Métis settlement, Inuit region, or other remote area have a design weight of 1. The final weight for these private households corresponds to the design weight and remains at the initial value of 1. All the other private households have a design weight greater than 1. All of the results presented in this chapter were calculated for the subset of households with a design weight greater than 1.

In short, each household was assigned a design weight that was determined by the long-form sample design. Some adjustments were then necessary to address coverage and total non-response. Non-linear calibration was performed during this adjustment to estimate the parameters for non-response models. After being adjusted for coverage and total non-response, the weights were adjusted further in the final calibration process to produce the final weights. The final weights enabled generally better agreement between the census counts and the long-form estimates.

The next few sections examine the distribution of the weights and, for various characteristics, the discrepancies between the census counts and the sample.

5.1 Distribution of the weights

Chart 5.1.1 and tables 5.1.1 and 5.1.2 illustrate the distribution of the design weights, the weights adjusted for coverage and non-response, and the weights adjusted in the final calibration. The weights are grouped by 0.5 length intervals, apart from the first and last intervals. The chart shows the percentage of times the weights appear in each interval. All the design weights ranged from 3.75 to 4.25. This is because of the long-form sample design, in which approximately one in four households received a long-form questionnaire in most areas. The impact on the weights of the coverage and non-response adjustments and the final calibration adjustments can also be seen. A very noticeable difference is shown in the 3.75 to 4.25 interval. In fact, 55% of households had their coverage and non-response adjusted weights between 3.75 and 4.25 compared with 100% of design weights in this interval. After the final calibration adjustment only 30% of households had their final weights in this interval. The final weights were more evenly distributed across all categories compared with the design and coverage and non-response adjusted weights.

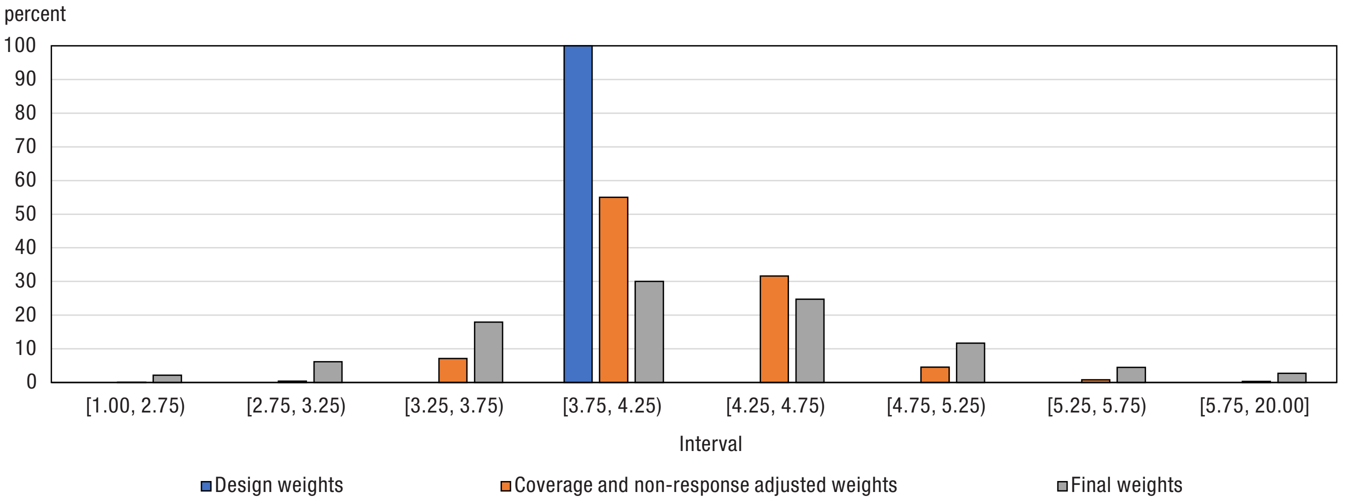
Logically, the non-response adjustment process should tend to increase the weights to compensate for the non-responding units. This did occur for most cases. The changes between the design weights and the coverage and non-response adjusted weights can be observed in Table 5.1.1. This table shows that most of the units that left the [3.75, 4.25) range moved to the [4.25, 4.75) or [4.75, 5.25) ranges. However, the coverage and non-response adjustment process also moved some weights from the [3.75, 4.25) range to the [1.00, 2.75), [2.75, 3.25) or [3.25, 3.75) ranges for some units. The main reason is that the procedure included an adjustment for overcoverage and undercoverage. To the extent that a few population groups may have experienced overcoverage, the weights would have been reduced in those areas.

An important element to notice in Table 5.1.1 is that the non-responding units originally had positive weights, since they were selected for the sample. The non-response adjustment process assigned them a weight of 0 and redistributed their original weights among responding units. The non-responding units correspond to the line labelled “0 (non-respondents)” in Table 5.1.1 and were removed from Table 5.1.2, since they were not used in the calibration process. Table 5.1.2 presents the changes between the coverage and non-response adjusted weights and the calibrated weights.

According to Table 5.1.2, most weights experienced only a small modification during the calibration process. In fact, 78.1% of cases either stayed in the same range or moved only one range up or down. The most stable range was 1.00 to 2.75, where 64% of the households with a coverage and non-response adjusted weight between 1.00 and 2.75 stayed in that category after calibration. The second most stable category was 5.75 to 12.25 where 61.9% of households with a coverage and non-response adjusted weight between 5.75 and 12.25 stayed in that category (although the calibration range goes up to 20.00 rather than 12.25).

Finally, whereas the coverage and non-response adjusted weights varied from 1.00 to 12.25, the range of the final weights was from 1.00 to 20.00.

Chart 5.1.1
Distribution of design weights, coverage and non-response adjusted weights, and final weights



Notes: All households with a design weight of 1 were excluded from the weighting process. These households were either located in First Nations communities, Métis Settlements, Inuit regions or other remote areas, or were private households attached to a collective dwelling.
The "[" symbol means the number is included in the interval and the ")" symbol means it is not included in the interval.
Source: Statistics Canada, 2021 Census long-form sample.

Table 5.1.1
Distribution of design weights and coverage and non-response adjusted weights

	Design weights	
Coverage and non-response adjusted weights	**[3.75, 4.25)**	**Total**
0 (non-respondents)	149,945	149,945
[1.00, 2.75)	1,555	1,555
[2.75, 3.25)	14,575	14,575
[3.25, 3.75)	252,583	252,583
[3.75, 4.25)	1,949,875	1,949,875
[4.25, 4.75)	1,121,300	1,121,300
[4.75, 5.25)	161,359	161,359
[5.25, 5.75)	29,512	29,512
[5.75, 20.00]	11,441	11,441
Total	**3,692,145**	**3,692,145**

Notes: All households with a design weight of 1 were excluded from the weighting process. These households either were located in First Nations communities, Métis Settlements, Inuit regions or other remote areas, or were private households attached to a collective dwelling.
The “[“ symbol means the number is included in the interval and the “)” symbol means it is not included in the interval.
Source: Statistics Canada, 2021 Census long-form sample.

Table 5.1.2
Distribution of coverage and non-response adjusted weights and final weights

	Coverage and non-response adjusted weights								
Final weights	**[1.00, 2.75)**	**[2.75, 3.25)**	**[3.25, 3.75)**	**[3.75, 4.25)**	**[4.25, 4.75)**	**[4.75, 5.25)**	**[5.25, 5.75)**	**[5.75, 12.25)**	**Total**
[1.00, 2.75)	999	4,045	19,737	38,100	11,595	1,706	364	174	76,720
[2.75, 3.25)	350	4,559	47,337	131,632	31,612	2,910	395	115	218,910
[3.25, 3.75)	144	3,710	84,405	426,528	111,790	7,757	909	193	635,436
[3.75, 4.25)	42	1,527	65,228	690,146	286,261	18,837	1,895	394	1,064,330
[4.25, 4.75)	13	514	25,637	451,501	359,080	36,037	3,507	678	876,967
[4.75, 5.25)	6	151	7,430	154,852	205,139	41,044	5,725	1,166	415,513
[5.25, 5.75)	1	45	2,020	42,165	77,392	28,209	6,127	1,642	157,601
[5.75, 20.00]	0	24	789	14,951	38,431	24,859	10,590	7,079	96,723
Total	**1,555**	**14,575**	**252,583**	**1,949,875**	**1,121,300**	**161,359**	**29,512**	**11,441**	**3,542,200**

Notes: All households with a design weight of 1 were excluded from the weighting process. These households either were located in First Nations communities, Métis Settlements, Inuit regions or other remote areas, or were private households attached to a collective dwelling. The "[" symbol means the number is included in the interval and the ")" symbol means it is not included in the interval.
Source: Statistics Canada, 2021 Census long-form sample.

5.2 Discrepancies between census counts and long-form estimates, Canada

Chapter 4 describes the methods used to calculate the final household weights, and Section 5.1 shows some of the relationships between design weights, coverage and non-response adjusted weights and final weights. The coverage and non-response adjustment reduced the discrepancies between the census counts and the corresponding long-form estimates for the constraints considered (see Appendix C). Following those adjustments, calibration further reduced or eliminated those discrepancies for certain variables (constraints). However, some discrepancies remain, since constraints are sometimes excluded. The relative difference between census counts and long-form estimates, called the discrepancy, is defined as:

$$\text{Discrepancy} = \frac{(\text{long-form estimate} - \text{census count})}{\text{census count}} \times 100\%$$

This ratio represents the percentage that the characteristic was overestimated (a positive value) or underestimated (a negative value). For comparison reasons, it is also useful to look at the absolute values of the discrepancy and difference, hereafter referred to as the absolute discrepancy and absolute difference, respectively.

Table 5.2.1 shows the 2021 Canada-level differences between census counts and long-form estimates for the constraints considered for the design weights, the coverage and non-response adjusted weights and the final weights.

Table 5.2.1 also shows the discrepancy for estimates based on final weights. Looking at these discrepancies sheds more light on the differences. Over 94% of the cases in Table 5.2.1 had a discrepancy from -1% to 1%, and over 99.5% of them had discrepancies ranging from -5% to 5%.

Chart 5.2.2 shows, for all the constraints, the difference between the census counts and each of the three estimates: design weights (blue), coverage and non-response adjusted weights (orange), and final weights (green). The x-axis represents the population size of the constraint, in thousands, and the three series of dots show for each constraint:

- the difference between the sum of the design weights and the census count
- the difference between the sum of the coverage and non-response adjusted weights and the census count
- the difference between the sum of the final weights and the census count.

The constraints are sorted, from left to right, by increasing population size.

Chart 5.2.3 shows the percentage discrepancies between census counts and final estimates for all the constraints by population size. For the medium-sized and large-sized constraints, the discrepancies are all small. Only certain small-sized constraints have relatively large discrepancies.

The most important observation from Chart 5.2.2 is that the coverage and non-response adjustment carries a big improvement over the design-weighted estimates, in terms of reducing the discrepancy. Although it is not apparent in the chart, the coverage adjustment does most of the job. The difference between census counts and long-form estimates for design weights tended to be (much) greater than the difference between census counts and long-form estimates for the coverage and non-response adjusted weights. This, in turn, tended to be greater than the corresponding difference using the final weights. This shows the importance of the non-response adjustment and calibration processes. A difference between the census count and long-form estimate could occur in a SADA or ADA for a characteristic if its constraint is excluded during calibration. In other words, the process did not control on the excluded constraint for a given area. If the constraint is excluded in many areas, these differences could partially cancel each other out, or they could cumulate to create a large difference at the Canada level. Total persons (TOTPERS) and total households (TOTHHLD) were the only mandatory constraints for which agreement between census counts and long-form estimates had to be guaranteed for all ADAs. As a result, the final weight difference and discrepancy for these characteristics were 0. However, all other constraints had to be excluded in some areas.

Appendix C along with Table 5.2.1 illustrates that constraints that were excluded frequently tended to exhibit high differences or discrepancies. Looking at the constraints defined only at the SADA level, there was a positive relationship between the number of times a constraint was excluded and the absolute difference between census counts and long-form estimates. The “Persons in a couple (married or common-law)” (COUPLE) constraint exhibited the largest absolute difference of 8,642 while being the fourth most excluded constraint having been excluded 627 times between the coverage and non-response adjustment and final calibration. Lastly, among the SADA-only constraints with the top 10 largest relative differences, 7 constraints were also a part of the top 10 most excluded.

Across all possible constraints, the top 10 largest absolute differences were between 2,466 and 8,642. However, because the census counts were so high, the discrepancies for these constraints were small (ranging from -0.45% to 0.28%). The largest discrepancies were observed for some of the place of birth categories. Many place of birth categories are uncommon in Canada and are therefore frequently excluded during calibration. This resulted in some large absolute differences and particularly large absolute discrepancies.

Table 5.2.1
Census counts and long-form estimate differences and discrepancies, Canada

Characteristic	Census counts	Design weights		Coverage and non-response adjusted weights		Final weights		
		estimates	difference	estimates	difference	estimates	difference	discrepancy (%)
ADULTCF	18,645,248	18,561,008	-84,240	18,646,762	1,514	18,646,559	1,311	0.01
AGE00_14	5,863,246	5,825,360	-37,886	5,862,200	-1,046	5,863,461	215	0.00
AGE14	2,074,620	2,061,284	-13,336	2,075,061	441	2,074,101	-519	-0.03
AGE15_24	4,110,940	4,060,056	-50,884	4,109,533	-1,407	4,111,450	510	0.01
AGE15_29	6,479,017	6,395,416	-83,601	6,475,962	-3,055	6,479,062	45	0.00
AGE19	1,960,356	1,942,036	-18,320	1,965,980	5,624	1,965,564	5,208	0.27
AGE24	2,150,584	2,118,020	-32,564	2,143,553	-7,031	2,145,886	-4,698	-0.22

Table 5.2.1
Census counts and long-form estimate differences and discrepancies, Canada

		Design weights		Coverage and non-response adjusted weights		Final weights		
Characteristic	**Census counts**	estimates	difference	estimates	difference	estimates	difference	discrepancy (%)
AGE25_34	4,835,713	4,774,696	-61,017	4,833,889	-1,824	4,835,382	-331	-0.01
AGE29	2,368,077	2,335,360	-32,717	2,366,429	-1,648	2,367,612	-465	-0.02
AGE30_49	9,549,900	9,477,840	-72,060	9,547,899	-2,001	9,550,076	176	0.00
AGE34	2,467,636	2,439,336	-28,300	2,467,460	-176	2,467,770	134	0.01
AGE35_44	4,820,439	4,784,580	-35,859	4,819,469	-970	4,820,173	-266	-0.01
AGE39	2,464,600	2,441,428	-23,172	2,464,015	-585	2,464,706	106	0.00
AGE4	1,783,383	1,769,160	-14,223	1,782,623	-760	1,783,298	-85	0.00
AGE44	2,355,839	2,343,152	-12,687	2,355,454	-385	2,355,468	-371	-0.02
AGE45_54	4,582,978	4,553,408	-29,570	4,581,504	-1,474	4,582,969	-9	0.00
AGE49	2,261,825	2,253,924	-7,901	2,260,970	-855	2,262,132	307	0.01
AGE50_64	7,432,010	7,390,072	-41,938	7,431,294	-716	7,431,838	-172	0.00
AGE54	2,321,153	2,299,484	-21,669	2,320,534	-619	2,320,837	-316	-0.01
AGE55_64	5,110,857	5,090,588	-20,269	5,110,759	-98	5,111,001	144	0.00
AGE59	2,593,703	2,581,512	-12,191	2,593,232	-471	2,593,861	158	0.01
AGE64	2,517,154	2,509,076	-8,078	2,517,527	373	2,517,141	-13	0.00
AGE65PL	6,534,621	6,522,260	-12,361	6,535,110	489	6,534,356	-265	0.00
AGE74	3,958,758	3,953,664	-5,094	3,959,122	364	3,957,893	-865	-0.02
AGE75PL	2,575,863	2,568,596	-7,267	2,575,987	124	2,576,463	600	0.02
AGE9	2,005,243	1,994,916	-10,327	2,004,516	-727	2,006,063	820	0.04
APT5PLUS	1,593,869	1,585,684	-8,185	1,593,702	-167	1,593,648	-221	-0.01
APTLT5	2,728,558	2,713,976	-14,582	2,727,869	-689	2,728,625	67	0.00
CHILD	10,442,414	10,361,584	-80,830	10,442,904	490	10,443,251	837	0.01
CHILDFAM	5,884,937	5,846,660	-38,277	5,883,853	-1,084	5,884,257	-680	-0.01
COMLAWNO_DIV	1,859,786	1,849,532	-10,254	1,859,958	172	1,859,981	195	0.01
COMLAWNO_OTHERS	4,008,592	3,990,936	-17,656	4,008,187	-405	4,008,588	-4	0.00
COMLAWNO_SEP	716,514	711,864	-4,650	716,241	-273	716,757	243	0.03
COMLAWNO_SINGLE	14,508,945	14,350,816	-158,129	14,504,993	-3,952	14,508,422	-523	0.00
COMLAWNO_SINGLE_GE15	8,645,699	8,525,456	-120,243	8,642,793	-2,906	8,644,960	-739	-0.01
COMLAWNO_SINGLE_LT15	5,863,246	5,825,360	-37,886	5,862,200	-1,046	5,863,461	215	0.00

Table 5.2.1
Census counts and long-form estimate differences and discrepancies, Canada

		Design weights		Coverage and non-response adjusted weights		Final weights		
Characteristic	Census counts	estimates	difference	estimates	difference	estimates	difference	discrepancy (%)
COMLAWNO_WID	1,432,292	1,429,540	-2,752	1,431,989	-303	1,431,849	-443	-0.03
COMLAWYE_MARRIED	17,341,257	17,269,196	-72,061	17,339,283	-1,974	17,341,785	528	0.00
COMLAW_YE	3,836,830	3,812,728	-24,102	3,835,614	-1,216	3,835,160	-1,670	-0.04
COUPLE	17,005,592	16,940,552	-65,040	17,009,739	4,147	17,014,234	8,642	0.05
EMPIN_GT50	10,466,983	10,418,360	-48,623	10,468,477	1,494	10,466,948	-35	0.00
EMPIN_LE50	10,470,465	10,370,160	-100,305	10,466,820	-3,645	10,470,574	109	0.00
EMPIN_P0	14,921,346	14,822,428	-98,918	14,917,167	-4,179	14,921,272	-74	0.00
EMPIN_P0_GE15	9,058,100	8,997,068	-61,032	9,054,967	-3,133	9,057,810	-290	0.00
EMPIN_P0_LT15	5,863,246	5,825,360	-37,886	5,862,200	-1,046	5,863,461	215	0.00
EMPIN_P100	5,232,212	5,212,280	-19,932	5,231,362	-850	5,231,417	-795	-0.02
EMPIN_P25	5,237,240	5,184,568	-52,672	5,233,499	-3,741	5,238,205	965	0.02
EMPIN_P50	5,233,225	5,185,592	-47,633	5,233,321	96	5,232,369	-856	-0.02
EMPIN_P75	5,234,771	5,206,080	-28,691	5,237,115	2,344	5,235,531	760	0.01
EMPIN_SADA_GT50	10,468,377	10,421,676	-46,701	10,468,529	152	10,468,160	-217	0.00
EMPIN_SADA_LE50	10,469,071	10,366,844	-102,227	10,466,767	-2,304	10,469,363	292	0.00
EMPIN_SADA_P0	14,921,346	14,822,428	-98,918	14,917,167	-4,179	14,921,272	-74	0.00
EMPIN_SADA_P0_GE15	9,058,100	8,997,068	-61,032	9,054,967	-3,133	9,057,810	-290	0.00
EMPIN_SADA_P0_LT15	5,863,246	5,825,360	-37,886	5,862,200	-1,046	5,863,461	215	0.00
EMPIN_SADA_P100	5,234,004	5,216,312	-17,692	5,234,300	296	5,234,269	265	0.01
EMPIN_SADA_P25	5,235,162	5,181,240	-53,922	5,231,079	-4,083	5,235,190	28	0.00
EMPIN_SADA_P50	5,233,909	5,185,604	-48,305	5,235,688	1,779	5,234,173	264	0.01
EMPIN_SADA_P75	5,234,373	5,205,364	-29,009	5,234,229	-144	5,233,890	-483	-0.01
FEMALE	18,157,552	18,051,988	-105,564	18,155,596	-1,956	18,157,618	66	0.00
FEMALEGE15	15,303,878	15,217,160	-86,718	15,302,343	-1,535	15,303,972	94	0.00
FEMALELT15	2,853,674	2,834,828	-18,846	2,853,253	-421	2,853,646	-28	0.00
HHADA	14,826,894	14,768,580	-58,314	14,826,894	0	14,826,894	0	0.00
HHADACSD	14,826,894	14,768,580	-58,314	14,826,894	0	14,826,894	0	0.00
HHINC_GT50	7,412,136	7,372,056	-40,080	7,414,714	2,578	7,412,314	178	0.00
HHINC_LE50	7,414,758	7,396,524	-18,234	7,412,180	-2,578	7,414,580	-178	0.00

Table 5.2.1
Census counts and long-form estimate differences and discrepancies, Canada

	Census	Design weights		Coverage and non-response adjusted weights		Final weights		
Characteristic	**counts**	estimates	difference	estimates	difference	estimates	difference	discrepancy (%)
HHINC_P100	3,704,978	3,678,564	-26,414	3,703,026	-1,952	3,704,942	-36	0.00
HHINC_P25	3,708,673	3,704,396	-4,277	3,708,008	-665	3,708,664	-9	0.00
HHINC_P50	3,706,085	3,692,128	-13,957	3,704,172	-1,913	3,705,915	-170	0.00
HHINC_P75	3,707,158	3,693,492	-13,666	3,711,687	4,529	3,707,372	214	0.01
HHINC_SADA_GT50	7,413,255	7,373,352	-39,903	7,413,008	-247	7,413,093	-162	0.00
HHINC_SADA_LE50	7,413,639	7,395,228	-18,411	7,413,886	247	7,413,801	162	0.00
HHINC_SADA_P100	3,706,522	3,682,152	-24,370	3,706,067	-455	3,706,454	-68	0.00
HHINC_SADA_P25	3,706,961	3,702,248	-4,713	3,707,028	67	3,706,238	-723	-0.02
HHINC_SADA_P50	3,706,678	3,692,980	-13,698	3,706,857	179	3,707,564	886	0.02
HHINC_SADA_P75	3,706,733	3,691,200	-15,533	3,706,941	208	3,706,638	-95	0.00
HHSIZE1	4,356,317	4,363,644	7,327	4,356,599	282	4,354,798	-1,519	-0.03
HHSIZE2	5,086,584	5,065,324	-21,260	5,087,377	793	5,086,686	102	0.00
HHSIZE3	2,172,884	2,160,960	-11,924	2,172,802	-82	2,172,690	-194	-0.01
HHSIZE4	1,983,642	1,970,564	-13,078	1,984,050	408	1,984,337	695	0.04
HHSIZE5	777,142	770,908	-6,234	779,026	1,884	779,459	2,317	0.30
HHSIZEGE5	1,227,467	1,208,088	-19,379	1,226,066	-1,401	1,228,382	915	0.07
HHSIZEGE6	450,325	437,180	-13,145	447,040	-3,285	448,924	-1,401	-0.31
INEFAM	29,985,843	29,806,124	-179,719	29,989,178	3,335	29,991,606	5,763	0.02
IR_LINK_NO	35,456,960	35,217,972	-238,988	35,451,968	-4,992	35,457,267	307	0.00
IR_LINK_YE	401,834	392,976	-8,858	400,496	-1,338	401,527	-307	-0.08
LIM_NO	31,967,342	31,741,028	-226,314	31,962,884	-4,458	31,966,263	-1,079	0.00
LIM_YE	3,891,452	3,869,920	-21,532	3,889,579	-1,873	3,892,531	1,079	0.03
LONEPAR	1,639,656	1,620,456	-19,200	1,637,023	-2,633	1,632,325	-7,331	-0.45
MALE	17,701,242	17,558,960	-142,282	17,696,868	-4,374	17,701,176	-66	0.00
MALEGE15	14,691,670	14,568,428	-123,242	14,687,920	-3,750	14,691,361	-309	0.00
MALELT15	3,009,572	2,990,532	-19,040	3,008,948	-624	3,009,815	243	0.01
MARRIED	13,504,427	13,456,468	-47,959	13,503,668	-759	13,506,625	2,198	0.02
NB_NOTINCF	6,771,132	6,688,356	-82,776	6,762,799	-8,333	6,768,984	-2,148	-0.03
NOCLDFAM	4,257,515	4,244,072	-13,443	4,258,039	524	4,255,185	-2,330	-0.05

Table 5.2.1
Census counts and long-form estimate differences and discrepancies, Canada

Characteristic	Census counts	Design weights		Coverage and non-response adjusted weights		Final weights		
		estimates	difference	estimates	difference	estimates	difference	discrepancy (%)
NOINEFAM	5,872,951	5,804,824	-68,127	5,863,285	-9,666	5,867,188	-5,763	-0.10
NOINEFAMHHSIZEEQ1	4,356,317	4,363,644	7,327	4,356,599	282	4,354,798	-1,519	-0.03
NOINEFAMHHSIZEGT1	1,516,634	1,441,180	-75,454	1,506,687	-9,947	1,512,390	-4,244	-0.28
NOTINFAM	6,771,132	6,688,356	-82,776	6,762,799	-8,333	6,768,984	-2,148	-0.03
NOTINFAMHHSIZEEQ1	4,356,317	4,363,644	7,327	4,356,599	282	4,354,798	-1,519	-0.03
NOTINFAMHHSIZEGT1	2,414,815	2,324,712	-90,103	2,406,200	-8,615	2,414,186	-629	-0.03
OLN_BI	6,523,298	6,482,368	-40,930	6,523,158	-140	6,524,917	1,619	0.02
OLN_EN	24,660,168	24,472,152	-188,016	24,656,197	-3,971	24,658,926	-1,242	-0.01
OLN_FR	3,996,708	3,982,912	-13,796	3,996,280	-428	3,996,783	75	0.00
OLN_NO	678,620	673,516	-5,104	676,829	-1,791	678,168	-452	-0.07
OTHERDTYPE	2,742,490	2,728,540	-13,950	2,742,407	-83	2,742,753	263	0.01
POBG2_1	18,188	18,436	248	18,747	559	18,398	210	1.15
POBG2_10	69,328	68,104	-1,224	68,893	-435	69,005	-323	-0.47
POBG2_11	100,824	100,972	148	100,654	-170	100,771	-53	-0.05
POBG2_16	39,401	39,044	-357	39,043	-358	39,386	-15	-0.04
POBG2_17	37,600	36,964	-636	36,661	-939	37,076	-524	-1.39
POBG2_18	4,705	4,568	-137	4,571	-134	4,660	-45	-0.96
POBG2_19	26,175	25,972	-203	26,157	-18	26,087	-88	-0.34
POBG2_20	27,622	27,504	-118	27,542	-80	27,646	24	0.09
POBG2_21	973,117	971,244	-1,873	974,259	1,142	973,913	796	0.08
POBG2_22	138,803	138,148	-655	140,247	1,444	140,100	1,297	0.93
POBG2_24	45,496	44,620	-876	44,556	-940	44,886	-610	-1.34
POBG2_25	24,018	23,724	-294	23,849	-169	23,699	-319	-1.33
POBG2_26	90,748	90,332	-416	91,084	336	91,009	261	0.29
POBG2_27	70,378	69,804	-574	70,429	51	70,554	176	0.25
POBG2_28	159,482	157,392	-2,090	158,758	-724	159,151	-331	-0.21
POBG2_29	87,715	88,556	841	88,327	612	88,550	835	0.95
POBG2_3	39,318	39,292	-26	39,516	198	39,455	137	0.35
POBG2_30	43,855	43,428	-427	44,366	511	44,233	378	0.86

Table 5.2.1
Census counts and long-form estimate differences and discrepancies, Canada

	Census counts	Design weights		Coverage and non-response adjusted weights		Final weights		
Characteristic		estimates	difference	estimates	difference	estimates	difference	discrepancy (%)
POBG2_31	145,516	145,352	-164	146,258	742	145,899	383	0.26
POBG2_32	208,157	207,616	-541	208,788	631	208,633	476	0.23
POBG2_33	133,732	133,192	-540	133,176	-556	133,053	-679	-0.51
POBG2_34	24,611	24,160	-451	24,537	-74	24,468	-143	-0.58
POBG2_35	75,006	74,360	-646	75,190	184	75,623	617	0.82
POBG2_36	82,061	80,744	-1,317	81,632	-429	82,056	-5	-0.01
POBG2_37	135,441	133,692	-1,749	134,832	-609	135,038	-403	-0.30
POBG2_38	97,604	97,816	212	98,419	815	98,205	601	0.62
POBG2_39	33,558	33,732	174	33,517	-41	33,587	29	0.09
POBG2_4	69,501	69,680	179	69,323	-178	69,454	-47	-0.07
POBG2_40	147,988	148,988	1,000	147,968	-20	147,920	-68	-0.05
POBG2_41	2,707	2,676	-31	2,736	29	2,713	6	0.21
POBG2_42	207,251	203,500	-3,751	206,111	-1,140	207,270	19	0.01
POBG2_43	759,124	760,536	1,412	761,959	2,835	761,199	2,075	0.27
POBG2_45	93,239	92,184	-1,055	93,552	313	93,699	460	0.49
POBG2_46	73,961	73,020	-941	73,263	-698	73,343	-618	-0.84
POBG2_47	182,036	179,428	-2,608	180,255	-1,781	180,423	-1,613	-0.89
POBG2_48	453,197	451,084	-2,113	454,014	817	453,697	500	0.11
POBG2_50	1,117	1,104	-13	1,121	4	1,091	-26	-2.29
POBG2_51	20,746	20,264	-482	20,457	-289	20,452	-294	-1.42
POBG2_54	114,417	114,916	499	114,900	483	114,799	382	0.33
POBG2_55	21,665	21,680	15	22,065	400	21,929	264	1.22
POBG2_56	87,057	86,636	-421	87,125	68	87,376	319	0.37
POBG2_57	31,161	30,388	-773	30,862	-299	30,901	-260	-0.84
POBG2_59	50,363	51,048	685	51,538	1,175	51,260	897	1.78
POBG2_6	38,714	38,148	-566	38,503	-211	38,242	-472	-1.22
POBG2_60	129,348	128,936	-412	129,813	465	129,745	397	0.31
POBG2_63	12,109	12,044	-65	11,964	-145	12,044	-65	-0.54
POBG2_64	1,282,470	1,274,528	-7,942	1,284,077	1,607	1,281,244	-1,226	-0.10

Table 5.2.1
Census counts and long-form estimate differences and discrepancies, Canada

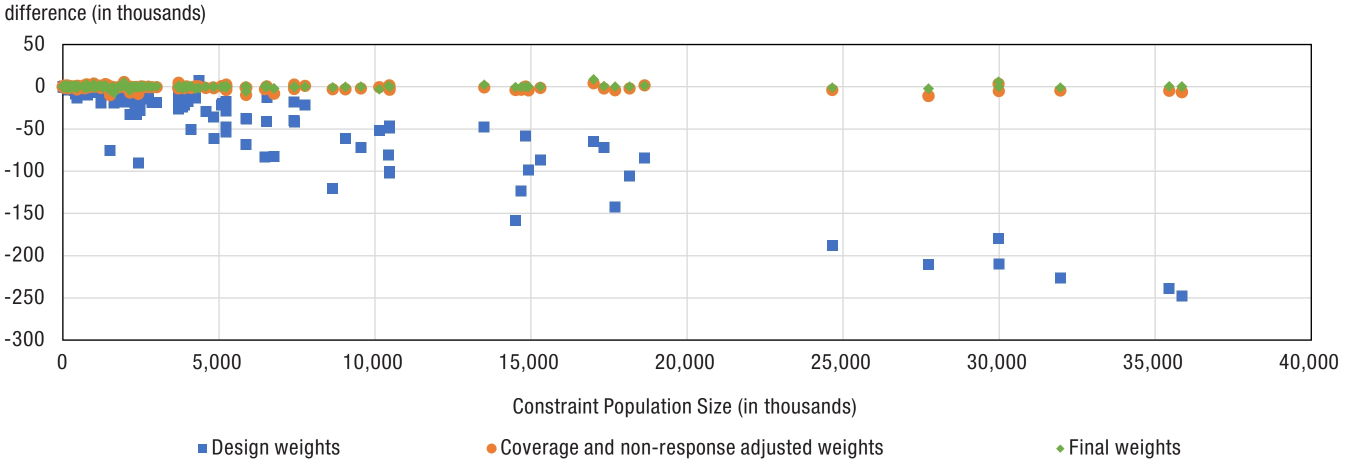
Characteristic	Census counts	Design weights		Coverage and non-response adjusted weights		Final weights		
		estimates	difference	estimates	difference	estimates	difference	discrepancy (%)
POBG2_65	33,796	33,840	44	33,947	151	33,820	24	0.07
POBG2_66	340,472	340,244	-228	340,373	-99	340,034	-438	-0.13
POBG2_67	225,939	224,856	-1,083	225,835	-104	226,274	335	0.15
POBG2_68	30,781	29,976	-805	30,454	-327	30,502	-279	-0.91
POBG2_69	132,839	131,732	-1,107	133,079	240	132,934	95	0.07
POBG2_7	77,130	76,736	-394	76,908	-222	76,934	-196	-0.25
POBG2_70	62,339	61,204	-1,135	61,677	-662	61,229	-1,110	-1.78
POBG2_71	108,119	107,468	-651	108,495	376	108,324	205	0.19
POBG2_9	1,800	1,896	96	1,906	106	1,942	142	7.91
POBG3_10	356,921	356,844	-77	356,937	16	356,600	-321	-0.09
POBG3_12	60,983	60,972	-11	61,581	598	61,384	401	0.66
POBG3_14	372,701	368,480	-4,221	373,187	486	373,777	1,076	0.29
POBG3_15	991,957	992,892	935	995,727	3,770	994,720	2,763	0.28
POBG3_16	52,163	52,944	781	53,444	1,281	53,202	1,039	1.99
POBG3_17	1,436,629	1,427,828	-8,801	1,438,445	1,816	1,435,473	-1,156	-0.08
POBG3_18	333,700	330,792	-2,908	333,977	277	334,217	517	0.15
POBG3_19	238,714	236,492	-2,222	238,111	-603	238,401	-313	-0.13
POBG3_2	371,246	369,500	-1,746	371,278	32	371,108	-138	-0.04
POBG3_20	225,939	224,856	-1,083	225,835	-104	226,274	335	0.15
POBG3_21	197,885	195,612	-2,273	197,493	-392	196,876	-1,009	-0.51
POBG3_22	269,123	268,664	-459	269,484	361	268,968	-155	-0.06
POBG3_3	81,706	80,576	-1,130	80,276	-1,430	81,123	-583	-0.71
POBG3_4	188,533	187,180	-1,353	188,954	421	188,989	456	0.24
POBG3_5	230,116	228,392	-1,724	230,849	733	230,586	470	0.20
POBG3_6	1,156,486	1,155,964	-522	1,157,691	1,205	1,157,397	911	0.08
POBG3_7	260,737	260,508	-229	260,109	-628	260,456	-281	-0.11
POBG3_8	709,194	703,532	-5,662	707,532	-1,662	707,463	-1,731	-0.24
POBG3_9	294,966	292,056	-2,910	294,438	-528	295,820	854	0.29
PPADA	35,858,794	35,610,948	-247,846	35,852,464	-6,330	35,858,794	0	0.00

Table 5.2.1
Census counts and long-form estimate differences and discrepancies, Canada

		Design weights		Coverage and non-response adjusted weights		Final weights		
Characteristic	**Census counts**	estimates	difference	estimates	difference	estimates	difference	discrepancy (%)
PPADACSD	35,858,794	35,610,948	-247,846	35,852,464	-6,330	35,858,794	0	0.00
SNGLDET	7,761,977	7,740,380	-21,597	7,762,916	939	7,761,868	-109	0.00
TOTCFAM	10,142,452	10,090,732	-51,720	10,141,892	-560	10,139,442	-3,010	-0.03
TOTHHLD	14,826,894	14,768,580	-58,314	14,826,894	0	14,826,894	0	0.00
TOTPERS	35,858,794	35,610,948	-247,846	35,852,464	-6,330	35,858,794	0	0.00
TPERGE15	29,995,548	29,785,588	-209,960	29,990,263	-5,285	29,995,333	-215	0.00
TPERLT15	5,863,246	5,825,360	-37,886	5,862,200	-1,046	5,863,461	215	0.00
YRIMD_1900	1,057,784	1,051,708	-6,076	1,057,998	214	1,057,899	115	0.01
YRIMD_1980	336,657	335,244	-1,413	335,330	-1,327	336,180	-477	-0.14
YRIMD_1986	478,539	477,856	-683	479,476	937	479,823	1,284	0.27
YRIMD_1991	704,709	705,148	439	706,313	1,604	705,588	879	0.12
YRIMD_1996	664,518	662,160	-2,358	665,920	1,402	665,189	671	0.10
YRIMD_2001	834,315	830,188	-4,127	833,798	-517	833,822	-493	-0.06
YRIMD_2006	938,128	935,132	-2,996	938,038	-90	938,252	124	0.01
YRIMD_2011	1,045,048	1,038,256	-6,792	1,045,197	149	1,044,095	-953	-0.09
YRIMD_2016	1,265,550	1,261,424	-4,126	1,265,880	330	1,265,578	28	0.00
YRIMD_M3	27,740,538	27,530,424	-210,114	27,729,416	-11,122	27,738,072	-2,466	-0.01
YRIMD_M5	793,008	783,408	-9,600	795,099	2,091	794,295	1,287	0.16
YRIMG1_1900	1,057,784	1,051,708	-6,076	1,057,998	214	1,057,899	115	0.01
YRIMG1_1980	815,196	813,100	-2,096	814,806	-390	816,003	807	0.10
YRIMG1_1991	1,369,227	1,367,308	-1,919	1,372,232	3,005	1,370,777	1,550	0.11
YRIMG1_2001	1,772,443	1,765,320	-7,123	1,771,836	-607	1,772,075	-368	-0.02
YRIMG1_2011	1,045,048	1,038,256	-6,792	1,045,197	149	1,044,095	-953	-0.09
YRIMG1_2016	1,265,550	1,261,424	-4,126	1,265,880	330	1,265,578	28	0.00
YRIMG1_M3	27,740,538	27,530,424	-210,114	27,729,416	-11,122	27,738,072	-2,466	-0.01
YRIMG1_M5	793,008	783,408	-9,600	795,099	2,091	794,295	1,287	0.16

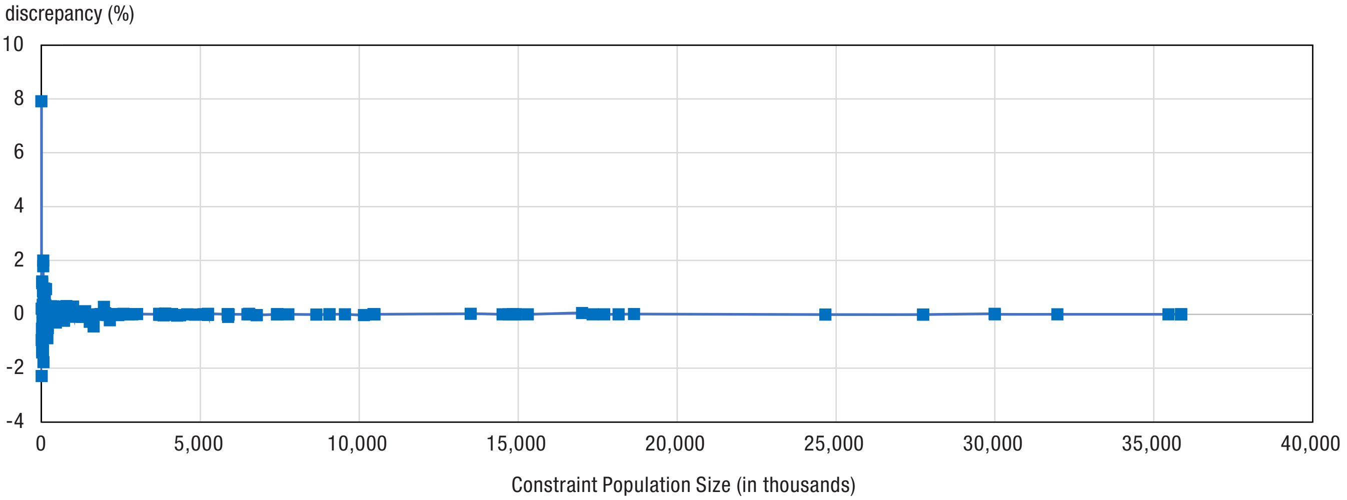
Note: All households with a design weight of 1 were excluded from the weighting process. These households either were located in First Nations communities, Métis Settlements, Inuit regions or other remote areas, or were private households attached to a collective dwelling.
Source: Statistics Canada, 2021 Census long-form sample.

Chart 5.2.2
Differences between census counts and counts estimated using design, coverage and non-response adjusted, and final weights



Note: All households with a design weight of 1 were excluded from the weighting process. These households were either located in First Nations communities, Métis Settlements, Inuit regions or other remote areas, or were private households attached to a collective dwelling.
Source: Statistics Canada, 2021 Census long-form sample.

Chart 5.2.3
Discrepancy between census counts and final estimates, as a percentage of census counts



Note: All households with a design weight of 1 were excluded from the weighting process. These households either were located in First Nations communities, Métis Settlements, Inuit regions or other remote areas, or were private households attached to a collective dwelling.
Source: Statistics Canada, 2021 Census long-form sample.

6. Variance estimation

The error in an estimate is the difference between the estimate and the actual value of what is being estimated. All estimates from the census questionnaires are subject to non-sampling errors such as total non-response error. Estimates from the long-form questionnaire are also subject to sampling error. Sampling error stems from the fact that the estimates are based on observations from a sample and not from the Census of Population. Total non-response error occurs when households selected in the sample do not respond to the survey.

The error in an estimate has a random component, measured by variance, and a systematic component, measured by bias. Variance measures how much the estimate varies from the average that would result from hypothetical repetitions of the survey process. Variance can be estimated using data from the sample. Bias is the difference between the average estimate that would result from hypothetical repetitions of the survey process and the actual value of the characteristic being estimated. The sampling and estimation methods used in the long-form sample survey all aim to minimize the bias.

Some estimation methods are more precise than others in estimating a particular characteristic of the population, so they can affect error. The estimated variance can be used to produce several quality indicators that are often used to measure the accuracy of an estimate. For example, it can be used to calculate standard errors, confidence intervals and coefficients of variation. The confidence interval was selected as a variance-based quality indicator to support the 2021 Census of Population long-form estimates, because it helps users easily make a statistical inference. Confidence intervals therefore generally accompany long-form estimates in the 2021 Census data products.

These measures of variability must be carefully distinguished from other measures of quality that are not, strictly speaking, measures of variability. Examples of such measures are the final response rates presented in Section 3.11, or item imputation rates. For more information, see Chapter 9 of the *Guide to the Census of Population, 2021,* Statistics Canada Catalogue no. 98-304-X and the *2021 Census Data quality Guidelines*, Statistics Canada Catalogue no. 98-26-0006.

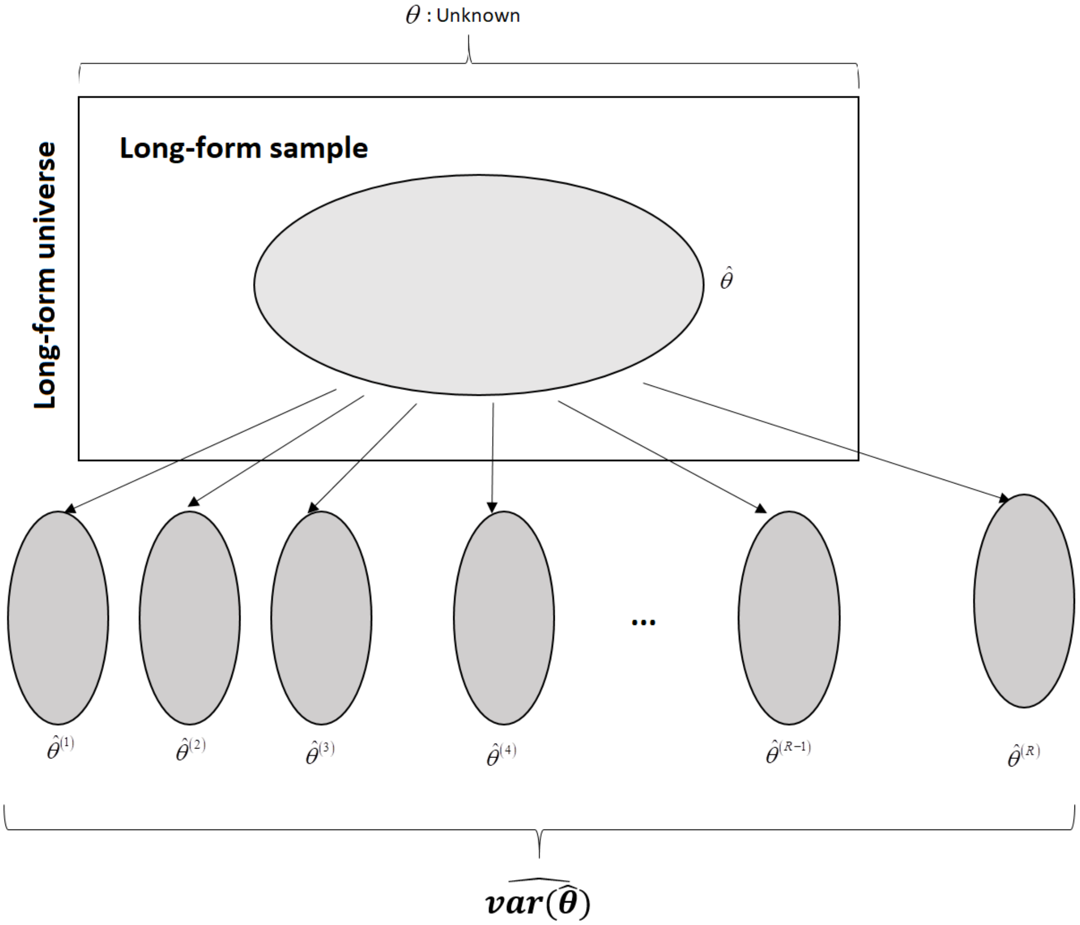
Since the long-form sample is geographically stratified into take-some strata (mail-out, list/leave and mail-out with drop-off CUs) and take-all strata (CUs in First Nations communities, Métis settlements, Inuit regions and other remote areas), two variance estimators are used. The first variance estimator is used to estimate the variance in take-some geographic areas (see Section 6.3.1), and the second estimator is used to estimate the variance attributable to total non-response in take-all areas (see Section 6.3.2). For the remainder of this chapter, the term variance is used to designate the sampling and total non-response variance in take-some geographic areas or the total non-response variance in take-all areas.

6.1 Elements to consider in choosing a variance estimation method

A very high number of diverse estimates were produced, and quality indicators for these estimates needed to be established within a reasonable time frame. As a result, a resampling variance estimator was used, which was derived from the modified partially balanced repeated replication method (Judkins 1990). This method was first introduced with the 2016 Census and was largely maintained for the 2021 Census. The method consists of drawing samples (or replicates) from the original sample. Weights are calculated for each replicate, and the weights undergo the same coverage, non-response and calibration adjustments as the original sample. Henceforth, the weights corresponding to the original sample are called the main weights. The weights resulting from each replicate sample are called replicate weights. Estimates are then produced for each replicate, and the variance is estimated using replicate estimates and the main weight estimate.

Figure 6.1.1 gives an overview of replication variance estimation when R samples are used.

Figure 6.1.1
Overview of replication variance estimation



Source: Statistics Canada.

Full description:

Figure 6.1.1 gives an overview of the replication variance estimation methodology used in the 2021 Census. The replication variance estimation method simulates the selection of several samples to estimate sampling variance.

More specifically, the figure shows the long-form questionnaire universe representing the population of interest and the long-form questionnaire sample. The sample is situated within the universe to indicate that it corresponds to a subset of the population of interest. This sample is used to estimate a characteristic of the population of interest, such as the number of persons who are members of a visible minority.[8] The theta symbol is used to represent the true value of this characteristic. A circumflex on the theta indicates that the value is an estimate of this characteristic. This value is known as theta hat.

The R other samples placed outside the universe are linked to the long-form questionnaire sample with arrows. The arrows indicate that these samples are taken from the long-form questionnaire sample. The characteristic of interest is re-estimated based on these R subsamples. The R theta hat values, referred to as theta hat one, theta hat two, up to theta hat R, are used to calculate the estimated theta hat variance.

The following are defined:

- theta, θ, the true value of the characteristic in the population, which can be a total, an average, a quantile, etc.
- theta hat, $\hat{\theta}$, the value of θ estimated using the main weights
- theta hat r, $\hat{\theta}^{(r)}$, the value of θ estimated using the replication weights r, $r = 1, ..., R$
- theta hat bar, $\overline{\hat{\theta}}$, the average value of the R replication estimates $\hat{\theta}^{(r)}$
- $\widehat{\text{var}}(\hat{\theta})$, the estimated value of the variance of $\hat{\theta}$.

6.2 Variance estimator

The replicate estimator chosen for the long-form sample survey was derived from Fay's balanced half-sample method (Judkins 1990). This method determines the creation of replicates, the calculation of replicate weights and the multiplication factor used to estimate variance.

To produce variance estimates for the long-form sample estimates, two sets of replicate weights were created: the first had 32 replicate weights and the second had 100 replicate weights. The set of 32 replicate weights was produced to estimate the standard errors of standard products that are calculated under operational constraints (i.e., the need to publish a large number of confidence intervals within a reasonable time frame). The set of 100 replicate weights was made available to Statistics Canada analysts and research data centre analysts who have access to microdata to provide more precise variance estimators.

[8] In 2021 Census analytical and communications products, the term "visible minority" has been replaced by the terms "racialized population" or "racialized groups," reflecting the increased use of these terms in the public sphere.

The replication variance estimator can be calculated in two ways, one of which is more conservative than the other. The first method consists of taking the sum of the squared differences between the replication estimates, $\hat{\theta}^{(r)}$, and the average of the replication estimates, $\bar{\hat{\theta}}$. The second method consists of taking the sum of squared differences between the replication estimates, $\hat{\theta}^{(r)}$, and the estimate from the original sample, $\hat{\theta}$. With both methods, the sum of squared differences is multiplied by a certain factor. The second method, which uses the estimate from the primary sample, is more conservative. In the computer system used to publish statistics, the variance estimator is calculated using the estimate from the primary sample.

For example, two variance estimators of an estimator $\hat{T}$ of a total T from a set of R replicates are given in the equations below:

$$\widehat{\mathrm{Var}}_1\left(\hat{T}\right)=\frac{1}{R/2}\sum_{r=1}^{\mathrm{R}}\left(\hat{T}^{(r)}-\bar{\hat{T}}\right)^2,$$

$$\widehat{\mathrm{Var}}_2\left(\hat{T}\right)=\frac{1}{R/2}\sum_{r=1}^{\mathrm{R}}\left(\hat{T}^{(r)}-\hat{T}\right)^2$$

where

$$\hat{T}=\sum\nolimits_{k\in s}w_k y_k \quad ,$$

$$\hat{T}^{(r)}=\sum\nolimits_{k\in s}w_k^{(r)} y_k \quad \text{, and}$$

$$\bar{\hat{T}}=\sum\nolimits_{r=1}^{R}\hat{T}^{(r)}\Big/\mathrm{R} \quad .$$

The final weight of the sample is represented by w_k, $w_k^{(r)}$ is the final weight of replicate r, y_k is the value of characteristic y for unit k, and s is the long-form sample.

The number of degrees of freedom of the variance estimator is approximated by the number of squared differences $\left(\hat{T}^{(r)}-\bar{\hat{T}}\right)^2$ for the variance estimator, i.e., 32 or 100. The number of degrees of freedom gives an idea of the precision of the variance estimator and is used in calculating confidence intervals for long-form estimates. See Chapter 7 for more details.

6.3 Replicate weight adjustment

6.3.1 Mail-out and list/leave collection units

As mentioned in Section 6.2, replicate weights were calculated for all long-form sample households. The replicates were partially balanced. They were balanced by resampling strata, which were created by combining CUs to obtain 600 to 1,800 households per resampling stratum.

Fay's modified balanced half-sample method, as described by Rao and Shao (1999), requires an epsilon value in the calculation of replicate weights to control the perturbation of the replicate weights. This perturbation results in all sampled households participating in every replicate, unlike other more popular replication methods. This facilitates the calibration of the replicate weights and, occasionally, the calculation of point estimates for each replicate (e.g., the denominator of a ratio estimator for a given replicate will not have a nil value if the corresponding denominator was not nil with the final weight). Adding an epsilon factor to the calculation of replicate weights meant the large survey fraction used to select the long-form sample could be taken into account. The technical details of the variance estimation process were provided by Devin and Verret (2016).

The replicate weights underwent the same adjustments as the primary sample design weight. They were adjusted for coverage and total non-response following the same methodology that was used for the primary sample weight (see Section 4.4). The resulting replicate weights were then calibrated to census counts, once again following the same methodology that was used for the main weight (see Section 4.5).

6.3.2 Collection units in First Nations communities, Métis settlements, Inuit regions and other remote areas[9]

As described in Chapter 2, all the households in First Nations communities, Métis settlements, Inuit regions and other remote areas CUs were selected with certainty. As such, they originally had a design weight equal to 1. A coverage adjustment was not needed. All these households were selected for the long-form questionnaire, and therefore differential coverage between the short-form and long-form questionnaire could not occur. Total non-response in these areas was treated with the process of whole household imputation (WHI), described in Chapter 3. In other words, the data of a non-responding household were replaced by the data of another responding household from the same CU (except for geography variables, which were known for non-respondents). As a consequence, reweighting for households in those CUs was not needed.

Calibration was not needed in these areas, because the long-form questionnaire was a census. Consequently, all households in First Nations communities, Métis settlements, Inuit regions and other remote areas CUs maintained their original weight equal to 1 in the final weighting scheme.

Although sampling variability did not occur for households in - First Nations communities, Métis settlements, Inuit regions and other remote areas CUs, WHI variability did occur. Variance estimation in these areas was computed using a similar method to that of the rest of the country, with the following exceptions. First, the response probability by household size combination in each census division was estimated as the number of responding households divided by the number of in-scope households. Then, the base replicate weights were created as in the rest of the country, except that all respondents for which the response probability was equal to 1 were placed in every replicate. Respondents with estimated response probabilities less than 1 were not considered certainties and were treated as sampled elements (i.e., they were randomly divided among the replicates). Non-respondent households imputed by WHI were also divided among replicates, and each one was assigned the replicate inclusion indicator corresponding to its donor in a manner similar to that of Shao and Tang (2011). This caused the weights to vary from one replicate to the other. Finally, the replicate weights were calibrated to the number of households and number of persons in the SADA. As a result, the estimated variance of those two quantities was equal to 0 at the SADA level and at more aggregate levels, such as Canada (since those two constraints are mandatory in the rest of the country).

[9] The exception to this characteristic was the units in incompletely enumerated reserves and settlements, which were excluded from the target population and whose weight was set to 0, without any further modification to the dataset or weights. For more information on incompletely enumerated reserves and settlements, please refer to Appendix 1.5 of the *Guide to the Census of Population, 2021,* Statistics Canada Catalogue no. 98-304-X.

7. Statistical inference

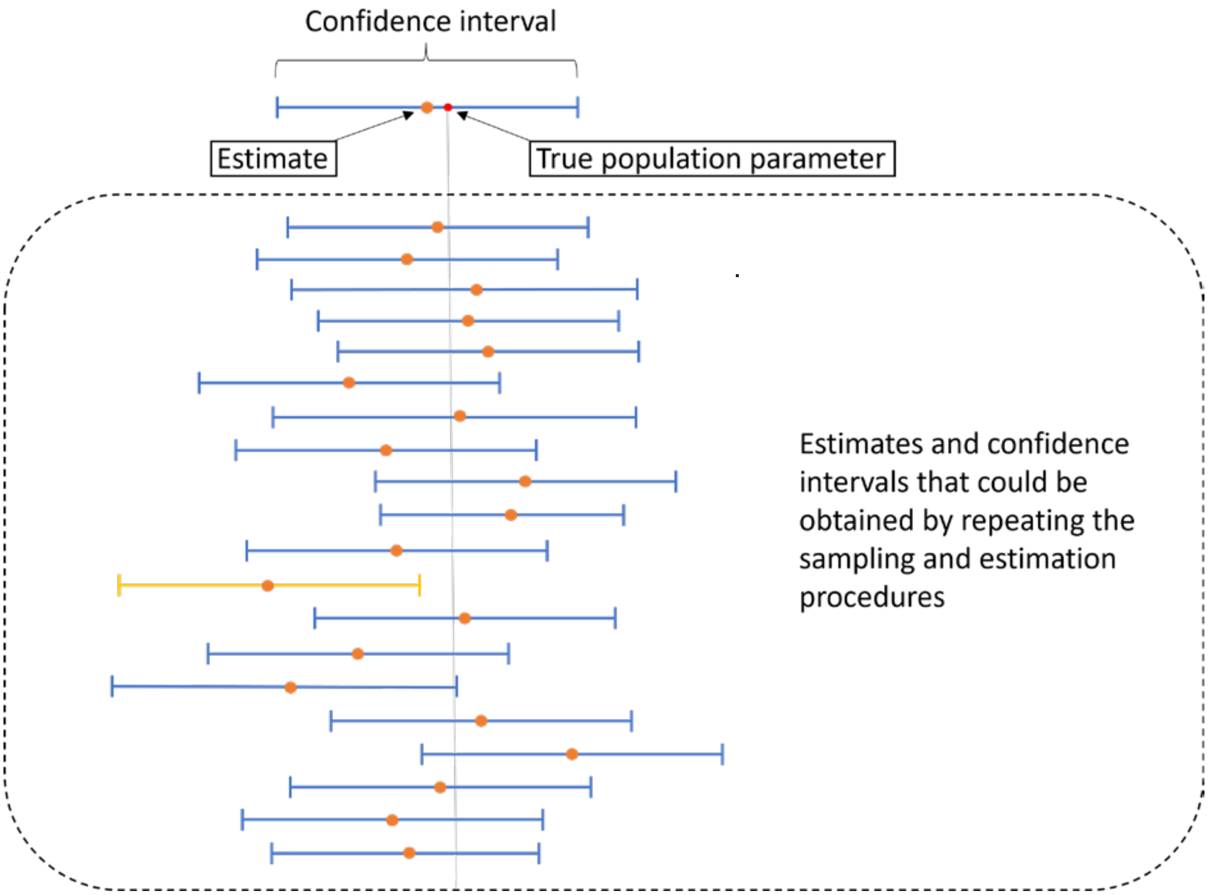
Statistical inference in the context of survey sampling is the process of drawing conclusions about the population based on data collected from survey respondents. To draw conclusions about the target population of the long-form sample survey using estimates produced from the sample, the uncertainty of these estimates must be taken into consideration. As described in Chapter 6, estimates produced from the long-form sample are subject to variability because of sampling and non-response. The variance of each estimate is a measure that quantifies this variability. Estimates of the variance of survey statistics can be used to produce other measures of the quality of the statistics that reflect their variability, but are more easily interpreted than the variance estimates themselves. These measures include standard errors, coefficients of variation and confidence intervals. Among these measures, confidence intervals have the advantage that they allow users to easily perform statistical inferences. For this reason, the confidence interval was chosen as the variance-based quality indicator to accompany long-form estimates for the 2021 Census.

7.1 Confidence intervals and their interpretation

A confidence interval for an estimate is an interval constructed around the estimate that reflects the estimate's uncertainty. A confidence interval is associated with a confidence level, which is expressed as a percentage. The confidence level describes the degree to which one can be confident that the true population parameter is contained in the confidence interval. A default confidence level is generally set for a survey or in a field of study based on user needs. A commonly used confidence level is 95%, and this is the default confidence level for the census dissemination system. Given an estimate together with a 95% confidence interval for the estimate, a user can infer with 95% confidence that the true population parameter is contained within the interval.

A rigorous interpretation of confidence intervals relies on hypothetically repeating the sampling and estimation procedures. This interpretation is illustrated in Figure 7.1.1.

Figure 7.1.1
Interpretation of confidence intervals


Confidence interval
Estimate
True population parameter
Estimates and confidence intervals that could be obtained by repeating the sampling and estimation procedures

Source: Statistics Canada, 2021 Census of Population.

Full description:

The top part of Figure 7.1.1 depicts a confidence interval for the estimate of a population parameter. This particular confidence interval contains the true population parameter.

The figure also shows different estimates that could be produced by hypothetically repeating the sampling and estimation procedures several times, together with their corresponding confidence intervals. In the case of the long-form questionnaire, this repeated sampling and estimation would involve drawing a very large number of samples from the long-form sample universe according to the sampling design described in Chapter 2. Each one of the samples would undergo the same processing, weighting and estimation steps as the actual sample. The estimates and the confidence intervals produced for a given characteristic would generally be different for the different samples. However, if the underlying assumptions of the given confidence interval method are valid, the percentage of the confidence intervals that contain the true population value would be approximately equal to the confidence level.

In the example depicted in the figure, all but one of the confidence intervals contain the true population parameter.

The width of the confidence interval for an estimate is an indication of the degree of uncertainty of the estimate. If two estimates are equal, but one has a wider 95% confidence interval than the other, then the estimate with the wider interval has greater uncertainty.

The types of uncertainty reflected in the confidence intervals for long-form estimates differ according to the stratum type. Since confidence intervals are based on variance estimates, they reflect the same types of uncertainty as the underlying variance estimates. In mail-out, list-leave, and mail-out with drop-off collection units (CUs), where the sampling fraction is one-quarter, the uncertainty measured is due to sampling and total non-response. In CUs in First Nations communities, Métis settlements, Inuit regions and other remote areas, where all households are sampled, the uncertainty measured is due only to total non-response.

7.2 Constructing confidence intervals

Confidence intervals are generally based on the properties of a mathematical expression called a pivot. Depending on the pivot used and on which assumptions are made about the properties of the pivot, different types of confidence intervals will result from it. When constructing a confidence interval for a population parameter θ , the pivot used will typically involve the estimate of theta, denoted by $\hat{\theta}$, and the standard error of $\hat{\theta}$. The standard error of an estimate is defined to be the square root of the estimated variance of the estimate. The standard error of $\hat{\theta}$ is denoted by $\widehat{SE}(\hat{\theta})$. The pivot may also involve additional quantities. The most commonly used type of confidence interval, the Wald interval, is based on the following pivot:

$$\frac{\hat{\theta}-\theta}{\widehat{SE}(\hat{\theta})}\,.$$

The assumption underlying the Wald interval is that this pivot approximately follows a standard normal distribution, i.e., a normal distribution with a mean of zero and a standard deviation of one. The lower bound (LB) and upper bound (UB) of the 95% Wald confidence interval for the population parameter θ are given by:

$$LB = \hat{\theta} - z \times \widehat{SE}(\hat{\theta}),$$

$$UB = \hat{\theta} + z \times \widehat{SE}(\hat{\theta}),$$

where z is the 97.5th percentile of the standard normal distribution.

In many situations, the assumption that the Wald pivot is approximately normal is not satisfied. When the assumptions underlying the construction of a confidence interval are violated, this can lead to undercoverage of the confidence interval. In other words, if the sampling and estimation procedures were repeated a very large number of times and corresponding confidence intervals were constructed for each estimate using the same method, the proportion of these confidence intervals containing the true population value could be less than the stated confidence level.

To minimize the risk of undercoverage, more advanced methods than the Wald construction have been used to produce the confidence intervals for long-form estimates. These methods are known to achieve coverage closer to the nominal rate. However, all methods of constructing confidence intervals rely on assumptions that are generally not possible to explicitly verify in specific use cases. When working with confidence intervals, data users should be mindful of the scenarios that can lead to violation of the assumptions underlying the confidence interval construction. The different methods used to produce confidence intervals for long-form estimates, as well as the required assumptions, are described in detail in the following sections.

7.3 Student's confidence interval

The Student's confidence interval is used for all long-form statistics except proportions and counts. Since most estimates disseminated for the long-form are in fact for proportion and count statistics, this method accounts for a minority of disseminated confidence intervals.

The Student's confidence interval is based on the same pivot as the Wald interval. However, rather than assuming that the distribution of this pivot can be approximated by a normal distribution, the approximating distribution is assumed to be a Student's t-distribution. This distribution is known to be a more suitable approximating distribution for the pivot in cases where the sample size is small. The Student's t-distribution is specified by a single parameter known as the "degrees of freedom." The number of degrees of freedom of the Student's t-distribution is influenced by the sampling design, the number of sampled units and the variance estimation method. The number of degrees of freedom affects the width of the confidence interval. For the 2021 Census, the degrees of freedom were approximated by the number of replicates used for variance estimation (see Section 6.2) and denoted by R .

The lower bound and the upper bound of a 95% Student's confidence interval for a population parameter of interest θ are given by:

$$LB = \hat{\theta} - t \times \widehat{SE}(\hat{\theta}),$$
$$UB = \hat{\theta} + t \times \widehat{SE}(\hat{\theta}),$$

where

- $\hat{\theta}$ is the estimate of θ .
- t is the 97.5th percentile of the Student's t-distribution with R degrees of freedom.
- $\widehat{SE}(\hat{\theta})$ is the standard error of $\hat{\theta}$.

7.3.1 Properties of the Student's confidence interval

The Student's t-distribution is nearly identical to a standard normal distribution when the number of degrees of freedom is very large. When the number of degrees of freedom is small, the Student's t-distribution is wider than the standard normal distribution. This leads to the Student's confidence interval being wider than the Wald interval for the same estimate. Wald intervals often suffer from undercoverage when the sample size is small. Although the census long-form has a large sample size for the entire country, the sample size in small geographic areas or small domains of interest may be small. The Student's confidence interval will generally have better coverage than the Wald interval in these cases.

In practice, Student's confidence intervals may still suffer from undercoverage for very small sample sizes. This is due to failure of the assumptions when the sample size is very small. For instance, the distribution of the pivot may not be well approximated by a Student's t-distribution, or the approximation of the degrees of freedom by the number of replicates may substantially overestimate the actual degrees of freedom of the distribution. The breakdown of these assumptions when the sample size is very small will generally lead to undercoverage of the Student's confidence intervals.

7.4 Modified Wilson confidence interval for proportions

There are several different methods for constructing confidence intervals for proportions. For the 2021 Census, the modified Wilson confidence interval method was chosen because of its generally superior coverage and its practicality for implementation. This method is used for all proportion-type statistics. The method is based on the Wilson confidence interval for a simple random sampling with replacement (SRSWR) sample design (Wilson 1927). For the census, a modified version of this confidence interval is used that has been adapted to complex sample designs (Kott and Carr 1997). Extensive simulation studies have shown that this method performs better than the Wald and Student confidence intervals in situations where those confidence intervals exhibit undercoverage for proportion-type statistics (Neusy and Mantel 2016; Statistics Canada 2023).

For proportions, the assumption that the pivot used to construct the Student's confidence interval follows a Student's t-distribution breaks down for small sample sizes and when the statistic takes values near zero or one.

The modified Wilson confidence interval for a proportion p is based instead on the following pivot:

$$\frac{\hat{p}-p}{\sqrt{p(1-p)/n_e}}$$

where

- $\hat{p}$ is the estimate of p
- $n_e = \min\left(n/\mathrm{deff}(\hat{p}), n\right)$ is the effective sample size
- $\mathrm{deff}(\hat{p}) = \dfrac{\hat{V}(\hat{p})}{\hat{p}(1-\hat{p})/n}$ is the estimated design effect of $\hat{p}$ with respect to an SRSWR sample design
- n is the in-scope sample size
- $\hat{V}(\hat{p})$ is the estimated variance of $\hat{p}$.

The in-scope sample size is defined to be the number of sampled units which are in-scope for the question corresponding to the proportion p , i.e., the number of sampled units for which the question is applicable and which belong to the population of interest for the question. The expression under the square root sign in the denominator of the pivot is the variance of the proportion estimate under an SRSWR sample design, but with the sample size replaced by the effective sample size. By using the effective sample size in this expression, the variance is adjusted to account for the complex sample design of the census long-form. The modified Wilson confidence interval is based on the assumption that this pivot approximately follows a Student's t-distribution. As with the Student's confidence interval, for the 2021 Census, the degrees of freedom of the Student's t-distribution are approximated by R , the number of replicates used for variance estimation.

The lower bound and the upper bound of a 95% modified Wilson confidence interval for the proportion p are given by:

$$LB = \frac{\hat{p}+t^2/2n_e}{1+t^2/n_e} - \frac{t\sqrt{\hat{p}(1-\hat{p})+t^2/4n_e}}{\sqrt{n_e}(1+t^2/n_e)},$$

$$UB = \frac{\hat{p}+t^2/2n_e}{1+t^2/n_e} + \frac{t\sqrt{\hat{p}(1-\hat{p})+t^2/4n_e}}{\sqrt{n_e}(1+t^2/n_e)},$$

where t is the 97.5th percentile of the Student's t-distribution with R degrees of freedom and the other terms are as defined above.

7.4.1 Properties of the modified Wilson confidence interval for proportions

In addition to achieving better coverage than the Wald and Student intervals for small sample sizes and when the population parameter is near zero or one, the modified Wilson confidence interval for a proportion has the desirable property that its lower bound is never less than zero and that its upper bound is never greater than one. Since proportions cannot take on values outside of the interval between zero and one, it is reasonable that confidence intervals for proportions would exclude negative values and values greater than one.

It should also be noted that, unlike the Wald and Student intervals, the modified Wilson confidence interval for proportions is asymmetric, meaning that the estimate will not be exactly at the centre of the interval. The asymmetry is small when the effective sample size is large or when the estimated proportion is near 0.5.

Much like the Wald and Student confidence intervals, the modified Wilson confidence interval for proportions may suffer from some undercoverage, particularly when the sample size is very small, the value of the proportion is near zero or one, or there is high correlation between members of the same household. However, the modified Wilson method generally achieves nominal coverage rates in extreme situations, compared with the Wald and Student methods. It generally maintains coverage as good as or better than the Wald and Student methods in those situations.

7.5 Modified Wilson confidence interval for counts

For long-form estimates of counts, the confidence interval method used is a modified Wilson method similar to the method used for proportion-type statistics. For counts, the Wald and Student confidence intervals often perform poorly when the sample size is small, and when the value for the variable of interest is zero for almost all sampled units or when the value is one for almost all sampled units. In these situations, the distribution of the pivot used to construct the Wald and Student's confidence intervals is generally not well approximated by either a normal distribution or a Student's t-distribution.

The modified Wilson confidence interval for counts was developed for the 2021 Census as an alternative to the Wald and Student's confidence intervals which achieves coverage closer to the nominal rate. It has been tested in a simulation environment similar to the census long-form and has been shown to typically achieve good coverage (Neusy et al. 2021).

7.5.1 Modified Wilson confidence interval for counts: Theoretical form

The version of the modified Wilson confidence interval for counts used for the 2021 Census is an approximation of a theoretical form of the interval. The theoretical form can be derived in a similar manner to the modified Wilson confidence interval for proportions. In the case of the modified Wilson confidence interval for counts, the formulation relies on the notion of a calibration group of interest. A **calibration group** is a collection of units for which survey weights are calibrated with respect to control totals (for the census long-form, the calibration groups are ADAs and SADAs), and a **calibration group of interest** for a count Y is a calibration group that could potentially contain units with the characteristic of interest corresponding to Y .

The modified Wilson confidence interval for a count Y is based on the following pivot:

$$\frac{\hat{Y}-Y}{\sqrt{Y(N_C-Y)/n_e}}$$

where

- $\hat{Y}$ is the estimate of Y
- N_C is the total population size of the calibration groups of interest
- $n_e = \min(n_C / \text{deff}(\hat{Y}), n_C)$ is the effective total sample size in the calibration groups of interest
- n_C is the total sample size in the calibration groups of interest

- $\text{deff}(\hat{Y}) = \dfrac{\hat{V}(\hat{Y})}{\hat{Y}(N_C - \hat{Y})/n_C}$ is the estimated design effect of $\hat{Y}$ with respect to an SRSWR sample design, with population and sample size terms based on the calibration groups of interest
- $\hat{V}(\hat{Y})$ is the estimated variance of $\hat{Y}$.

The population size and sample size terms are defined with respect to the calibration groups of interest because this leads to a confidence interval with good properties. Specifically, simulations show that the resulting confidence interval has better coverage than for alternative ways of defining the size terms (Neusy et al. 2021).

Similar to the modified Wilson confidence interval for proportions, the modified Wilson confidence interval for counts is based on the assumption that the pivot approximately follows a Student's t-distribution. Like all confidence intervals for the 2021 Census, the degrees of freedom of the Student's t-distribution are approximated by R , the number of replicates used for variance estimation.

From the above pivot, the theoretical version of the 95% modified Wilson confidence interval for the count Y can be derived. This version of the interval has the following lower bound and upper bound:

$$LB = \frac{\hat{Y} + N_C t^2 / 2n_e}{1 + t^2 / n_e} - \frac{t\sqrt{\hat{Y}(N_C - \hat{Y}) + N_C{}^2 t^2 / 4n_e}}{\sqrt{n_e}(1 + t^2 / n_e)},$$

$$UB = \frac{\hat{Y} + N_C t^2 / 2n_e}{1 + t^2 / n_e} + \frac{t\sqrt{\hat{Y}(N_C - \hat{Y}) + N_C{}^2 t^2 / 4n_e}}{\sqrt{n_e}(1 + t^2 / n_e)},$$

where the terms are as defined above. In this version of the interval, the similarity to the modified Wilson confidence interval for proportions is evident.

7.5.2 Modified Wilson confidence interval for counts: Approximate form

The approximate form of the modified Wilson confidence interval for counts that was implemented for the 2021 Census is based on the theoretical interval above. An approximation was used because of limitations of the census tabulation system.

In the approximate form, the lower bound and the upper bound of a 95% modified Wilson confidence interval for a count Y are given by:

$$LB = \hat{Y} + t^2 \frac{1}{2} \frac{\hat{V}(\hat{Y})}{\hat{Y}} - \sqrt{t^2 \hat{V}(\hat{Y}) + \left(t^2 \frac{1}{2} \frac{\hat{V}(\hat{Y})}{\hat{Y}}\right)^2},$$

$$UB = \hat{Y} + t^2 \frac{1}{2} \frac{\hat{V}(\hat{Y})}{\hat{Y}} + \sqrt{t^2 \hat{V}(\hat{Y}) + \left(t^2 \frac{1}{2} \frac{\hat{V}(\hat{Y})}{\hat{Y}}\right)^2},$$

where

- $\widehat{Y}$ is the estimate of Y
- t is the 97.5th percentile of the Student's t-distribution with R degrees of freedom
- $\widehat{V}(\widehat{Y})$ is the estimated variance of $\widehat{Y}$.

7.5.3 Assumptions for approximating the modified Wilson confidence interval for counts

It has been demonstrated empirically that the approximation implemented for the census behaves well for sample designs similar to that of the long-form (Neusy et al. 2021). The approximation is based on the following assumptions:

1. The total sample size in the calibration groups of interest n_C is sufficiently large so that $1/n_e$ is close to zero.
2. The estimated count $\widehat{Y}$ is much smaller than the total population size of the calibration groups of interest N_C .

The first assumption is generally valid for the census long-form questionnaire because the calibration groups of interest correspond to SADAs or ADAs, which always have a very large sample size. The second assumption may not be met, but only for common characteristics and for very large domains of interest, such as SADAs. In this context, the approximate version of the modified Wilson confidence interval is nearly identical to the Student confidence interval, and both methods perform well enough for large domain sizes. Therefore, not meeting the second assumption is not a concern for the census long-form.

7.5.4 Properties of the modified Wilson confidence interval for counts

The modified Wilson confidence interval for counts has the advantage over the Wald and Student confidence intervals that the lower bound of the confidence interval is never less than zero. This property, as well as the other properties described in this section, applies to both the theoretical and the approximate version of the confidence interval. Since count statistics cannot be negative, it is appropriate that the confidence interval does not contain negative values. Similarly to the modified Wilson confidence interval for proportions, the modified Wilson confidence interval for counts is asymmetric. This asymmetry will be small when the estimated variance of a count is small in relation to the estimated count itself.

Much like the Wald and Student confidence intervals, the modified Wilson confidence interval for counts may suffer from some undercoverage, particularly when the sample size is very small, the value of the count is close to zero or close to the size of the population of the domain of interest, or there is high correlation between members of the same household. However, the modified Wilson method generally achieves nominal coverage rates in extreme situations, compared with the Wald and Student methods. It generally maintains coverage as good as or better than the Wald and Student methods in those situations.

8. Conclusion

The 2021 Census long-form questionnaire saw the introduction of considerable new content, and its collection took place during a pandemic. Both of these factors posed methodological and operational challenges. In particular, significant efforts were made to achieve response rates similar to those in 2016. Additionally, administrative data were introduced in the imputation process to try to minimize non-response bias in the rare areas where high response rates could not be achieved. Despite those challenges, sampling and weighting methods were by and large a continuation of the 2016 Census long-form methods.

The introduction of a new dissemination system allowed Statistics Canada to revamp and improve the dissemination of data quality indicators. In an effort to empower data users to make valid statistical inference, new confidence interval methods were developed and implemented. Confidence intervals became the key variance-based data quality indicator and are disseminated with the majority of long-form data tables. Additionally, detailed non-response and imputation rate statistics per question were made available to data users to make the data quality more transparent. For more information, see the *2021 Census Data Quality Guidelines*, Statistics Canada Catalogue no. 98-26-0006.

Appendix A – Glossary

The definitions of the main census terms, variables and concepts mentioned in this document are presented here. Users can also refer to the *Dictionary, Census of Population, 2021*, Catalogue no. 98-301-X for additional information.

Aggregate dissemination area: An aggregate dissemination area (ADA) is a dissemination geography created for the census. ADAs cover the entire country and, where possible, have a population between 5,000 and 15,000 based on the previous census population counts. ADAs are created by grouping existing dissemination geographic areas, including census tracts (CTs), census subdivisions (CSDs) or dissemination areas (DAs). ADA boundaries respect provincial, territorial, census division (CD), census metropolitan area (CMA) and census agglomeration (CA) boundaries.

The intent of the ADA geography is to ensure the availability of census data, where possible, across all regions of Canada.

Census division: Group of neighbouring municipalities joined together for the purposes of regional planning and managing common services (such as police or ambulance services). These groupings are established under laws in effect in certain provinces of Canada. Census division (CD) is the general term for provincially legislated areas (such as county, *municipalité régionale de comté* (MRC) and regional district) or their equivalents. In other provinces and the territories where laws do not provide for such areas, Statistics Canada defines equivalent areas for statistical reporting purposes in cooperation with these provinces and territories. Census divisions are intermediate geographic areas between the province or territory level and the municipality (census subdivision).

Census family: Census family is defined as a married couple and the children, if any, of either and/or both spouses; a couple living common law and the children, if any, of either and/or both partners; or a parent of any marital status in a one-parent family with at least one child living in the same dwelling and that child or those children. All members of a particular census family live in the same dwelling. Children may be biological or adopted children regardless of their age or marital status as long as they live in the dwelling and do not have their own married spouse, common-law partner or child living in the dwelling. Grandchildren living with their grandparent(s) but with no parents present also constitute a census family.

Census subdivision: Census subdivision (CSD) is the general term for municipalities (as determined by provincial/territorial legislation) or areas treated as municipal equivalents for statistical purposes (e.g., Indian reserves, Indian settlements and unorganized territories). Municipal status is defined by laws in effect in each province and territory in Canada.

Census tract: Census tracts (CTs) are small, relatively stable geographic areas that usually have a population of fewer than 7,500 persons, based on data from the previous Census of Population Program. They are located in census metropolitan areas (CMAs) and in census agglomerations (CAs) that had a core population of 50,000 or more in the previous census.

A committee of local specialists (for example, municipal planners and others) initially delineates CTs in conjunction with Statistics Canada. Once a CMA or CA has been subdivided into CTs, the CTs are maintained even if the core population subsequently declines below 50,000.

Collection unit:[10] Collection units (CUs) are small geographic units used for the collection of census data. CUs cover all the territory of Canada.

[10] This definition is not found in the *Dictionary, Census of Population, 2021*, because the dictionary consists mainly of dissemination terms and this is a collection term.

Collective dwelling: Collective dwelling refers to a dwelling of a commercial, institutional, or communal nature in which a person or group of persons reside or could reside. It must provide care or services or have certain common facilities, such as a kitchen or bathroom, which are shared by the occupants. Examples include lodging or rooming houses, hotels, motels, tourist establishments, nursing homes, residences for senior citizens, hospitals, staff residences, military bases, work camps, correctional facilities and group homes.

Dissemination area: A dissemination area (DA) is a small, relatively stable geographic unit composed of one or more adjacent dissemination blocks with an average population of 400 to 700 persons based on data from the previous Census of Population Program. It is the smallest standard geographic area for which all census data are disseminated. DAs cover all the territory of Canada.

Dwelling: A dwelling is defined as a set of living quarters. Two types of dwellings are identified in the census, collective dwellings and private dwellings. The former pertains to dwellings which are institutional, communal or commercial in nature. The latter, private dwellings, refers to a separate set of living quarters with a private entrance either from outside the building or from a common hall, lobby, vestibule or stairway inside the building. The entrance to the dwelling must be one that can be used without passing through the living quarters of some other person or group of persons.

Economic family: “Economic family” refers to a group of two or more persons who live in the same dwelling and are related to each other by blood, marriage, common-law union, adoption or a foster relationship.

By definition, all persons who are members of a census family are also members of an economic family. Examples of the broader concept of economic family include the following: two co-resident census families who are related to one another are considered one economic family; co-resident siblings who are not members of a census family are considered as one economic family; and nieces or nephews living with aunts or uncles are considered one economic family.

Household: Household refers to a person or group of persons who occupy the same dwelling and do not have a usual place of residence elsewhere in Canada or abroad. The dwelling may be either a collective dwelling or a private dwelling. The household may consist of a family group such as a census family, of two or more families sharing a dwelling, of a group of unrelated persons or of a person living alone. Household members who are temporarily absent on reference day are considered part of their usual household.

Private dwelling: “Private dwelling” refers to a separate set of living quarters with a private entrance either from outside the building or from a common hall, lobby, vestibule or stairway inside the building. The entrance to the dwelling must be one that can be used without passing through the living quarters of some other person or group of persons.

The dwelling must meet the two conditions necessary for year-round occupancy:

1. a source of heat or power (as evidenced by chimneys, power lines, oil or gas pipes or meters, generators, woodpiles, electric lights, heating pumps, or solar panels)
2. an enclosed space that provides shelter from the elements, as evidenced by complete and enclosed walls and a roof, and by doors and windows that provide protection from wind, rain and snow.

Dwellings that do not meet the conditions necessary for year-round occupancy are marginal dwellings. Private dwellings are classified into regular private dwellings and occupied marginal dwellings. Regular private dwellings are further classified into three major groups: occupied dwellings (occupied by usual residents), dwellings occupied solely by foreign residents or by temporarily present persons, and unoccupied dwellings. Marginal dwellings are classified as occupied by usual residents or occupied solely by foreign residents or by temporarily present persons. Marginal dwellings that were unoccupied on May 11, 2021, are not counted in the housing stock.

Private dwelling occupied by usual residents: A private dwelling occupied by usual residents refers to a private dwelling in which a person or a group of persons is permanently residing. Also included are private dwellings whose usual residents are temporarily absent on May 11, 2021. Unless otherwise specified, all data in housing products are for private dwellings occupied by usual residents, rather than for unoccupied private dwellings or dwellings occupied solely by foreign residents or by temporarily present persons.

Private household: Private household refers to a person or group of persons who occupy the same dwelling and do not have a usual place of residence elsewhere in Canada or abroad. The household universe is divided into two sub-universes on the basis of whether the household is occupying a collective dwelling or a private dwelling. The latter is a private household.

For census purposes, households are classified into three groups: private households, collective households and households outside Canada.

Unless otherwise specified, all data in census products are for private households only.

Super aggregate dissemination area:[11] Super aggregate dissemination areas (SADAs) are a geography created specifically for weighting census data. They respect pre-established rules, some of which are mandatory and others optional. SADAs are created by combining aggregate dissemination areas (ADAs) and are contained within provincial and territorial boundaries. All individuals living in census collection units (CUs) in First Nations communities, Métis settlements, Inuit regions and other remote areas are excluded from the SADA population. SADAs are created, in as much as possible, with a target population of between 50,000 and 150,000. Census divisions (CD) with a population of 40,000 to 50,000 comprise their own SADA. In addition, where possible, SADAs respect the boundaries of—in order of priority—census division (CDs), census metropolitan areas (CMAs), census agglomerations (CAs) and census subdivisions (CSD). Lastly, SADAs should be created by combining adjacent ADAs (where possible) and must be as compact as possible.

[11] This definition is not found in the *Dictionary, Census of Population, 2021*, because the dictionary consists mainly of dissemination terms and this is a weighting term.

Appendix B – The history of sampling in the Canadian census

Sampling was first used in the Canadian census in 1941. A housing schedule was completed for every 10th dwelling. The information from 27 questions on the separate housing schedule was integrated with the data in the personal and household section of the population schedule for the same dwelling. This enabled cross-tabulation of sample and basic characteristics. Also, in the 1941 Census, sampling was used at the processing stage to obtain early estimates of earnings of wage-earners, of the distribution of the population of working age and of the composition of families in Canada. In this case, a sample of every 10th enumeration area across Canada was selected and all population schedules in these areas were processed in advance.

The census of housing was again conducted on a sample basis in 1951. This time, every fifth dwelling (those whose identification numbers ended in a 2 or 7) was selected to complete a housing document containing 24 questions. In the 1961 Census, persons aged 15 years and older in a 20% sample of private households were required to complete a population sample questionnaire containing questions on internal migration, fertility and income. Sampling was not used in the smaller censuses of 1956 and 1966.

The 1971 Census saw several major innovations in the method of census-taking. The primary change was from the traditional canvasser method of enumeration to the use of self-enumeration for the majority of the population. This change was prompted by the results of several studies in Canada and elsewhere (Fellegi 1964; Hansen et al. 1959), which indicated that the effect of the enumerator was a major contribution to the variance of census figures in a canvasser census. Consequently, the use of self-enumeration was expected to reduce the variance of census figures by reducing the effect of the enumerator and by giving the respondent more time and privacy in which to answer the census questions—factors that might be expected to yield more accurate responses.

The second aspect of the 1971 Census that differentiated it from any earlier census was its content. The number of topics covered and the number of questions asked were greater than in any previous census. Considerations of cost, respondent burden and timeliness versus the level of data quality to be expected using self-enumeration and sampling led to a decision to collect all but certain basic characteristics on a one-third sample basis in the 1971 Census. In all but the more remote areas of Canada, every third private household received the “long questionnaire,” which contained all the census questions. The remaining private households received the “short questionnaire” containing only the basic questions covering name, relationship to head of household, sex, date of birth, marital status, mother tongue, type of dwelling, tenure, number of rooms, water supply, toilet facilities and certain census coverage items. All households in pre-identified remote enumeration areas and all collective dwellings received the long questionnaire. A more detailed description of the consideration of the use of sampling in the 1971 Census is given in *Sampling in the Census* (Dominion Bureau of Statistics 1968).

The 1976 Census had considerably less content than the 1971 Census. Furthermore, the 1976 questionnaire did not include the questions that cause the most difficulty in collection (e.g., income) or that are costly to code (e.g., occupation, industry and place of work). Therefore, the benefits of sampling in terms of cost savings and reduced respondent burden were less clear than for the 1971 Census. Nevertheless, after estimating the potential cost savings to be expected with various sampling fractions and considering the public relations issues related to a reversion to 100% enumeration after a successful application of sampling in 1971, Statistics Canada decided to use the same sampling procedure in 1976 as in 1971.

Most of the methodology used in the 1971 and 1976 censuses was kept for the 1981 Census, except that the sampling rate was reduced from every third occupied private household to every fifth. Studies done at the time showed that the resulting reduction in data quality (measured in terms of variance) would be tolerable and would not be significant enough to offset the benefits of reduced cost and respondent burden and improved timeliness (Royce 1983). The one-in-five sampling rate was maintained for every census from 1981 to 2006.

In 2011, information previously collected by the mandatory long-form census questionnaire was collected on a voluntary basis, via the National Household Survey (NHS). With this change, every household was required to answer the 10 questions that were contained in the 2011 Census questionnaire, while 30% of households were selected to respond to the NHS. As well, NHS non-responding households were subsampled for follow-up at a rate of one in three. The increased sampling fraction to 30% was implemented in anticipation of a lower response rate to the NHS. For the 2016 Census, the government reinstated the census long-form questionnaire as mandatory, replacing the NHS. The sampling fraction was changed in 2016 to one in four, compared with one in five for the previous census long-form questionnaire in 2006, to mitigate the risk of the response rate not recovering to its previously high levels.

In 2021, the sample design and sampling fraction of one in four remained nearly identical to those of the 2016 Census. Only a small operational improvement to support a random start to the systematic sampling procedure was made. This was done in an effort to ensure the sample of the two cycles were statistically independent, since dependent samples would have made historical comparisons more difficult to make.

Appendix C – Constraints used in or excluded from the weighting process

The following is a list of the possible constraints defined at the aggregate dissemination area (ADA) and super aggregate dissemination area (SADA) levels. A total of 271 possible constraints were defined at the ADA level and 203 at the SADA level. The table includes the number of times the constraint was calibrated on and the number of times it was excluded. In total, 408 SADAs and 4,207 ADAs were subjected to the weighting processes. The constraints HHADACSD and PPADACSD, while labelled ADA constraints, could possibly be calibrated for multiple census subdivisions within an ADA.

Table C.1
Statistics on the use of calibration constraints, by constraint

		Coverage and non-response adjustment			Final calibration		
Constraint variable name	Description	Area	Number of calibrated constraints	Number of excluded constraints	Area	Number of calibrated constraints	Number of excluded constraints
ADULTCF	Adults in a census family	SADA	155	253	SADA	8	400
AGE00_14	Persons aged 0 to 14 years	SADA	350	58	Both	4,360	255
AGE14	Persons aged 10 to 14 years	SADA	264	144	Both	1,022	3,593
AGE15_24	Persons aged 15 to 24 years	SADA	317	91	SADA	312	96
AGE15_29	Persons aged 15 to 29 years	…	…	…	ADA	4,059	148
AGE19	Persons aged 15 to 19 years	SADA	82	326	Both	710	3,905
AGE24	Persons aged 20 to 24 years	SADA	79	329	Both	726	3,889
AGE25_34	Persons aged 25 to 34 years	SADA	384	24	SADA	385	23
AGE29	Persons aged 25 to 29 years	SADA	365	43	Both	1,426	3,189
AGE30_49	Persons aged 30 to 49 years	…	…	…	ADA	4,143	64
AGE34	Persons aged 30 to 34 years	SADA	380	28	Both	1,736	2,879
AGE35_44	Persons aged 35 to 44 years	SADA	405	3	SADA	402	6
AGE39	Persons aged 35 to 39 years	SADA	389	19	Both	1,706	2,909
AGE4	Persons aged 0 to 4 years	SADA	277	131	Both	796	3,819
AGE44	Persons aged 40 to 44 years	SADA	388	20	Both	1,547	3,068
AGE45_54	Persons aged 45 to 54 years	SADA	383	25	SADA	381	27
AGE49	Persons aged 45 to 49 years	SADA	373	35	Both	1,520	3,095
AGE50_64	Persons aged 50 to 64 years	…	…	…	ADA	4,022	185
AGE54	Persons aged 50 to 54 years	SADA	378	30	Both	1,665	2,950
AGE55_64	Persons aged 55 to 64 years	SADA	393	15	SADA	365	43
AGE59	Persons aged 55 to 59 years	SADA	382	26	Both	1,685	2,930
AGE64	Persons aged 60 to 64 years	SADA	379	29	Both	1,743	2,872

Table C.1
Statistics on the use of calibration constraints, by constraint

		Coverage and non-response adjustment			Final calibration		
Constraint variable name	Description	Area	Number of calibrated constraints	Number of excluded constraints	Area	Number of calibrated constraints	Number of excluded constraints
AGE65PL	Persons aged 65 years and older	SADA	406	2	Both	4,295	320
AGE74	Persons aged 65 to 74 years	SADA	406	2	Both	2,428	2,187
AGE75PL	Persons aged 75 years and older	SADA	404	4	Both	2,320	2,295
AGE9	Persons aged 5 to 9 years	SADA	288	120	Both	776	3,839
APT5PLUS	Households living in an apartment in a building that has 5 or more storeys	SADA	261	147	Both	1,210	3,405
APTLT5	Households living in an apartment in a building with less than five storeys	SADA	401	7	Both	2,585	2,030
CHILD	Children in a census family	SADA	93	315	SADA	13	395
CHILDFAM	Census families with children	SADA	238	170	Both	3,056	1,559
COMLAWNO_DIV	Divorced persons not in a common-law couple	SADA	369	39	Both	1,190	3,425
COMLAWNO_OTHERS	Divorced, separated or widowed persons not in a common-law couple	SADA	369	39	Both	3,097	1,518
COMLAWNO_SEP	Separated persons not in a common-law couple	SADA	292	116	Both	575	4,040
COMLAWNO_SINGLE	Never-married persons not in a common-law couple	SADA	261	147	Both	3,648	967
COMLAWNO_SINGLE_GE15	Never-married persons aged 15 years and older not in a common-law couple	SADA	259	149	Both	3,656	959
COMLAWNO_SINGLE_LT15	Never-married persons aged less than 15 years not in a common-law couple	SADA	350	58	Both	4,360	255
COMLAWNO_WID	Widowed persons not in a common-law couple	SADA	290	118	Both	679	3,936
COMLAWYE_MARRIED	Persons married or in a common-law couple	SADA	245	163	Both	3,370	1,245
COMLAW_YE	Persons in a common-law couple	SADA	260	148	Both	2,739	1,876
COUPLE	Persons in a couple (married or common-law)	SADA	156	252	SADA	33	375
EMPIN_GT50	Persons with an annual employment income above the 50th percentile for the ADA	…	…	…	ADA	4,127	80

Table C.1
Statistics on the use of calibration constraints, by constraint

		Coverage and non-response adjustment			Final calibration		
Constraint variable name	Description	Area	Number of calibrated constraints	Number of excluded constraints	Area	Number of calibrated constraints	Number of excluded constraints
EMPIN_LE50	Persons with an annual employment income equal to or below the 50th percentile for the ADA	…	…	…	ADA	4,135	72
EMPIN_P0	Persons with no annual employment income, at the ADA level	…	…	…	ADA	4,143	64
EMPIN_P0_GE15	Persons aged 15 years and older with no annual employment income, at the ADA level	…	…	…	ADA	4,057	150
EMPIN_P0_LT15	Persons aged less than 15 years with no annual employment income, at the ADA level	…	…	…	ADA	4,033	174
EMPIN_P100	Persons with an annual employment income above the 75th percentile for the ADA	…	…	…	ADA	3,707	500
EMPIN_P25	Persons with an annual employment income equal to or below the 25th percentile for the ADA	…	…	…	ADA	2,922	1,285
EMPIN_P50	Persons with an annual employment income above the 25th percentile and equal to or below the 50th percentile for the ADA	…	…	…	ADA	2,918	1,289
EMPIN_P75	Persons with an annual employment income above the 50th percentile and equal to or below the 75th percentile for the ADA	…	…	…	ADA	3,708	499
EMPIN_SADA_GT50	Persons with an annual employment income above the 50th percentile for the SADA	SADA	408	0	SADA	324	84
EMPIN_SADA_LE50	Persons with an annual employment income equal to or below the 50th percentile for the SADA	SADA	407	1	SADA	292	116
EMPIN_SADA_P0	Persons with no annual employment income, at the SADA level	SADA	407	1	SADA	360	48
EMPIN_SADA_P0_GE15	Persons aged 15 years and older with no annual employment income, at the SADA level	SADA	351	57	SADA	328	80

Table C.1
Statistics on the use of calibration constraints, by constraint

		Coverage and non-response adjustment			Final calibration		
Constraint variable name	**Description**	**Area**	**Number of calibrated constraints**	**Number of excluded constraints**	**Area**	**Number of calibrated constraints**	**Number of excluded constraints**
EMPIN_SADA_P0_LT15	Persons aged less than 15 years with no annual employment income, at the SADA level	SADA	350	58	SADA	327	81
EMPIN_SADA_P100	Persons with an annual employment income above the 75th percentile for the SADA	SADA	407	1	SADA	295	113
EMPIN_SADA_P25	Persons with an annual employment income equal to or below the 25th percentile for the SADA	SADA	272	136	SADA	126	282
EMPIN_SADA_P50	Persons with an annual employment income above the 25th percentile and equal to or below the 50th percentile for the SADA	SADA	271	137	SADA	145	263
EMPIN_SADA_P75	Persons with an annual employment income above the 50th percentile and equal to or below the 75th percentile for the SADA	SADA	407	1	SADA	288	120
FEMALE	Women+	SADA	403	5	Both	4,365	250
FEMALEGE15	Females aged 15 years and older	SADA	371	37	Both	3,744	871
FEMALELT15	Females aged less than 15 years	SADA	368	40	Both	3,636	979
HHADA	Households in the ADA	ADA	3,592	615	…	…	…
HHADACSD	Households that fall within the CSD and the ADA	…	…	…	ADA	4,866	2,066
HHINC_GT50	Households with an annual income above the 50th percentile for the ADA	…	…	…	ADA	4,178	29
HHINC_LE50	Households with an annual income at or below the 50th percentile for the ADA	…	…	…	ADA	4,178	29
HHINC_P100	Households with an annual income above the 75th percentile for the ADA	…	…	…	ADA	4,033	174
HHINC_P25	Households with an annual income at or below the 25th percentile for the ADA	…	…	…	ADA	4,052	155
HHINC_P50	Households with an annual income above the 25th percentile and at or below the 50th percentile for the ADA	…	…	…	ADA	4,052	155

Table C.1
Statistics on the use of calibration constraints, by constraint

		Coverage and non-response adjustment			Final calibration		
Constraint variable name	Description	Area	Number of calibrated constraints	Number of excluded constraints	Area	Number of calibrated constraints	Number of excluded constraints
HHINC_P75	Households with an annual income above the 50th percentile and at or below the 75th percentile for the ADA	…	…	…	ADA	4,033	174
HHINC_SADA_GT50	Households with an annual income above the 50th percentile for the SADA	SADA	408	0	SADA	364	44
HHINC_SADA_LE50	Households with an annual income at or below the 50th percentile for the SADA	SADA	408	0	SADA	364	44
HHINC_SADA_P100	Households with an annual income above the 75th percentile for the SADA	SADA	406	2	SADA	320	88
HHINC_SADA_P25	Households with an annual income at or below the 25th percentile for the SADA	SADA	408	0	SADA	372	36
HHINC_SADA_P50	Households with an annual income above the 25th percentile and at or below the 50th percentile for the SADA	SADA	408	0	SADA	352	56
HHINC_SADA_P75	Households with an annual income above the 50th percentile and at or below the 75th percentile for the SADA	SADA	406	2	SADA	333	75
HHSIZE1	One-person households	SADA	259	149	Both	888	3,727
HHSIZE2	Two-person households	SADA	237	171	Both	3,904	711
HHSIZE3	Three-person households	SADA	288	120	Both	3,391	1,224
HHSIZE4	Four-person households	SADA	389	19	Both	3,527	1,088
HHSIZE5	Five-person households	SADA	193	215	Both	848	3,767
HHSIZEGE5	Five-or-more-person households	SADA	7	401	Both	135	4,480
HHSIZEGE6	Six-or-more-person households	SADA	15	393	Both	42	4,573
INEFAM	Persons in an economic family	SADA	282	126	SADA	177	231
IR_LINK_NO	Persons who could not be linked to the Indian Register	SADA	152	256	Both	403	4,212
IR_LINK_YE	Persons who could be linked to the Indian Register	SADA	152	256	Both	334	4,281
LIM_NO	Persons not in a low income household (after tax)	SADA	403	5	Both	2,820	1,795

Table C.1
Statistics on the use of calibration constraints, by constraint

		Coverage and non-response adjustment			Final calibration		
Constraint variable name	**Description**	**Area**	**Number of calibrated constraints**	**Number of excluded constraints**	**Area**	**Number of calibrated constraints**	**Number of excluded constraints**
LIM_YE	Persons in a low income household (after tax)	SADA	403	5	Both	2,820	1,795
LONEPAR	Parents in one-parent families	SADA	159	249	SADA	14	394
MALE	Men+	SADA	403	5	Both	4,365	250
MALEGE15	Males aged 15 years and older	SADA	362	46	Both	3,708	907
MALELT15	Males aged less than 15 years	SADA	360	48	Both	3,606	1,009
MARRIED	Married persons	SADA	373	35	Both	2,831	1,784
NB_NOTINCF	Persons not in a census family	SADA	220	188	Both	3,732	883
NOCLDFAM	Census families without children	SADA	223	185	Both	1,630	2,985
NOINEFAM	Persons not in an economic family	SADA	282	126	SADA	177	231
NOINEFAMHHSIZEEQ1	Persons not in an economic family - In a one-person household	SADA	259	149	SADA	50	358
NOINEFAMHHSIZEGT1	Persons not in an economic family - In a two-or-more-person household	SADA	205	203	SADA	59	349
NOTINFAM	Persons not in a census family	SADA	220	188	SADA	174	234
NOTINFAMHHSIZEEQ1	Persons not in a census family - In a one-person household	SADA	259	149	SADA	50	358
NOTINFAMHHSIZEGT1	Persons not in a census family - In a two-or-more-person household	SADA	163	245	SADA	32	376
OLN_BI	Official languages—English and French	SADA	301	107	Both	1,859	2,756
OLN_EN	Official language—English	SADA	69	339	Both	495	4,120
OLN_FR	Official language—French	SADA	80	328	Both	552	4,063
OLN_NO	Official language—neither	SADA	157	251	Both	248	4,367
POBG2_1	Place of birth—Albania	SADA	0	408	Both	5	4,610
POBG2_10	Place of birth—Brazil	SADA	16	392	Both	58	4,557
POBG2_11	Place of birth—Bulgaria and Romania	SADA	26	382	Both	52	4,563
POBG2_16	Place of birth—Democratic Republic of the Congo and Republic of the Congo	SADA	1	407	Both	9	4,606
POBG2_17	Place of birth—Cameroon, Central African Republic, Chad and Gabon	SADA	2	406	Both	12	4,603

Table C.1
Statistics on the use of calibration constraints, by constraint

		Coverage and non-response adjustment			Final calibration		
Constraint variable name	**Description**	**Area**	**Number of calibrated constraints**	**Number of excluded constraints**	**Area**	**Number of calibrated constraints**	**Number of excluded constraints**
POBG2_18	Place of birth—Angola, and Sao Tome and Principe	SADA	0	408	Both	1	4,614
POBG2_19	Place of birth—Kazakhstan, Kyrgyzstan, Tajikistan, Turkmenistan and Uzbekistan	SADA	0	408	Both	2	4,613
POBG2_20	Place of birth—Chile	SADA	1	407	Both	9	4,606
POBG2_21	Place of birth—China, Hong Kong, Macao and Taiwan	SADA	83	325	Both	162	4,453
POBG2_22	Place of birth—Colombia, Ecuador and Peru	SADA	47	361	Both	107	4,508
POBG2_24	Place of birth—Czech Republic, Hungary and Slovakia	SADA	9	399	Both	16	4,599
POBG2_25	Place of birth—Burundi and Rwanda	SADA	2	406	Both	4	4,611
POBG2_26	Place of birth—Eritrea, Kenya, Tanzania, Uganda and Zambia	SADA	19	389	Both	51	4,564
POBG2_27	Place of birth—Comoros, Djibouti, Madagascar, Malawi, Mauritius, Seychelles, Somalia and Zimbabwe	SADA	6	402	Both	16	4,599
POBG2_28	Place of birth—Belarus, Moldova, Russian Federation and Ukraine	SADA	37	371	Both	67	4,548
POBG2_29	Place of birth—Egypt, South Sudan and Sudan	SADA	13	395	Both	10	4,605
POBG2_3	Place of birth—Australia and New Zealand	SADA	0	408	Both	3	4,612
POBG2_30	Place of birth—Ethiopia	SADA	10	398	Both	12	4,603
POBG2_31	Place of birth—France, Luxembourg and Monaco	SADA	18	390	Both	32	4,583
POBG2_32	Place of birth—Cambodia, Laos and Viet Nam	SADA	16	392	Both	24	4,591
POBG2_33	Place of birth—Cuba, Dominican Republic and Haiti	SADA	8	400	Both	13	4,602
POBG2_34	Place of birth—Greece	SADA	2	406	Both	10	4,605
POBG2_35	Place of birth—Guyana and Suriname	SADA	18	390	Both	39	4,576

Table C.1
Statistics on the use of calibration constraints, by constraint

		Coverage and non-response adjustment			Final calibration		
Constraint variable name	Description	Area	Number of calibrated constraints	Number of excluded constraints	Area	Number of calibrated constraints	Number of excluded constraints
POBG2_36	Place of birth—Holy See (Vatican City State), Italy and San Marino	SADA	29	379	Both	47	4,568
POBG2_37	Place of birth—Bahamas, Jamaica and Puerto Rico	SADA	43	365	Both	71	4,544
POBG2_38	Place of birth—Antigua and Barbuda, Aruba, Barbados, Bonaire, Sint Eustatius and Saba, Curaçao, Dominica, Grenada, Saint Kitts and Nevis, Saint Lucia, Sint Maarten (Dutch part), Saint Vincent and the Grenadines, Trinidad and Tobago, and United States Virgin Islands	SADA	8	400	Both	15	4,600
POBG2_39	Place of birth—Japan	SADA	12	396	Both	18	4,597
POBG2_4	Place of birth—Austria, Germany and Liechtenstein	SADA	19	389	Both	47	4,568
POBG2_40	Place of birth—North Korea and South Korea	SADA	10	398	Both	19	4,596
POBG2_41	Place of birth—Liberia	SADA	0	408	Both	2	4,613
POBG2_42	Place of birth—Algeria, Libya, Morocco and Tunisia	SADA	14	394	Both	10	4,605
POBG2_43	Place of birth—Brunei Darussalam, Indonesia, Malaysia, Philippines, Singapore and Timor-Leste	SADA	96	312	Both	118	4,497
POBG2_45	Place of birth—Mexico	SADA	12	396	Both	26	4,589
POBG2_46	Place of birth—Bahrain, Qatar, Saudi Arabia, United Arab Emirates and Yemen	SADA	16	392	Both	27	4,588
POBG2_47	Place of birth—Lebanon, Syria	SADA	26	382	Both	49	4,566
POBG2_48	Place of birth—Afghanistan, Cyprus, Iran, Iraq, Israel, Jordan, Kuwait, Oman, Turkey, and West Bank and Gaza Strip (Palestine)	SADA	87	321	Both	107	4,508
POBG2_50	Place of birth—Mozambique	SADA	0	408	Both	1	4,614
POBG2_51	Place of birth—Nepal	SADA	1	407	Both	2	4,613
POBG2_54	Place of birth—Poland	SADA	37	371	Both	52	4,563

Table C.1
Statistics on the use of calibration constraints, by constraint

		Coverage and non-response adjustment			Final calibration		
Constraint variable name	**Description**	**Area**	**Number of calibrated constraints**	**Number of excluded constraints**	**Area**	**Number of calibrated constraints**	**Number of excluded constraints**
POBG2_55	Place of birth—Oceania region (excluding Australia and New Zealand)	SADA	0	408	Both	2	4,613
POBG2_56	Place of birth—Portugal	SADA	30	378	Both	40	4,575
POBG2_57	Place of birth—Argentina, Bolivia, Paraguay and Uruguay	SADA	2	406	Both	2	4,613
POBG2_59	Place of birth—Namibia and Republic of South Africa	SADA	0	408	Both	2	4,613
POBG2_6	Place of birth—Belgium and Netherlands	SADA	4	404	Both	9	4,606
POBG2_60	Place of birth—Sri Lanka	SADA	7	401	Both	16	4,599
POBG2_63	Place of birth—Armenia, Azerbaijan and Georgia	SADA	0	408	Both	1	4,614
POBG2_64	Place of birth—Bangladesh, India and Pakistan	SADA	39	369	Both	69	4,546
POBG2_65	Place of birth—Union of Soviet Socialist Republics, Former	SADA	5	403	Both	17	4,598
POBG2_66	Place of birth—Guernsey, Ireland, Isle of Man, Jersey, Sark and United Kingdom	SADA	5	403	Both	0	4,615
POBG2_67	Place of birth—United States	SADA	171	237	Both	245	4,370
POBG2_68	Place of birth—Venezuela	SADA	4	404	Both	6	4,609
POBG2_69	Place of birth—Gambia, Ghana, Nigeria and Sierra Leone	SADA	4	404	Both	3	4,612
POBG2_7	Place of birth—Belize, El Salvador, Guatemala and Honduras	SADA	8	400	Both	15	4,600
POBG2_70	Place of birth—Benin, Burkina Faso, Côte d’Ivoire, Guinea, Mali, Mauritania, Niger, Senegal and Togo	SADA	2	406	Both	2	4,613
POBG2_71	Place of birth—Bosnia and Herzegovina, Croatia, Kosovo, Republic of Macedonia, Montenegro, Serbia, and Slovenia	SADA	40	368	Both	65	4,550
POBG2_9	Place of birth—Botswana, Lesotho and Swaziland	SADA	1	407	Both	0	4,615

Table C.1
Statistics on the use of calibration constraints, by constraint

		Coverage and non-response adjustment			Final calibration		
Constraint variable name	Description	Area	Number of calibrated constraints	Number of excluded constraints	Area	Number of calibrated constraints	Number of excluded constraints
POBG3_10	Place of birth—Northern Europe	SADA	220	188	Both	275	4,340
POBG3_12	Place of birth—Oceania	SADA	29	379	Both	48	4,567
POBG3_14	Place of birth—South America	SADA	151	257	Both	214	4,401
POBG3_15	Place of birth—Southeast Asia	SADA	204	204	Both	517	4,098
POBG3_16	Place of birth—Southern Africa	SADA	13	395	Both	33	4,582
POBG3_17	Place of birth—Southern Asia	SADA	181	227	Both	648	3,967
POBG3_18	Place of birth—Southern Europe	SADA	160	248	Both	173	4,442
POBG3_19	Place of birth—Union of Soviet Socialist Republics, Former	SADA	88	320	Both	142	4,473
POBG3_2	Place of birth—Caribbean and Bermuda	SADA	130	278	Both	162	4,453
POBG3_20	Place of birth—United States	SADA	171	237	Both	245	4,370
POBG3_21	Place of birth—Western Africa	SADA	65	343	Both	83	4,532
POBG3_22	Place of birth—Western Europe	SADA	169	239	Both	265	4,350
POBG3_3	Place of birth—Central Africa	SADA	11	397	Both	18	4,597
POBG3_4	Place of birth—Central America	SADA	79	329	Both	94	4,521
POBG3_5	Place of birth—Eastern Africa	SADA	61	347	Both	107	4,508
POBG3_6	Place of birth—Eastern Asia	SADA	164	244	Both	570	4,045
POBG3_7	Place of birth—Eastern Europe (excluding Union of Soviet Socialist Republics, Former)	SADA	116	292	Both	157	4,458
POBG3_8	Place of birth—West Central Asia and the Middle East	SADA	141	267	Both	330	4,285
POBG3_9	Place of birth—Northern Africa	SADA	69	339	Both	132	4,483
PPADA	Persons in the ADA	ADA	3,222	985	…	…	…
PPADACSD	Persons with geographic overlap between CSD and ADA	…	…	…	ADA	4,830	2,102
SNGLDET	Households living in a single-detached house	SADA	399	9	Both	3,639	976
TOTCFAM	Census families	SADA	367	41	Both	1,244	3,371
TOTHHLD	Households	SADA	408	0	Both	4,615	0

Table C.1
Statistics on the use of calibration constraints, by constraint

		Coverage and non-response adjustment			Final calibration		
Constraint variable name	**Description**	**Area**	**Number of calibrated constraints**	**Number of excluded constraints**	**Area**	**Number of calibrated constraints**	**Number of excluded constraints**
TOTPERS	Persons	SADA	408	0	Both	4,615	0
TPERGE15	Persons aged 15 years and older	SADA	350	58	Both	4,360	255
TPERLT15	Persons less than 15 years of age	SADA	350	58	Both	4,360	255
YRIMD_1900	Immigrants who landed prior to 1981	SADA	333	75	SADA	191	217
YRIMD_1980	Immigrants who landed from 1980 to 1985	SADA	194	214	SADA	100	308
YRIMD_1986	Immigrants who landed from 1986 to 1990	SADA	179	229	SADA	121	287
YRIMD_1991	Immigrants who landed from 1991 to 1995	SADA	202	206	SADA	97	311
YRIMD_1996	Immigrants who landed from 1996 to 2000	SADA	197	211	SADA	94	314
YRIMD_2001	Immigrants who landed from 2001 to 2005	SADA	243	165	SADA	115	293
YRIMD_2006	Immigrants who landed from 2006 to 2010	SADA	246	162	SADA	133	275
YRIMD_2011	Immigrants who landed from 2011 to 2015	SADA	233	175	SADA	168	240
YRIMD_2016	Immigrants who landed from 2016 to 2021	SADA	271	137	SADA	161	247
YRIMD_M3	Persons with no year of immigration	SADA	119	289	SADA	65	343
YRIMD_M5	Persons with no year of immigration	SADA	192	216	SADA	148	260
YRIMG1_1900	Immigrants who landed prior to 1981	SADA	333	75	SADA	191	217
YRIMG1_1980	Immigrants who landed from 1980 to 1990	SADA	254	154	SADA	143	265
YRIMG1_1991	Immigrants who landed from 1991 to 2000	SADA	251	157	SADA	127	281
YRIMG1_2001	Immigrants who landed from 2001 to 2010	SADA	300	108	SADA	142	266
YRIMG1_2011	Immigrants who landed from 2011 to 2015	SADA	233	175	SADA	168	240
YRIMG1_2016	Immigrants who landed from 2016 to 2021	SADA	271	137	SADA	161	247
YRIMG1_M3	Persons with no year of immigration	SADA	119	289	SADA	65	343
YRIMG1_M5	Persons with no year of immigration	SADA	192	216	SADA	148	260

... not applicable
CSD = Census subdivision
SADA = Super aggregate dissemination area
ADA = Aggregate dissemination area
Both = SADA and ADA
Men+ = includes men (and/or boys), as well as some non-binary persons
Women+ = includes women (and/or girls), as well as some non-binary persons
Source: Statistics Canada, 2021 Census long-form sample.

Bibliography

Devin, N., and F. Verret. 2016. “The development of a variance estimation methodology for large-scale dissemination of quality indicators for the 2016 Canadian census long form sample.” *JSM Proceedings.* Survey Research Methods Section, American Statistical Association. Chicago, United States.

Dominion Bureau of Statistics. 1968. *Sampling in the Census.* Internal report. Statistics Canada. Ottawa, Ontario.

Fellegi, I.P. 1964. “Response variance and its estimation.” *Journal of the American Statistical Association.* Vol. 59, no. 308. p. 1016–1041.

Folsom, R.E., Jr., and A.C. Singh. 2000. “The generalized exponential model for sampling weight calibration for extreme values, nonresponse, and poststratification.” *Proceedings of the Survey Research Methods Section.* American Statistical Association. p. 598–603.

Hansen, M.H., W.N. Hurwitz and M.A. Bershad. 1959. “Measurement errors in censuses and surveys.” *Bulletin of the International Statistical Institute.* Vol. 38. p. 359–374.

Judkins, D.R. 1990. “Fay’s method for variance estimation.” *Journal of Official Statistics.* Statistics Sweden. Vol. 6, no. 3. p. 223–239.

Kott, P.S. and Carr, D.A. 1997. “Developing an Estimation Strategy for a Pesticide Data Program.” *Journal of Official Statistics.* Vol. 13, no. 4. p. 367-383.

Neusy, E., and Mantel, H. 2016. “Confidence Intervals for Proportions Estimated from Complex Survey Data.” *Proceedings of the Survey Methods Section.* SSC Annual Meeting, June 2016.

Neusy E., Savard, S.-A., Hidiroglou, M., and Martin, V. 2021. “Modified Wilson Intervals for Estimated Counts with Application to Census 2021 Long Form Estimation.” Presentation to the Advisory Committee on Statistical Methods, May 2021. Internal document. Statistics Canada.

Rao, J.N.K., and J. Shao. 1999. “Modified balanced repeated replication for complex survey data.” *Biometrika.* Oxford University Press. Vol. 86, no. 2. p. 403–415.

Royce, D. 1983. *The Use of Sampling in the 1981 Canadian Census.* Internal report. Statistics Canada. Ottawa, Ontario.

Savard, S-A (2021). Inference for Census Long-Form Weighted Counts. Proceedings of the Survey Methods Section. SSC Annual Meeting, June 2021.

Shao, J., and Q. Tang. 2011. “Random group variance estimators for survey data with random hot deck imputation.” *Journal of Official Statistics.* Statistics Sweden. Vol. 27, no. 3. p. 507–526.

Statistics Canada. 2020. Canadian Census Edit and Imputation System (CANCEIS), version 5.4. Basic user guide. Ottawa, Ontario.

Statistics Canada. 2018. *Sampling and Weighting Technical Report, Census of Population, 2016*. Catalogue no. 98-306-X2016001. Ottawa, Ontario. (accessed January 6, 2023).

Statistics Canada. 2022. *Guide to the Census of Population, 2021*. Catalogue no. 98-304-X2021001. Ottawa, Ontario. Version updated November 30, 2022. (accessed January 6, 2023).

Statistics Canada. 2023. *Evaluating confidence interval methods for the 2021 Census using the Census Monte Carlo simulation environment.* Working document. Internal Document. Ottawa, Canada.

Wilson, E.B. 1927. “Probable Inference, the Law of Succession, and Statistical Inference.” *Journal of the American Statistical Association,* Vol. 22, no. 158. p. 209-212.