2021 Census Data Quality Guidelines

Census of Population, 2021

Release date: March 29, 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.

Published by authority of the Minister responsible for Statistics Canada

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

Release date: March 29, 2023
Catalogue number 98-26-0006, issue 2021001
ISBN 978-0-660-44668-4

Table of contents

1. Introduction

For the 2021 Census of Population, the dissemination strategy of quality indicators has been completely revamped, with the aim of providing more detailed information about data quality. New quality indicators are included with the data products, helping users to better assess the data quality and determine how well the available information meets their needs. The 2021 Census Data Quality Guidelines offer an overview of the available quality indicators, their definition and indications on how they can be interpreted.

A suite of quality indicators accompanies the 2021 Census of Population data outputs. The total non-response (TNR) rate associated with the geography of interest is presented with each table. For the first time in 2021, the non-response and imputation rates per question are available for more detailed standard geography in data quality indicator products. Moreover, confidence intervals are available for long-form questionnaire estimates whenever technically feasible.

The purpose of providing data quality indicators is to paint a detailed picture of the quality of the data, which can be impacted by non-response errors, processing errors and sampling errors. The indicators provided for the 2021 Census relate to data accuracy, which is one of the six dimensions of quality defined in the Statistics Canada Quality Guidelines. Accuracy is the degree to which the statistical information accurately describes what it should measure. The data quality indicators are part of the interpretability dimension of quality. They are available so that data users are informed about the quality of the statistical information provided and can determine the relevance and the limitations of the data relative to their needs.

This document provides all the information required to understand and interpret the data quality indicators for the 2021 Census. The data quality indicators are presented in detail, along with guidelines to enable their proper usage. Section 2 introduces some general considerations. Section 3 defines the TNR rate and how it compares to the global non-response (GNR) rate that was used to report non-response in past census cycles. Section 4 describes the data quality indicators based on place of residence (POR) tabulations. Section 5 discusses the quality indicators per question, and Section 6 discusses the quality indicators based on variance. Finally, Section 7 offers best practices and recommendations on the use of data quality indicators for the 2021 Census.

2. General considerations

This section provides some general considerations that apply to many or all of the quality indicators described in this document. These considerations are important to understand the context in which the data quality indicators are produced, as this influences the way they can be interpreted.

2.1 About the Census of Population

The Census of Population includes two main components, each having its own design and particularities. The two components are closely tied to the questionnaires used to collect information from the respondents: the short-form questionnaire and the long-form questionnaire. For the 2021 Census, the 2A short-form questionnaire was used to enumerate all usual residents of 75% of private dwellings. The 2A-L long-form questionnaire, which also includes the questions from the 2A short-form questionnaire, was used to enumerate a 25% sample of private households in Canada. For private households in First Nations communities, Métis Settlements, Inuit regions and other remote areas, the 2A-R questionnaire was used to enumerate 100% of the population.

The estimates produced from the responses to questions that are asked on both questionnaires (i.e., short-form content) are obtained from a **census** of population. As such, all households contribute to a given number. In this document, such estimates are referred to as “short-form estimates” (100% data). Similarly, data quality indicators calculated from all households are referred to as “short-form indicators.”

The estimates produced from the responses to at least one question specific to the long-form questionnaire are obtained from a **sample survey**. In this case, only the respondent households from the long-form sample contribute to the estimate. These estimates are referred to as “long-form estimates” (25% sample data). Similarly, data quality indicators calculated from sampled households are referred to as “long-form indicators.” Long-form estimates and indicators are weighted so that they represent the entire target population of the survey.

The **target population** of the census component is the total population of Canada, which consists of all persons who have a usual POR in Canada and certain Canadian citizens and landed immigrants who live outside of the country. This target population can be broken down into three parts: persons living in private dwellings, persons living in collective dwellings and persons living in dwellings outside Canada. The target population of the survey component consists only of persons living in private dwellings.

In most regions, the long-form sample is selected according to a stratified systematic sampling design: the long-form questionnaire is distributed to one-quarter of the households living in private dwellings and the selected households are assigned a design weight equal to four. The design weights are then adjusted to compensate for total non-response and are calibrated to chosen totals obtained from the census. The final weights are restricted to lie between 1 and 20. In First Nations communities, Métis Settlements, Inuit regions and other remote areas, the long-form questionnaire is distributed to every household. Total non-response is compensated by imputation, and households have a final weight equal to one.

Since long-form estimates are derived from a sample survey, they are subject to sampling error. Sampling variance reflects the variability in estimates because of the use of a sample as opposed to the total population. Sampling variance is therefore estimated using a statistically appropriate method, i.e., one that considers the sampling plan and the estimation strategy. Sampling variance is estimated with a replication method.

In First Nations communities, Métis Settlements, Inuit regions and other remote areas, the choice of treating total non-response by imputation instead of reweighting also has an impact on the variability of estimates. This variability is known as variance due to imputation. The replication method used for estimating sampling variance in other regions was adapted to estimate variance due to imputation in First Nations communities, Métis Settlements, Inuit regions and other remote areas. In both cases, replicate weights are used to produce variance estimates, which are used to derive confidence intervals.

Further details about weighting and estimation will be available in the *Sampling and Weighting Technical Report, Census of Population, 2021*, Statistics Canada Catalogue no. 98-306-X.

2.2 Weighting of quality indicators

The quality indicators related to long-form questionnaire data, namely the long-form TNR rate and the quality indicators per question for long-form content, are weighted so that they represent the target population of the survey and not only the units forming the sample. Weighted quality indicators provide a measure of the quality that would be expected if the entire population had been enumerated. They are more informative of the quality of the associated estimates than their unweighted counterpart. The design weight or the final weight can be used to construct weighted quality indicators; the weight used for a specific indicator is provided in its respective section.

2.3 Statistical units and population of interest

Some data quality indicators are directly associated to an estimate or to a group of estimates corresponding to the same variable of interest, such as the rates per question. In this case, the same units and population of interest are used in the calculation of estimates and of their associated data quality indicators. Detailed information about the statistical units and the population of interest by topic is provided in Appendix 1.3 of the *Guide to the Census of Population*, Statistics Canada Catalogue no. 98-304-X.

2.4 Evaluation of the quality at the estimation step

Non-response rates can be derived at the collection step or at the estimation step. The latter is generally more useful to data users interested in assessing whether the data is of sufficient quality for their needs. While collection response rates are also available at the national, provincial and territorial level for the 2021 Census, the non-response rates described in this document all reflect the estimation step.

In the census context, this means that the non-response rates are calculated considering the final classification of dwelling occupancy status. In other words, dwellings with a final status of occupied private dwelling, whether identified as occupied during collection or imputed as such in processing, count in the response rates at the estimation step.

The classification of dwelling occupancy status is based on the analysis of data collected by field operations and of data provided by respondents, and, in most areas of the country, it is further adjusted according to the results of the Dwelling Classification Survey (DCS). More information about the DCS is given in Chapter 9 of the *Guide to the Census of Population*. Once the occupancy status of dwellings is finalized, the whole household imputation procedure is used to impute data for non-respondent occupied dwellings.

2.5 Sources of error

The sources of error that are addressed by the currently available indicators are mainly non-response, imputation and sampling. There are other sources of errors, for example coverage and measurement, that are not measured by the available quality indicators but that may also affect the quality of the figures and estimates from the 2021 Census. More information about potential sources of errors in the census is provided in Chapter 9 of the *Guide to the Census of Population*. Information about coverage errors will be available in the *Coverage Technical Report, Census of Population, 2021,* Statistics Canada Catalogue no. 98-303-X.

3. Total non-response rate

Total non-response occurs when all questions are unanswered for a dwelling that received a questionnaire or when a returned questionnaire does not meet the minimum content (i.e., the information provided is not sufficient to continue processing). This type of non-response is measured by the TNR rate, which is the primary quality indicator that accompanies each disseminated output from the 2021 Census of Population. In this sense, it replaces the GNR rate, which was used for the 2016 Census of Population and for previous cycles.

The GNR rate combined total and partial non-response, while the TNR rate reflects only total non-response. Non-response is partial when answers to certain questions are not available for a respondent household. Partial non-response is now accounted for separately (see Section 5). This new approach allows for a comparison of data quality across variables that the GNR rate could not provide.

3.1 Definitions

For each geography, two different TNR rates are calculated. There are thus two definitions of the TNR rate: the unweighted TNR rate and the design-weighted TNR rate. Suppose that Q_i is a household-level variable which takes a negative value if unit i does not belong to the target population, takes a value of 0 if unit i belongs to the target population and did not respond and takes a value of 1 if unit i belongs to the target population and responded.

The unweighted TNR rate is used to calculate the short-form TNR rate. It is given by:

$$UTNR = 100 \times \left(1 - \frac{\sum_{(i:Q_i>0)} 1}{\sum_{(i:Q_i\geq 0)} 1} \right).$$

The weighted TNR rate is used to calculate the long-form TNR rate. Suppose that d_i denotes the design weight of unit i. The weighted TNR rate is given by:

$$WTNR = 100 \times \left(1 - \frac{\sum_{(i:Q_i>0)} d_i}{\sum_{(i:Q_i\geq 0)} d_i} \right).$$

3.2 Comparison with the global non-response rate used in previous census cycles

The 2021 TNR rate and the GNR rate from previous census cycles meet the same objective: to measure the scope of non-response in a given region. Conceptually, the difference observed between the GNR rate from a previous census cycle and the 2021 TNR rate for a given region can be broken down into two parts: the difference because of the change in definition and the actual difference in non-response rates between the two cycles. It is, however, impossible to describe exactly the relation between both indicators. On the one hand, the household size is taken into account in the calculation of the GNR rate but not in the calculation of the TNR rate. The impact of this difference in the definition should decrease as the population size increases. On the other hand, the GNR rate includes partial non-response and is thus generally higher than the TNR rate (though it is possible for it to be lower).

The TNR rates for the 2016 Census cycle were calculated for a comparative study of the GNR rate and the TNR rate. This study showed that there is a strong positive correlation between the two indicators and that their difference is generally less than 5%. Greater differences were observed more often for the long-form rates than for the short-form rates.

Recommendation: When the GNR rate from a previous cycle and the TNR rate from 2021 are compared, differences of less than 5% can be considered as being solely attributable to the change in definition.

In previous census cycles, areas with a GNR rate above a certain threshold were suppressed from disseminated products (the threshold used in 2016 was 50%). This type of suppression of data based on quality was abandoned in 2021.

Recommendation: Data in areas having a TNR rate above 50% should be used with caution.

4. Data quality indicators for tabulations based on place of residence geographic areas

To give a quick overview of data quality associated to a geographic area, a five-digit numeric code representing five data quality indicators is attached to each standard geographic area in the census database environment. For example, the code at the national level is 20000. These data quality indicators are included in the tabulations based on POR geographies. They can be used to identify regions for which data was suppressed for specific reasons and to get information about the level of TNR in the area. The five-digit numeric code and its components are further described below.

4.1 Incompletely enumerated reserves and settlements indicator

The first digit of the data quality indicators' five-digit numeric code indicates whether the geography of the table includes an incompletely enumerated area. In the 2021 Census of Population, as well as in previous censuses, enumeration could not be completed for some reserves and settlements. These incompletely enumerated reserves and settlements, as well as higher-level geographic areas containing these areas, are identified in the products. Although census data are not available for incompletely enumerated reserves and settlements, the areas themselves are included as part of the standard geographic hierarchies in the census database. For the list of 2021 incompletely enumerated reserves and settlements, refer to Appendix 1.5 of the *Guide to the Census of Population*.

4.2 Total non-response indicators

The magnitude of the TNR rate in the geographic area associated to a table has an impact on data quality. To guide users, its range was broken down into categories as shown in Section 4.4. The second digit of the data quality indicators' five-digit numeric code contains the category of the short-form TNR rate, and the fourth digit contains the category of the long-form TNR rate. As mentioned in Section 3.2, **data for areas having a TNR rate above 50% should be used with caution**. A note to this effect is included in the data products.

The second and fourth digits are also used when data was suppressed to meet the confidentiality requirements of the *Statistics Act*.

4.3 Income confidentiality suppression indicators

In some geographic areas, income data from the short-form or long-form questionnaire must be suppressed to meet the confidentiality requirements of the *Statistics Act*. The short-form and long-form income confidentiality suppression flags indicate whether income data were suppressed for a given area. These indicators are given, respectively, by the third and fifth digits of the data quality indicators' five-digit numeric code. The statistical disclosure control rules for income variables are different than for other types of variables and, as such, require distinct suppression indicators.

4.4 Summary of the data quality indicators for tabulations based on place of residence geographies

Table 1 below describes the data quality indicators' five-digit numeric code and its contents for tabulations from the short-form questionnaire, and Table 2 describes the indicators from the long-form questionnaire. Note that a zero in any of the five digits is the default for the respective indicator.

Table 1
2021 Census short-form data quality indicators

Digit	Description	Flag	Flag description
First (0XXXX)	Incomplete enumeration flag	0	Default. Not applicable.
		1	Incompletely enumerated reserve or settlement (suppressed).
		2	Excludes census data for one or more incompletely enumerated reserves or settlements.
Second (X0XXX)	Short-form data quality flag	0	Default. Data quality index showing a short-form total non-response rate lower than 10%.
		1	Data quality index showing a short-form total non-response rate higher than or equal to 10%, but lower than 20%.
		2	Data quality index showing a short-form total non-response rate higher than or equal to 20%, but lower than 30%.
		3	Data quality index showing a short-form total non-response rate higher than or equal to 30%, but lower than 40%.
		4	Data quality index showing a short-form total non-response rate higher than or equal to 40%, but lower than 50%.
		5	Data quality index showing a short-form total non-response rate higher than or equal to 50% (**use with caution**).
		9	Short-form data suppressed to meet the confidentiality requirements of the *Statistics Act*.
Third (XX0XX)	Short-form income confidentiality suppression flag	0	Default. Short-form income data not suppressed.
		9	Short-form income data suppressed to meet the confidentiality requirements of the *Statistics Act*.
Fourth (XXX0X)	Not applicable	0	Default. Not applicable.
Fifth (XXXX0)	Not applicable	0	Default. Not applicable.

Source: Statistics Canada, 2021 Census of Population.

Table 2
2021 Census long-form data quality indicators

Digit	Description	Flag	Flag description
First (0XXXX)	Incomplete enumeration flag	0	Default. Not applicable.
		1	Incompletely enumerated reserve or settlement (suppressed).
		2	Excludes census data for one or more incompletely enumerated reserves or settlements.
Second (X0XXX)	Short-form data quality flag	0	Default. Data quality index showing a short-form total non-response rate lower than 10%.
		1	Data quality index showing a short-form total non-response rate higher than or equal to 10%, but lower than 20%.
		2	Data quality index showing a short-form total non-response rate higher than or equal to 20%, but lower than 30%.
		3	Data quality index showing a short-form total non-response rate higher than or equal to 30%, but lower than 40%.
		4	Data quality index showing a short-form total non-response rate higher than or equal to 40%, but lower than 50%.
		5	Data quality index showing a short-form total non-response rate higher than or equal to 50% (**use with caution**).
		9	Short-form data suppressed to meet the confidentiality requirements of the *Statistics Act*.
Third (XX0XX)	Short-form income confidentiality suppression flag	0	Default. Short-form income data not suppressed.
		9	Short-form income data suppressed to meet the confidentiality requirements of the *Statistics Act*.
Fourth (XXX0X)	Long-form data quality flag	0	Default. Data quality index showing a long-form total non-response rate lower than 10%.
		1	Data quality index showing a long-form total non-response rate higher than or equal to 10%, but lower than 20%.
		2	Data quality index showing a long-form total non-response rate higher than or equal to 20%, but lower than 30%.
		3	Data quality index showing a long-form total non-response rate higher than or equal to 30%, but lower than 40%.
		4	Data quality index showing a long-form total non-response rate higher than or equal to 40%, but lower than 50%.
		5	Data quality index showing a long-form total non-response rate higher than or equal to 50% (**use with caution**).
		9	Long-form data suppressed to meet the confidentiality requirements of the *Statistics Act*.

Table 2
2021 Census long-form data quality indicators

Digit	Description	Flag	Flag description
Fifth (XXXX0)	Long-form income confidentiality suppression flag	0	Default. Long-form income data not suppressed.
		9	Long-form income data suppressed to meet the confidentiality requirements of the *Statistics Act*.

Source: Statistics Canada, 2021 Census of Population.

5. Quality indicators per question

Quality indicators per question are data quality measures specific to each question. In this context, "question" refers to questions on the short-form and long-form questionnaires and to variables related to income, immigration or mobility for which data come primarily from administrative records and from 2016 Census records rather than responses to the 2021 Census questionnaires. In this section, we distinguish between short-form rates, calculated from the short-form population, and long-form rates, calculated from the sample.

The quality indicators provided per question quantify two related sources of error in the data: non-response and imputation. Specifically, the quality indicators per question available for the 2021 Census are the non-response rate per question, the imputation rate per question and, for income variables, the impact of imputation per question. This section presents details about their availability, their definitions and their particularities, as well as guidelines on how they can be interpreted. More information about how they can be interpreted together is given in Section 7.

5.1 Availability

Quality indicators per question are calculated at some specific levels of geography. For the 2021 Census, they are publicly available in data tables specifically for data quality for the following standard geographic hierarchies:

- Canada, provinces and territories, census metropolitan areas (CMA), census agglomerations (CA) and census subdivisions within CMAs and CAs
- Canada, provinces and territories, census divisions, census subdivisions.

The quality indicators per question could also be obtained for other levels of geography through custom requests.

For the 2016 Census, the imputation rate per question and the impact of imputation per question were available at the national, provincial and territorial level in the respective reference guides. The non-response rates per question are new for the 2021 Census.

The available quality indicators per question are presented in Appendix 1. For variables that are derived from responses to more than one question, data quality indicators are not directly available. In these cases, users should refer to those indicators related to questions covering the topic of interest. For further information, refer to the relevant Reference Guides.

5.2 General definitions

The definitions used to produce the data quality indicators per question are given in this section. Guidelines on their interpretation are presented in the sections that follow. The calculation of the data quality indicators per question requires unit-level indicator variables associated with each question. Units are either persons or households depending on whether the question is about a person-level characteristic or a household-level characteristic. Suppose Y is the variable of interest associated to a question and y_i is the value it takes for unit i. Suppose also that Z is an indicator variable associated to Y and z_i is the value it takes for unit i.

For the calculation of the non-response rate, the indicator variable Z indicates whether the unit was a respondent or a non-respondent to the question: $z_i < 0$ if unit i is out-of-scope for variable Y, $z_i = 0$ if unit i responded and $z_i = 1$ if unit i did not respond. More detail is given about non-response in Section 5.3.

For the calculation of the imputation rate, the indicator variable Z indicates whether the response to the question was imputed or not for the unit; $z_i < 0$ if unit i is out-of-scope for variable Y, $z_i = 0$ if the response for unit i was not imputed and $z_i = 1$ if it was imputed. More detail is given about imputation in Section 5.4.

The quality indicators per question are calculated to be coherent with the set of units that is considered to be in-scope for each question. A unit is considered to be in-scope for a given question if the question is applicable to that unit and the unit belongs to the population of interest related to the question (see Section 2.3). In-scope units are the units that contribute to estimation.

5.2.1 Short-form rates

Short-form rates are calculated from the whole population. The available short-form rates are the non-response rate and the imputation rate. Using the above notation, the short-form rates are given by the following general formula:

$$NR \mid IMP = \frac{\sum_{(i:z_i=1)} 1}{\sum_{(i:z_i \geq 0)} 1}.$$

The non-response rate per question on the short-form questionnaire is given by the number of in-scope units in the population of interest who did not respond to the question divided by the number of in-scope units in the population of interest.

The imputation rate per question on the short-form questionnaire is given by the number of in-scope units in the population of interest for which the response to the question was imputed divided by the number of in-scope units in the population of interest.

5.2.2 Long-form rates

Long-form rates are calculated from the long-form sample. The available long-form rates are the non-response rate, the imputation rate and the impact of imputation. Using the above notation, the following general formula is used for long-form non-response and imputation rates:

$$NR \mid IMP = \frac{\sum_{(i:z_i=1)} w_i}{\sum_{(i:z_i \geq 0)} w_i},$$

where w_i is the final weight of unit i.

The long-form non-response rate per question is given by the sum of the final weights of in-scope units in the population of interest who did not respond to the question divided by the sum of the final weights of in-scope units in the population of interest.

The long-form imputation rate per question is given by the sum of the final weights of in-scope units in the population of interest for which the response to the question was imputed divided by the sum of the final weights of in-scope units in the population of interest.

The impact of imputation involves a continuous variable of interest Y. In some cases, the value of this variable is obtained by taking the sum of various components. Suppose that Z^* is a variable associated to Y and z_i^* the value it takes for unit i, such that $z_i^* < 0$ when unit i is out of scope for variable Y and z_i^* is between 0 and 1 otherwise.

Variable Z^* indicates to what extent the components of variable Y were imputed. More precisely, $z_i^* = 0$ if none of its components were imputed for unit i, $z_i^* = 1$ if all of its components were imputed and z_i^* takes a value between 0 and 1 if some but not all of its components were imputed. For variables of interest that are not obtained from separate components, z_i^* either takes a value 0 or 1.

For a continuous variable of interest Y, the impact of imputation is given by the following formula:

$$IMPACT = \frac{\sum_{(i:z_i^*>0)} w_i z_i^* y_i}{\sum_{(i:z_i^*\geq 0)} w_i y_i}.$$

5.3 Non-response rate per question

The non-response rate per question as defined above is a measure of missing information because of non-response to a specific question. If a response to a question is not provided for a given person or household, this could be due to total non-response or to partial non-response. Non-response is said to be total when no questionnaire is returned from a household or when a returned questionnaire does not meet the minimum content. Non-response is partial when answers to certain questions are not available for a respondent household.

The types of non-response taken into account by the non-response rate per question differ between the types of questionnaires. For short-form rates, both partial and total non-response contribute to the non-response rate. For long-form rates, only partial non-response is included, except in First Nations communities, Métis Settlements, Inuit regions and other remote areas where partial non-response and total non-response are taken into account.

This is because total non-response is treated differently for the two types of questionnaires. Total non-response to the short form is imputed, whereas the treatment of total non-response to the long-form questionnaire depends on the geographical area of the dwelling. In areas where the sampling fraction is equal to one-quarter, total non-response is compensated by reweighting the respondent households so they represent the non-respondents. In areas where all households are part of the long-form questionnaire sample (First Nations communities, Métis Settlements, Inuit regions and other remote areas), total non-response to the long-form is instead treated by imputation.

Interpretation: Generally, the non-response rate per question can be interpreted as the proportion of in-scope units in the population of interest for which the information was missing because of non-response. Long-form rates are weighted to reflect the fact that the long-form questionnaire is only distributed to a sample of the population, so in this case the proportion is estimated.

5.4 Imputation rate per question

The imputation rate per question provides a measure of the extent to which responses to a given question were imputed. Imputation is used to replace missing data in the event of non-response or when a response is found to be invalid. Imputation is also used for variables where data are obtained from administrative records. In cases where an administrative record cannot be linked to a respondent to provide the necessary information, imputation is applied to fill in the missing values. When carried out appropriately, imputation should reduce non-response bias.

Various imputation methods were used in the treatment of the 2021 Census data. These mostly random methods use the reported values of respondents to fill in missing information. Deterministic edits are not considered imputation and are not accounted for in the imputation rate per question or in the impact of imputation per question. When the problem is clear and unambiguous (i.e., there is only one reasonable value), deterministic edits assign a specific value to resolve the problem. This method of editing the data is used in certain situations to treat partial non-response and inconsistent responses to questions.

Interpretation: Generally, the imputation rate per question can be interpreted as the proportion of in-scope units in the population of interest for which the information was imputed rather than reported. It does not take deterministic edits into account. Long-form rates are weighted to reflect the fact that the long-form questionnaire is only distributed to a sample of the population, so in this case the proportion is estimated.

5.5 Impact of imputation per question

The impact of imputation is a measure that is available for income concepts.

Interpretation: The impact of imputation per question can be interpreted as the proportion of the total of the variable for which values were imputed. Like the imputation rate per question, the impact of imputation does not take deterministic edits into account. For variables that are derived from various components, the impact of imputation also takes into account the proportion of components that were imputed.

The impact of imputation incorporates the values of a variable, as opposed to measuring only the fraction of in-scope units for which the variable was imputed, as is the case for the imputation rate. The largest imputed values of the variable contribute more to the impact of imputation for that variable than its smallest imputed values. For instance, an imputed employment income value of $200,000 per year contributes more to the impact of imputation for employment income than an imputed value of $30,000 per year.

Some income variables can have negative values. In these cases, imputed values can also be negative. Negative imputed values tend to decrease the impact of imputation of a total positive value and could even lead to a negative impact of imputation. This indicator is more difficult to interpret in such cases. Alternative quality indicators, such as the imputation rate, or the impact of imputation of absolute income may be more helpful for interpretation. These indicators are not published but are available upon request.

6. Quality indicators based on variance

The variance is a measure of uncertainty of an estimate produced from a sample. As it is difficult to interpret, surveys typically provide other quality indicators derived from the variance estimator to data users, namely standard errors, coefficient of variation or confidence intervals. More detail about variance estimation is given in Section 2.1.

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.

6.1 Standard error

The standard error associated with an estimate is the square root of its estimated variance. It has the same scale as the estimate itself.

6.2 Coefficient of variation

The coefficient of variation associated with an estimate is the ratio of the standard error to the estimate. It is a standardized measure that is expressed as a percentage of the estimate.

6.3 Confidence interval

A confidence interval is associated with a confidence level. A default confidence level is generally set for a survey or in a field of study based on user needs. For the census dissemination system, the default confidence level was set to 95%. A 95% confidence interval is an interval constructed around the estimate so that if the process that generated the sample were repeated many times, the value of the parameter of interest in the population would be contained in 95% of these intervals.

The usual confidence interval, also known as the Wald interval, assumes that the sampling distribution of the estimator is a normal distribution. The Wald 95% confidence interval is constructed by subtracting or adding approximately twice the standard error to the estimate. When the sample size is small, as well as for certain statistics such as proportions and counts, the assumption that the estimator distribution is normal is often violated. Therefore, a confidence interval constructed in this manner is not appropriate; its true confidence level is less than the nominal 95% confidence level indicated.

Consequently, the confidence intervals presented with the 2021 Census of Population long-form estimates are produced using more elaborate methods that offer a true confidence level closer to the nominal level. The methods used to produce confidence intervals are described below. Although confidence intervals based on these methods generally have good properties, all confidence intervals are based on assumptions that cannot be verified. Further details on the different methods used to construct confidence intervals and their assumptions are provided in the *Sampling and Weighting Technical Report, Census of Population, 2021*, Statistics Canada Catalogue no. 98-306-X.

6.3.1 Student's confidence interval

The Student's confidence interval is used for all statistics except proportions and counts. It is based on Student's t-distribution which has one parameter—the number of degrees of freedom. When the number of degrees of freedom is very large, Wald and Student intervals are approximately identical. But this is often not the case with the census long-form estimates. 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. In the 2021 Census, the degrees of freedom were approximated by the number of replicates used for variance estimation and denoted as R.

The lower bound (LB) and the upper bound (UB) 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 97.5^{th} percentile of the Student's t-distribution with R degrees of freedom
- $\widehat{SE}(\hat{\theta})$ is the standard error of $\hat{\theta}$.

6.3.2 Modified Wilson confidence interval for proportions

The modified Wilson confidence interval method is used for proportion-type statistics. The LB and the UB of a 95% modified Wilson confidence interval for a proportion-type statistic 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

- $\hat{p}$ is the estimate of p
- t is the 97.5^{th} percentile of the Student's t-distribution with R degrees of freedom
- $n_e = \min(\mathrm{n}/\mathrm{deff}(\hat{p}),\ n)$ is the effective sample size
- $\mathrm{deff}(\hat{p}) = \frac{\hat{V}(\hat{p})}{\hat{p}(1-\hat{p})/\mathrm{n}}$ is the estimated design effect
- n is the in-scope sample size
- $\hat{V}(\hat{p})$ is the estimated variance of $\hat{p}$.

Theoretical developments and extensive simulation studies[1,2] have shown that this method has good properties in most situations and performs better than Wald and Student's confidence intervals when hypotheses do not hold.

6.3.3 Modified Wilson confidence interval for counts

The modified Wilson confidence interval method is used for estimated counts. The LB and the UB 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

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

1. 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, 367-383.
2. 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.

Theoretical developments and extensive simulation studies[3] have shown that this method has good properties in most situations and performs better than Wald and Student's confidence intervals when hypotheses do not hold.

6.4 Interpretation using confidence interval

When given a 95% confidence interval for a parameter of interest, a user can say that they are 95% confident that the true population parameter lies within the interval. For example, if the mean employment income estimate is $40,000 and its 95% confidence interval ranges from $35,000 to $45,000, the user can say that they are 95% confident that the mean employment income of the population lies between $35,000 and $45,000.

7. Best practices and recommendations

This section provides more detail about how the indicators can be used together and how they relate to each other, as well as general interpretation recommendations.

7.1 Comprehensive strategy

Taken together, the available data quality indicators provide information about the overall quality of the 2021 Census figures and estimates. To have the best possible picture of data quality, data users should consult the entire suite of relevant indicators. As a reminder, the TNR rate and the five-digit data quality indicators are provided with each table. For long-form questionnaire tables, the confidence intervals are usually included along with the estimates. Data quality indicators by question, including non-response and imputation rates by question, are available in data tables specifically for data quality.

7.2 Interpretation of non-response rates

Non-response is a potential source of bias in census counts and long-form estimates. Bias occurs when the characteristics of respondents differ from those of non-respondents. Unfortunately, bias cannot be directly measured as the characteristics of non-respondents are generally unknown.

The TNR rate and the non-response rates per question indicate the risk, and its potential magnitude, that a significant bias may be introduced by non-response. For a given profile of non-respondents, a lower non-response rate indicates a lower risk of non-response bias and, therefore, more reliable figures and estimates.

Both the TNR rate and the applicable non-response rate(s) by question should be consulted as they may offer different perspectives on data quality. Consider a region where the TNR rate is high and the non-response rate is low for a specific question. This may happen when a question is out-of-scope for a substantial subgroup of the population that had a low response rate. In such a situation, there is a risk of bias for related characteristics. For example, if the TNR rate is high for unemployed individuals and there are many unemployed individuals in the region, the TNR rate will be high. If at the same time employed individuals responded well, the non-response rate by question for labour questions could be low. Because there is a risk of bias, users should interpret labour data in this region with caution.

Conversely, consider a region where the TNR rate is low and the non-response rate by question for a specific question is high. If the imputation rate for the question is also high, it may indicate that there is a risk of bias for the characteristic of interest. In this case, users should interpret data with caution even if the TNR rate is low.

3. 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.

When comparing the TNR rate and the non-response rate per question, users should be aware of differences in their definition. First, they are not based on the same statistical units: the units used in the calculation of the TNR rate are households, and the units used in the calculation of the non-response rates per question can be either persons or households depending on the question. Moreover, the denominator for the TNR rate is the entire target population, whereas the denominator for the non-response rate per question is the subset of units that are in-scope for the question and are part of the population of interest.

7.3 Interpretation of imputation rates

The imputation rate indicates whether the quantity of imputed values is large relative to the quantity of reported values. The impact of imputation can be interpreted similarly by replacing the quantity by the sum. Generally, the higher these rates, the more reason there is to question the quality of the estimates and the potential for bias.

However, the rates themselves do not indicate the level of quality of the imputed data. If possible, evaluating the imputation strategy provides additional information about the quality of the estimates. When imputation models are based on the use of auxiliary information well correlated with the characteristic of interest (the preferred approach for the 2021 Census of Population) then one can conclude that the imputed values are quite accurate. In this case, a high imputation rate will not necessarily imply that the quality is questionable.

7.4 Relationship between the non-response rate and the imputation rate

As imputation is used to treat non-response, the imputation rate per question is strongly related to the non-response rate per question. Nonetheless, there are circumstances in which these two rates are not equivalent.

For instance, a unit could be considered respondent and imputed if the reported response was found to be invalid during treatment. If this is often the case for a question, it can cause the imputation rate to be higher than the non-response rate.

Inversely, it is possible for a unit to be considered neither respondent nor imputed. This happens when deterministic edits are applied to treat cases of non-response that can be resolved in a unique way. If this is often the case for a question, it can cause the non-response rate to be higher than the imputation rate.

8. Conclusion

In summary, there are many data quality indicators available in data products for the 2021 Census of Population. The TNR rate is provided in each table and further information on quality at the area level is available for tabulations based on POR geographies. The non-response and imputation rates per question are available for standard lower levels of geography in separate tables. Finally, for the majority of long-form questionnaire estimates, confidence intervals are also available.

Data quality indicators are provided so that users can assess the relevance of the data to their needs. Their definition as well as guidelines for their interpretation were presented in this document. In general, the quality of the 2021 Census of Population data is very good, but in some cases data have to be used with caution. It is strongly recommended that users consult all available data quality indicators to get a better sense of the quality of the data products in which they are interested.

Appendix 1

Table A1.1
Data quality indicators by question available for short-form content (excluding income questions)

Topic	Question	Non-response rate	Imputation rate	Impact of imputation
Age, sex at birth and gender	Age	Yes	Yes	No
	Sex at birth	Yes	Yes	No
	Gender	Yes	Yes	No
Marital status	Marital status (legal)	Yes	Yes	No
	Common law	Yes	Yes	No
Families and households	Relationship to Person 1	Yes	Yes	No
Language	Knowledge of official languages	Yes	Yes	No
	All languages spoken at home	Yes	Yes	No
	Language spoken most often at home	Yes	Yes	No
	Mother tongue	Yes	Yes	No
Canadian military experience	Canadian military experience	Yes	Yes	No
Type of dwelling (collective or private)	Structural type of dwelling	Yes	Yes	No

Source: Statistics Canada, 2021 Census of Population.

Table A1.2
Data quality indicators by question available for income topic

Question	Non-response rate	Imputation rate	Impact of imputation
Short-form content			
2020 Total income	Yes	No	No
2020 Market income	Yes	No	No
2020 Employment income	Yes	No	No
2020 Government transfers	Yes	No	No
2020 After-tax income	Yes	No	No
2019 Total income	Yes	No	No
2019 Employment income	Yes	No	No
Long-form content			
2020 Total income	Yes	No	Yes
2020 Market income	Yes	No	Yes
2020 Employment income	Yes	No	Yes
2020 Government transfers	Yes	No	Yes
2020 After-tax income	Yes	No	Yes
2019 Total income	Yes	No	Yes
2019 Employment income	Yes	No	Yes

Source: Statistics Canada, 2021 Census of Population.