Final report of the
2018 Census External Data Quality Panel

Acknowledgements

The 2018 Census External Data Quality Panel would like to thank staff at Stats NZ for their openness and responsiveness and for their help with providing the data, analysis and reports needed to compile the panel's initial report, this report and the assessments of variables report completed by the panel. We also thank an independent reviewer who provided useful comments on all of the panel's reports.

We were supported throughout the review by a small secretariat provided by Stats NZ.

January 2020

18 February 2020 corrections: We have made the following corrections to this document on the request of the External Data Quality Panel.

- Page 33, top paragraph: changed 'iwi' to 'absentees' in '... the only Priority 1 variable rated as poor or very poor was iwi'.
- Page 38, in the paragraph beginning 'For the measurement of ...': changed the year 1979 to 1997.
- Page 51, table 4.2: added 'high' as the panel internal consistency rating for L1 Ethnic group 'European'.

Citation
2018 Census External Data Quality Panel (2020). *Final report of the 2018 Census External Data Quality Panel*. Retrieved from www.stats.govt.nz

ISBN 978-1-98-858378-5

Published in February 2020 by
Stats NZ Tatauranga Aotearoa
Wellington, New Zealand

Introduction to the 2018 Census External Data Quality Panel and its reports

Stats NZ constituted the 2018 Census External Data Quality panel in August 2018.

Panel members are as follows:

- Richard Bedford, Emeritus Professor, recently retired Professor of Population Geography, Auckland University of Technology and University of Waikato (co-Chair)
- Alison Reid, Team Manager, Economic and Social Research and Evaluation, Auckland Council (co-Chair)
- Dr. Barry Milne, Director, COMPASS Research Centre, University of Auckland
- Dr. Donna Cormack, Senior Lecturer, Te Kupenga Hauora Māori, University of Auckland; Senior Research Fellow, Te Rōpū Rangahau Hauora a Eru Pomare, University of Otago, Wellington
- Ian Cope, international census expert, ex-Office of National Statistics (ONS), United Kingdom
- Len Cook, former New Zealand Government Statistician and former National Statistician of the United Kingdom
- Tahu Kukutai, Professor of Demography, National Institute of Demographic and Economic Analysis, University of Waikato
- Thomas Lumley, Professor of Biostatistics, University of Auckland.

Objectives

As set out in the Terms of Reference, the objectives of the panel were to provide independent advice to the Government Statistician about:

- whether the methodologies used to produce quality information from the census are based on sound research and a strong evidence base
- approaches to data processing and methodology, and increased use of administrative sources that affect the quality of the data
- data issues that may affect the usefulness of the data for Māori and iwi as Treaty partners
- any quality issues people will need to consider when using 2018 Census and related population data, and any further work required to assist customers.

Reporting

Originally, the panel was to convene from August 2018 until April 2019, and to produce a report at the time of the first release of data in April 2019. The delay in the first release of the results of the 2018 Census of Population and Dwellings (the 2018 Census) necessitated deferral of reporting.

On 23 September 2019, Stats NZ released the first official results of the 2018 Census. The panel released a substantive report to coincide with this release (*Initial Report of the 2018 Census External Data Quality Panel*). The full report is available here, and a brief summary of the key messages and some recommendations from it can be found below in Section 1 of this report.

Between September 2019 and January 2020, the panel continued its assessment of the quality of data for a wide range of variables relating to individuals and dwellings. In December 2019, a report assessing the quality of data for 31 variables (see 2018 Census external data quality panel: Assessment of variables) was completed. In January 2020, this report was completed, which marks the end of the work of the 2018 Census External Data Quality Panel.

In addition to these reports, there are two other outputs that the panel has prepared for users of 2018 Census data:

1) Interactive graphs of data sources for key 2018 Census individual variables, data sources for 2018 Census variables by geography and ethnicity, a series of plots showing the extent to which data for 2018 Census variables were sourced by responses to the 2018 Census or from other sources (administrative data, 2013 Census data, forms of imputation). Plots are produced down to SA2 level of geography, and for Level 1 of the ethnicity classification. These are available from 2018 Census external data quality panel: Data sources for key individual 2018 Census variables.
2) *Te reo data in the 2018 Census*, a short report describing the quality of te reo data in the 2018 Census, including guidelines and cautions regarding the use of te reo data. This report will be completed in March 2020.

Readers should note that the panel's quality assessments are based on the information that was available when the reports were compiled (August 2019 – December 2019). As at January 2020, Stats NZ are continuing to undertake assessments of the quality of a number of variables. It is not possible for the panel to take these ongoing assessments into account, nor can we speculate whether these ongoing assessments would change the panel's quality assessment or rating for any variable.

Coverage of the Terms of Reference

Of the nine focus areas listed in the Terms of Reference, the following five were covered in the panel's initial report which, as noted above, was published to coincide with Stats NZ's first release of 2018 Census data on 23 September 2019:

- a fit-for-purpose census file, taking into account methodological choices in its construction
- methodologies and implementation, including imputation and the use of administrative data sources
- use of data for electoral purposes, including Māori electorates
- impacts, issues, and implications of the data for Māori and iwi as Treaty partners
- census measurement estimation methodology.

The remaining focus areas were:

- demographic analysis and implications for key census data uses and customers
- data processing and evaluation problem reports
- census topic evaluations
- census product and services release schedule and metadata.

Following discussion with Stats NZ it was agreed that the following Terms of Reference would be addressed as follows:

- Stats NZ have not carried out demographic analyses, so the panel has not covered this topic in its reports.
- The panel reports briefly on data processing and evaluation problem reports in Section 9.
- No census topic evaluations had been completed by December 2019, so the panel has not been able to assess or report on them.
- Census product and services release schedule and metadata. There is a brief discussion in Section 9 of the data and metadata. The data release schedule, as this was known in December 2019, is summarised in Appendix 2.

The panel has seen no results from the 2018 Census Post Enumeration Survey and so has not commented on the 2018 PES in any of its reports.

Executive Summary and Recommendations

Executive summary

At the time this report was submitted for publication in January 2020, Stats NZ was finalizing its release schedule for census products that allow users to produce their own tabulations and statistical analyses using census data (see Appendix 2 for a summary of release dates for 2018 Census products).

In early 2020, Stats NZ will release 2018 Census data for use in the Data Lab, a distributed set of secure computer sites where users can access, with approval, microdata for research purposes. By March 2020, 2018 Census data will be in the Integrated Data Infrastructure (IDI) and available for use by approved researchers and policy analysts who wish to negotiate access to the IDI.

A very different data file

The data files for the 2018 Census have been constructed in a very different way from those for previous censuses. The impact on the quality of the 2018 Census data of a much lower than expected (88 percent) and highly uneven response rate across population sub-groups (e.g. 74 percent for Pacific populations) and areas became the focus of considerable inquiry both within Stats NZ and externally.

Stats NZ established the 2018 Census External Data Quality Panel (EDQP or 'the panel') in August 2018 to provide an independent assessment of the methods Stats NZ was employing to compensate for the lower than expected response rate to the census. The panel's job has been to assess the impact that application of these methods has had on the quality of data that is contained in the official census data file.

The panel's reviews of and reflections on the quality of the 2018 Census data can be found in three reports:

1) EDQP (2019a) Initial Report of the 2018 Census External Data Quality Panel, Stats NZ, September 2019.
2) EDQP (2019b) 2018 Census External Data Quality Panel: Assessment of variables , Stats NZ, November 2019
3) EDQP (2020) Final report of the 2018 Census External Data Quality Panel, Stats NZ, February 2020.

A summary of the quality ratings Stats NZ and the panel have assigned the variables about people and dwellings can be found in Section 2.

Guidance in the use of 2018 Census data

Policy makers, researchers and members of the public making use of the 2018 Census data need to have a good understanding of quality-related issues and impacts. The first substantive section in this report, Section 3 – Data quality considerations and some guidelines for users of the 2018 Census, contains important insights into how Stats NZ and the panel have assessed the quality of census data and some of the major issues those wishing to use these data need to consider and address.

In order to offset some of the loss in quality due to the low overall completion rate for census questionnaires compared to earlier censuses, Stats NZ has had to put in place a new methodology for producing the 2018 Census dataset. The panel has expressed its confidence in the methods used to ameliorate the situation in its initial report.

Section 3 covers issues such as improved coverage of key variables in 2018 compared to previous censuses, the impact of the uneven response rate on data for small areas and different ethnic groups, the challenge of breaks in the time series with previous censuses, the much wider use of imputation in 2018, and the impact of lower response (and the strategies adopted to compensate for this) on key analytical units such as families and households.

The fitness for use of information relating to characteristics that can be obtained only from census questionnaires cannot be presumed as in the past – it needs to be evaluated for specific applications. Supporting this need, Stats NZ has made available a wide range of quality measures about each characteristic such as in the DataInfo+ pages for each variable (Section 2 – Summary of quality ratings for variables).

In its separate report, 2018 Census External Data Quality Panel: Assessment of Variables, the panel has provided critical reviews of the information available on the quality of 31 variables detailing characteristics of individuals and dwellings. There are also detailed discussions of selected variables in its Initial Report and in this report.

A set of bullet points at the end of Section 3 contains a useful summary set of guidelines for users of the 2018 Census data. The final bullet point contains a very important message for those wishing to cross-tabulate 2018 Census data, or subject these data to analysis that uses multiple variables simultaneously. Users need to "consider the quality of each 2018 Census variable in isolation and then when they are taken together. For example, while data on age is of very high quality, age cross-tabulated by ethnicity may not be. The quality of such analyses will critically depend on which level of the ethnicity classification is used, and which specific ethnicities are being analysed". We discuss this issue in several places in this report.

Data on specific ethnic groups

New Zealand's Census of Population and Dwellings is the only source of data on most of the level 4 (L4) ethnic groups that are identified in Stats NZ's (2005) *Ethnicity New Zealand Standard Classification V2.0*. The classification is at four levels and in the panel's initial

report (Initial report of the 2018 Census External Data Quality Panel) there is a detailed discussion of the classification at levels 1 (L1) and 2 (L2) in section 5.

In Section 4 – Ethnicity and birthplace, further considerations the panel focuses on the level in the ethnicity classification (level 4) where most of New Zealand's specific ethnicities can be identified and their characteristics can be analysed. It is at this level that members of the groups who self-identify with particular ethnicities can find information relating to themselves in the census. There is a lot of interest amongst members of these groups, as well as amongst analysts and policy makers in central and local government and researchers in the public and private sectors, in the census data on characteristics of specific ethnicities.

In Section 4, the quality of level 4 (L4) ethnic data is examined in association with data on birthplace in order to differentiate between the New Zealand-born and the overseas-born components of specific populations defined on the basis of their self-declared ethnic identity/ies. Two key findings relating to the quality of these data are:

1) The full range of quality categories from **very high** to **very poor** apply to specific ethnic groups when assessing the quality of data at the L4 level of the classification. Stats NZ have rated the overall quality of data for the L1 and L2 ethnic groups as **high**. In its Initial Report the panel, using a broader definition of quality, rated the data as being of **moderate** quality**.** Once users start accessing data at levels 3 and 4 of the classification, they need to be aware that neither of these single measures of quality is appropriate for all groups.
2) In the light of evidence presented in Section 4, and in the panel's first two reports, it is very important that users of the 2018 Census data keep in mind that the general assessments of quality that have been produced by Stats NZ are designed to provide an overall indication of how well the data for the specific variables have been counted. The measures of quality are calculated at the lowest levels of the classification for many variables, but there are some exceptions including ethnicity and religion. These measures are calculated at the national level and the resultant ratings are not necessarily useful guides to quality when data are being examined for small areas or small population sub-groups.

Data on languages spoken with reference to te reo Māori

Census data on languages spoken, like ethnicity, are coded using a multi-level classification system. Stats NZ have given the languages spoken overall a quality category of **high**. This is a combined rating for all languages, but results are dominated by the English language. Stats NZ have given the 2018 Census data on te reo Māori, an official spoken language in New Zealand, a quality rating of **poor** on the basis of its Metric 1 score. As with the specific ethnicities at level 4 of the ethnic classification, the quality of the data for specific languages identified at level 4 of the languages spoken classification ranges from **high** to **poor**.

In Section 5 – Languages spoken**,** the panel concurs with the **poor** quality rating for 2018 Census data on te reo Māori. They point out that this is a critical failing of the 2018 Census

in the context of Te Tiriti o Waitangi, Stats NZ's obligation to provide robust data for Māori under Te Tiriti, and the Government's international human rights obligations relating to te reo Māori.

While use of te reo Māori is not legislatively required to be collected in the census, high quality time-series data on te reo is necessary for monitoring the health of the language, and for supporting the work of national bodies such as Te Mātāwai (the independent statutory entity charged with revitalising te reo Māori) and Te Taura Whiri i te reo Māori (Māori Language Commission), as well as government agencies, iwi and Māori communities. The poor quality of te reo data from Census 2018 is a major loss given the lack of nationally representative te reo data outside of the census with which to undertake time-series analysis.

On the basis of its analysis of the quality of data on te reo Māori, the panel does not judge the Census 2018 te reo data to be of sufficient quality to be used in conjunction with te reo data from previous censuses to undertake time series analysis. This would lead to very dubious findings of a significant increase in the number of te reo speakers between 2013 and 2018. This increase is an artefact of the methods used to compensate for the very low 2018 Census response rate for Māori, not evidence of increasing numbers of te reo speakers, or of increasing use of te reo Māori in the home.

Data on families and households

There are 29 variables or derived variables in the suite of family and household outputs that have been generated from 2018 Census (see Appendix 4). At the time of finalising this report Stats NZ rate these data as being of **very poor** quality. The panel is aware that further analysis is in progress and that this suggests that the quality of the data may be better than previously thought.[1] In Section 6 – Families and households the panel discusses the quality of the families and households variables on the basis of information available to members up to 6 December 2019.

In the warrant of fitness (WoF) for the families and households variables, Stats NZ state that little household data is fit for publication, except perhaps for counts of families at the national level and counts of households at the national and regional council levels. In their report on the potential for linked administrative data to provide household and family information, Stats NZ point out that: "Overall, the admin sources investigated show potential for providing household information on an aggregate level, despite some

[1] At the time of completing this report in January 2020 Stats NZ were working on better understanding the error in the families and households data, and potentially revising the quality ratings for some variables based on this investigation. Stats NZ note that some minor fixes may also be undertaken to improve family and household data quality. Users of data on families and households need to check with Stats NZ on the revised quality ratings and any changes made – the panel has not been in a position to assess these.

limitations. However, the lack of coverage of families in admin data means the potential for producing census-type information on families is currently minimal. Missing family information also affects our ability to replicate the census household composition variable."[2]

Unless Stats NZ can improve the quality of the families and households there will be a significant loss of information about families and households that is vital for public policy, for meeting Treaty obligations to Māori[3], for population projections and for the derivation of analytical measures such as household crowding and social deprivation.

In a recent paper shared with the panel on 10 December 2019, Stats NZ's re-evaluation indicates that the suite of families and households variables merit a quality rating of **high-moderate** depending on the variable (Metric 1), or **moderate-very poor** (Metric 3)**,** not their initial overall rating of **very poor**[4] for the full suite of variables. This is a very significant shift and the panel encourages Stats NZ to publish a paper on the website as soon as possible that explains to users what accounts for this very different assessment of data quality for the suite of variables.

Data on small areas

The census uniquely enables data on a range of census variables to be produced at small areas (and also for small population subgroups). Surveys can never produce the micro-data at a small area level that is typically available from the census. Central government, local government, businesses and community groups use highly localised data to allocate resources to local areas, to plan and deliver services (e.g. identifying future demand for care for the elderly), for retail store location decisions, and for local advocacy etc.

Section 7 – Small area considerations addresses implications for data quality of a highly variable response to the 2018 Census at the SA2 level of geography. As geographies become smaller, variation increases, with some areas with higher levels of response from census individual forms than the national average, and some with lower levels. Some small areas (SA2s, typically with populations of 1000–4000 people) were particularly affected by low response rates to the census field collection, and as such have lower quality data across many variables. For example, while 85 percent of the census dataset is made up of people who returned census individual forms, for the variable usual residence one year ago there were 24 SA2 areas (1 percent of the total) with less than 60 percent of data from census

[2] https://www.stats.govt.nz/research/the-potential-for-linked-administrative-data-to-provide-household-and-family-information

[3] For example *Stats NZ's Strategic Intentions 2019-2023* notes that Stats NZ, as the leader of the government data system, needs to ensure that decision-makers have the data and information they need to make decisions, and that there is "improved representation of Treaty partners in our data and products". https://www.stats.govt.nz/corporate/stats-nzs-strategic-intentions-201923

[4] Stats NZ (2019a) Quantifying data quality issues for household and family information, unpublished paper, 5 December 2019.

individual forms. There were a further 92 SA2s with between 60 and 70 percent of information coming from census individual forms.

The great majority of these areas are located in the North Island with Auckland accounting for 45 percent of the 116 SA2s where there was a response rate of 70 percent or less. Most of the low response SA2s have high Māori and/or Pacific populations, highlighting what the panel stated in its initial report, that Māori and Pacific were the worst affected by the low response to the 2018 Census.

Low response rates have been compensated for, in part, by incorporating data from the 2013 Census, a range of administrative sources, or by imputation methods. The mixes of data derived from these sources, and the prevalence of residual "no information" cases, varies considerably by variable. This has a major impact on data quality at the SA2 level. Examples of this issue for three variables with different mixes of alternative data sources and prevalence of "no information" cases are given for the 24 SA2s with response rates under 60 percent in Section 7.

The extent of 'no information' and use of alternative data sources for selected variables for the lowest responding SA2s in each region of New Zealand is presented in a series of graphs developed by the panel that can be found on 2018 Census external data quality panel: Data sources for key 2018 Census variables. The panel encourages users wishing to understand specific communities and small areas to consult these data sources graphs to get an indication of how data quality for the variables in which they are interested varies across New Zealand.

Data of 'poor' and 'very poor' quality

Stats NZs quality rating framework (see Appendix 3) covers three metrics: data sources and coverage, consistency and coherence, and data quality. Each quality metric is rated separately, and Stats NZ set the overall quality rating as the lowest of the three metric ratings.

The most important assessment of the quality of any variable has to be that at the level of the classification that will be used frequently by users. This will often be the lowest level of the classification for which information is available. This is particularly important for hierarchical classifications including ethnicity, birthplace, religion, language, occupation and industry.

Stats NZ consider that it is acceptable to generally release data at the appropriate geographic levels and levels of the relevant classification that has been assessed as having a **very high, high**, or **moderate** quality rating overall. Stats NZ will release data rated as **poor** quality if it passes an assessment process prior to release. Given that it is clearly desirable to release as much data as possible for use by users, this has to be weighed against the risk of perhaps inadvertent uses of the data that lead to false conclusions because the data are not of sufficient quality to support the relevant analyses.

The panel believes that caution needs to be exercised by both Stats NZ and users when considering release of data that is rated as being of **poor** or **very poor** quality (at any level of the relevant classification), even where at higher levels of a classification the results meet tests which meet quality ratings of **very high, high**, or **moderate**.

In Section 8 – Poor and very poor quality data the panel reviews some of the key findings relating to variables that have been rated in these two quality categories. The panel believes that there are a number of options open to Stats NZ when releasing data rated as of **poor** or **very poor** quality at any level for which it is available. Their relevance will depend on the uniqueness of the particular census question, the significance of the consequent decisions that will be based on the analysis, and the expertise of those using the information.

The panel consider that Stats NZ has these options:

- Release the disaggregated data as official statistics without additional information – this is not advised.
- Release the disaggregated data with well signposted guidance, quality ratings and metadata (including data about the quality of the data) at the level of geography or classification users will use the data – this is recommended.
- Release the data through restricted access mechanisms only to users who have been briefed in depth on the quality considerations of their proposed analyses or research – this is recommended for variables rated as poor quality overall.
- Do not release the information as official statistics – this is already Stats NZ's approach for variables rated as very poor quality overall. This should not rule out informed investigation of the data (as with iwi data).

Stats NZ have already announced that iwi data (which is of **very poor** quality), will not be released as official statistics. They are, however, working with representatives of the Māori community to allow access to the data for analyses to see what use, if any, might be made of the data collected. The panel believes that this is a sensible approach to determine whether any value can be obtained from this data.

Stats NZ has released data rated as poor, but the panel considers that such data has the potential to mislead. The panel encourages Stats NZ to control access to data rated as poor or very poor overall to accredited individuals working in controlled databases who are able to work closely with Stats NZ to understand the quality characteristics of the data.

Data processing and metadata

In Section 9 – Outstanding items in the Terms of Reference, the panel briefly reviews the data processing and evaluation approaches that Stats NZ used during development of the census data file.

In broad terms, once initial processing of the data was completed, there was a variable-by-variable assessment of the quality of the resulting data. If data issues were found and there

was still time in the processing timetable, then there was the potential to make fixes (either automated or manual) and to rerun the data. This is a normal census process.

During the evaluation process Stats NZ analysed the data and checked the data quality. This included conducting:

- Time series checks – 2006, 2013, 2018
- Checks against expectations – e.g. population estimates and projections
- Checks at lower levels of the classification
- Checks at lower levels of geography (Priority 1 variables down to SA2; Priority 2 and Priority 3 to start at Regional Council level then lower if necessary)
- Consistency checks
- Checks of key cross-tabs, e.g. individual variables by age, sex, ethnicity

At the end of this process Stats NZ wrote internal assessments of quality for individual variables termed 'Warrants of Fitness' (WoFs). The WoFs include:

- A quality rating assigned to each variable on the basis of the three quality metrics – data sources and coverage, consistency and coherence, and data quality (see Appendix 3)
- A breakdown of the data sources, non- response rates, and a data quality rating calculation for Metric 1
- An outline of the edits, including data edits
- Sections on analysis and problems, checks and outcomes
- Recommendations for using the data (including recommendations for the next census)

To complete its evaluation of variables (e.g. in the Assessment of Variables Report) the panel had access to 63 of these separate WoFs. The panel also had access to the Datainfo+ pages for the 2018 Census which provide standardised information on each variable, including data sources used, quality rating, and information on any changes to the question or classification.

The relative contributions of different data sources (2018 Census, 2013 Census, admin data, imputation and 'no information') vary significantly by variable, by level of classification for a particular variable (e.g. ethnicity) and by level of geography (e.g. regional compared to SA2). It is therefore important that users have readily available to them metadata on the relative contributions of data sources, and the associated quality rating at least for Metric 1 (data sources and coverage) at these low levels of disaggregation.

Looking ahead to the 2023 Census

A great deal has been learned about the assessment of census data quality during the 18 months the Expert Data Quality Panel has been interacting with Stats NZ on methods for producing and evaluating the data that began to be released to users and the general public from late September 2019. In Section 10 – Towards the 2023 Census, the panel offers some suggestions as Stats NZ prepares for the 35th enumeration of the country's population in 2023.

Delays in producing substantive outputs from the 2018 Census have generated a considerable amount of criticism of the value of the census. A key challenge facing Stats NZ is rebuilding public confidence in both the ability of the Department to deliver a modernised census that enumerates at least 94 percent of the population and is not compromised by significant under-enumeration of Māori and Pacific peoples.

Regaining the trust and support of the key users of census data and those people who can help support the census, including local government and community groups, will be critical for the 2023 Census. If Stats NZ wants to rebuild their trust and enlist them as advocates for the census operation, it will need to give such groups more of a voice in reviewing the operational arrangements. Stats NZ will need to draw on feedback on whether their proposed operational approach (and allocation of resources) will work locally.

Regaining the trust and support of Māori especially will be a challenging task for Stats NZ. There will need to be a demonstrable shift in perception amongst iwi and other Māori organisations that Stats NZ is committed to a meaningful partnership to deliver on its Tiriti o Waitangi obligations that are specified in *Stats NZ's Strategic Intentions 2019-2023.*[5] While comment on operational aspects of the 2018 Census is not part of the panel's brief, the need to improve coverage of the 2023 Census in order to deliver data of acceptable quality on iwi affiliation and use of te reo Māori is essential. The census is the only source of information at a national level on these, amongst many other, critically important dimensions of New Zealand society.

Building trust and confidence in the 2023 Census will require significant additional investment and meaningful action in the lead up to and the execution of the 2023 Census. This investment in time, resources and commitment will be essential to provide the capacity both to plan and deliver on a census that is developed in partnership with Māori. Delivering effectively for Māori in 2023 should also ensure strategies are implemented that will also result in a much better response from Pacific peoples to the census, the other main group that has been compromised by very low response rates in 2018.

[5] In *Stats NZ's Strategic Intentions 2019-2023* it is recognised that as the leader of the government data system, the Department needs to ensure that decision-makers have the data and information they need to make decisions, and that there is "improved representation of Treaty partners in our data and products". https://www.stats.govt.nz/corporate/stats-nzs-strategic-intentions-201923

To secure a very high response from Pacific residents will require culturally appropriate initiatives that, in themselves will have prioritisation as well as investment implications. In the panel's view, the 2023 Census cannot be a cost-cutting census – if it is to achieve the sorts of response rates indicated above, there will need to be significantly more investment in the 2023 Census than there was in the 2018 Census.

In addition to these investment-related observations, Section 10 includes a series of recommendations relating to census instruments, data sources and quality-related initiatives that Stats NZ should consider as the Department plans for the 2023 Census.

Recommendations

This section sets out the panel's recommendations arising from their assessments of the methods used by Stats NZ to improve coverage of the 2018 Census, and the wide-ranging implications for data quality of the much lower response rates for particular groups and areas than had occurred in previous censuses.

The recommendations are informed by information contained in the panel's three reports that are listed in the Executive summary. The recommendations are positioned in the section of the report to which they relate. Some are repeated in more than one part of the report where they have specific relevance. A small number arise from material contained in the Initial Report and the Assessment of Variables Report. These are listed in Sections 1 and 2. The rest relate to material presented in this report.

All recommendations are mentioned in this report and they are numbered sequentially from 1 (in Section 1) to 24 (in Section 10). The page numbers that are cited at the end of the recommendations indicate where they can be found in this report.

In the summary below, the recommendations have been grouped under four headings in order to assist readers to interpret the panel's advice to Stats NZ that arises from its comprehensive examination of issues relating to quality of the 2018 Census data. The headings are:

1) Preparing for the wider release of the 2018 Census results
2) Ensuring the adequacy of information for users about quality of the 2018 Census data
3) Ensuring greater comparability of data collected in the 2013 and 2018 Censuses
4) Securing the quality of data collected in the 2023 Census

Preparing for the wider release of 2018 Census results

In recent months Stats NZ has provided users of 2018 Census information with a timetable for the release of data, and developed a strategy for the release of information about the quality of 2018 Census data including the panel's three reports. The panel considers some additional actions are needed and these are covered in the following recommendations:

R 15. Stats NZ should ensure that users of 2018 Census data have readily available to them:

15a Metadata on the relative contributions of different data sources for every variable at all levels in their coding classifications for levels of data aggregation down to SA2, and for level 4 ethnic groups.

15b The associated quality rating (at least for Metric 1, data sources and coverage) should be provided at all levels of spatial aggregation (SA2 to region) and for all levels in coding classifications that are frequently employed by users (e.g. L4 for language, L3 religion, L4 for ethnicity, etc.). (Section 4, pg. 59)

R 16. Stats NZ should work with key users such as Te Taura Whiri i te reo Māori and Te Mātāwai to clearly communicate the problems with te reo Māori data from Census 2018, the limitations around its usage, and options to futureproof te reo data moving forward. (Section 5, pg. 64)

R 18. Stats NZ should report on 2018 Census SA2 unit source indicator and item source indicator by variable and calculate the metric 1 data quality to support users with an interest in small area analyses, because of the number of small areas experiencing very low coverage and consequently lower data quality across variables. (Section 7, pg. 81)

R 19. Stats NZ should only make data rated as being of **poor** or **very poor** quality overall available where project proposals are considered by Stats NZ on a case-by-case basis, similar to the current procedures for accessing Stats NZ microdata such as the Integrated Data Infrastructure (IDI). This includes data relating to families and households unless Stats NZ determine and advertise a higher quality rating for these data (Section 8, pg. 87)

Ensuring the adequacy of information for users about the quality of the 2018 Census data

There are a number of actions that the panel considers are necessary to complement the initiatives already taken by Stats NZ to inform users about quality issues with the 2018 Census. The additional actions are:

R 4. Stats NZ should further investigate whether its key methodological assumptions apply at smaller geographic scales. (Section 1, pg. 25)

R 5. Stats NZ should consider the costs and benefits of creating a multiply-imputed data set for use by experts working with microdata so that researchers using these data can take account of situations where the relatively high level of imputation will tend to distort relationships between variables and cause uncertainty in estimates to be underestimated. (Section 1, pg. 26)

R 8. Stats NZ should continue to investigate and report on issues linked with quality of the 2018 Census data, including up-dating their assessments of variables in the light of analyses of data quality at lower levels in coding classifications. (Section 3, pg. 35)

R 9. Stats NZ should consult with key users of census data to ensure the information they require on data quality is readily available on the Stats NZ website. Information relating to data quality in the 2018 Census should remain visible on the website until at least the results of the 2023 Census are published (Section 3, pg. 35)

R 14. Stats NZ should collate queries from users about data quality issues and the advice Stats NZ provide, and make this information available to all users, to give users comprehensive information about the quality of 2018 Census data. (Section 3, pg. 44)

Ensuring greater comparability of data collected in the 2013 and 2018 Censuses

The 2018 Census data pose several challenges for the measurement of population intercensal change, especially for small areas and at low levels in variable coding classifications. The following actions would ensure greater comparability with 2013 Census data:

R 7. Stats NZ should undertake further analysis of the impact for comparisons with 2013 Census data of the changes in the wording or response options of questions in the individual and dwelling forms. Examples include: Tenure of Household, Main means of travel to work, Main types of heating and fuel types used to heat dwellings. A comparison between responses to 'Main means of travel to work' and the Ministry of Transport Household Travel Survey data should also be undertaken. (Section 2, pg. 31)

R 10. Stats NZ should investigate the possibility of recalculating 2013 Census results using (as far as possible) 2018 Census methods. The retrospective use of government administrative records from around the time of the 2013 Census could contribute responses where non-response has been high, and enable more consistent measurement of changes that have occurred in comparable census data between 2013 and 2018 than is currently possible. (Section 3, pg. 36)

R 12. Stats NZ should systematically investigate the impact of the use of alternative data sources (previous census data, data from a range of admin sources, imputed data) on the quality of data across variables. Analyses should focus not just on whether population distributions are in line with expectations, but also impacts on estimates of inter-censal change, the impact on the sizes of ethnic groups and small areas (e.g. SA2s), and the impact on bivariate associations between variables. (Section 3, pg. 38; Section 7, pg. 81)

Securing the quality of the 2023 Census

Once completed, it is normal to analyse the quality of each census as if it were a pilot test for the census which will follow it. The panel has the following advice for the planning of the 2023 Census:

R 1. **Stats NZ should** ensure data collection in future censuses is comprehensive enough to accurately measure iwi affiliation, and should take responsibility, in partnership with iwi, for

investigating alternative ways to measure iwi affiliation so that the census is not the only source. (Section 1, pg. 25; Section 10, pg. 94)

R 2. Stats NZ should prioritise engagement and investment to ensure:

2a There is genuine partnership with Māori communities, organisations and iwi to develop and implement decision-making and governance mechanisms, to ensure meaningful involvement of Māori in future censuses. This includes Stats NZ actively addressing the acceptability of the extensive use of administrative data in future censuses and issues of social license and Māori data sovereignty specifically for the 2023 Census.

2b There is a real voice for members of all communities, especially Pacific peoples and new migrants, in decision-making on data about them, including the use of admin data in the census. (Section 1, pg. 25; Section 3, pg.38; Section 10, pg. 95)

R 3. **Stats NZ should** ensure individual census responses from prisoners are obtained in the 2023 Census. (Section 1, pg. 25; Section 10, p 96)

R 6. **Stats NZ should** ensure that all collection instruments (paper and online forms), systems and processes are thoroughly reviewed, tested, and made fit-for-purpose for 2023, including an assessment of the equity implications of all collection instruments (paper and online forms), systems and processes. It is essential that issues with the 2018 collection instruments are addressed for the 2023 Census. In this context, Stats NZ should review the extent to which the way the online forms were administered contributed to missing responses in 2018, with a focus on the differential impacts for different population groups and consider whether changes are needed for the 2023 Census. (Section 1, pg. 26; Section 10, pg. 95)

R 11. Stats NZ should report on data quality at the small area (SA2) level to support analysts and policy makers with an interest in small area analyses and build quality rating calculations by Level 1 ethnicity for every variable relating to individuals into their quality assurance and evaluation plans for the 2023 Census. (Section 3, pg. 37; Section 7, pg. 81; Section 10, pg. 95)

R 13. Stats NZ should ensure that the methodology to be adopted for the 2023 Census makes explicit provision for high quality measures of intercensal change between 2018 and 2023. The following actions are recommended to avoid further breaks in time-series for census data:

- A high-quality Address Register should form the basis of a management information system for the field and online enumeration to support a quality management strategy that would allow early intervention when things go wrong or not as planned.
- Because the method of field collection of responses will remain a critical part of the next census the enumeration model for 2023 must be developed by building on that

which worked in 2013 and earlier rather than that which failed in 2018 and led to unprecedented quality problems.

- Continue to impute for missing item non-response.
- Assess any new changes to methodology very carefully against whether they would lead to further disruption to the census time-series.
- Review the content of the census forms to ensure the information needed to assess the quality and integrity of data is present. An example is reinstating the question on the total count of people in a household/dwelling.
- With appropriate public consultation and attention to privacy and data justice impacts, make use of access to the relevant government administrative records in advance of the next census to maximise response rates by, amongst other things, targeting field operations, the distribution of paper forms, the number and location of field staff etc.
- Should reliance on administrative records grow, reassess how far statistical surveys (Household Labour Force Survey, New Zealand General Social Survey, and Household Economic Survey), may be better placed to obtain some of the information traditional gathered by the five-yearly census, such as household and family statistics.
- Provide where possible for measures of quality of the enumeration to be an integral part of the census collection and estimation stages as they proceed.
- Consider the role of the PES in 2023, given the changes in census methodology introduced for the 2018 Census.
- Consider producing household and families data by ethnicity. (Section 3, pg. 40; Section 10, pg. 95)

R 17. Stats NZ should support a dedicated team for the 2023 Census to undertake post-processing for families and households data, and other complex variables, and not divert this team to other tasks. (Section 6, pg. 74; Section 10, pg. 96)

R 20. Stats NZ should continue to undertake WoF assessments for variables in the 2023 Census and should implement a systematic template for WoF content which should include quality ratings for Regional Councils and level 1 ethnicity. Quality control processes should be implemented to ensure assessments are of a consistently high standard. (Section 9, pg. 90; Section 10, pg. 96)

R 21. Stats NZ should have an organisational commitment to, and focus on, achieving effective partnership with Māori to develop a census delivery model that will achieve a very high response (>94 percent) from Māori in the 2023 Census. (Section 10, pg. 94)

R 22. Stats NZ should set response rate targets for particular Territorial Authority and Auckland Local Board areas and ethnic groups that had low response rates in 2018. These targets will drive a focus and resources into areas/groups needed to achieve a much more

balanced and complete response profile in 2023 than was achieved in 2018. (Section 10, pg. 94)

R 23. Stats NZ should review the priority ratings that it gives to variables for the 2023 Census so that key statutory duties, including in respect of Māori (e.g. te reo Māori), and variables used to construct units of analyses (such as households/families), are reflected in higher priority for the associated variables. (Section 10, pg. 96)

R 24. Stats NZ should ensure that external scrutiny, in advance of the 2023 Census, is focussed on methodology, field planning and systems for interaction with the public as well as quality management and quality measures. Ex post reviews, such as that of the External Data Quality Panel, can contribute to improvements in future censuses, but for the 2023 Census it would be more cost effective if some of that external expertise were applied well in advance of the enumeration. (Section 10, pg. 96)

Table of Contents

1 Key messages and recommendations from the initial report

This section provides a summary of the key messages from the initial report which is available here. A copy of the Executive Summary in the Initial Report can be found in Appendix 1. Six recommendations arising from this report follow the key messages.

The **key messages** from this report were:

- One in six New Zealand residents did not complete a questionnaire for the 2018 Census.
- In response to this unexpectedly high level of non-response, Stats NZ initiated a large-scale census mitigation project that involved the extensive use of alternative government data to fill the gaps.
- While census mitigation has enabled Stats NZ to produce a range of statistical outputs from Census 2018, there are also long-standing key statistics that remain unavailable.
- The panel endorses the statistical approaches used to mitigate non-response.
- The use of administrative and 2013 Census data has improved the quality of results that we would otherwise have had from the 2018 Census.
- Because the core demographic elements were measured differently in 2018 than in 2013, measures of change have been distorted by both methodology and response rate variations.
- The addition of administrative records reduced the 2018 Census undercount compared to previous censuses for the population as a whole, and for Māori and Pacific ethnic groups in particular.
- The unprecedented use of administrative data to augment census data raises questions around ethics, social licence (i.e. tacit approval from the New Zealand public), cultural licence (collective mandate for the trusted use of Māori data), and Māori data sovereignty.
- The panel assesses that the linking of government records has improved the coverage and, in some cases, the accuracy of counts of selected core demographic elements of a census: age, sex, place of usual residence and ethnicity. Nearly all of the population in the 2018 Census have been given responses on these core variables.
- While use of data from an earlier census and from a range of administrative sources, along with imputation methods to fill remaining gaps, has improved coverage this does not necessarily mean improvements in accuracy of all of the responses. There are a range of quality-related issues when data are disaggregated by ethnicity and analysed at the SA2 level.

- The panel has taken a broader view of the needs of users of ethnicity data than simply the ethnicity variable itself. We rate the quality of ethnicity data as moderate, rather than Stats NZ's rating of ethnicity as high quality.
- There is significant variability in the quality of ethnicity data by ethnic group. This reflects different patterns of non-response, and the reliance on different alternative data sources, which have different quality characteristics. The quality of ethnicity data becomes increasingly variable as the level of ethnic and spatial specificity increases.
- Having reviewed the methodology and examined the sensitivity tests initiated by Stats NZ, the panel is confident that the measure of the Māori descent population for electoral purposes meets the accuracy requirements.
- The use of administrative records, 2013 Census data, and the estimation methodology has resulted in a larger increase in the Māori descent electoral population from 2013 than occurred between 2006 and 2013.
- Users of Census 2018 data/statistics will need to explicitly consider the fitness for purpose of the information they wish to use, by consulting the rich array of documentation and quality measures.

1.1 Recommendations arising from the panel's initial report

R 1. **Stats NZ should** ensure data collection in future censuses is comprehensive enough to accurately measure iwi affiliation, and should take responsibility, in partnership with iwi, for investigating alternative ways to measure iwi affiliation so that the census is not the only source.

R 2. Stats NZ should prioritise engagement and investment to ensure:

2a There is genuine partnership with Māori communities, organisations and iwi to develop and implement decision-making and governance mechanisms, to ensure meaningful involvement of Māori in future censuses. This includes Stats NZ actively addressing the acceptability of the extensive use of administrative data in future censuses and issues of social license and Māori data sovereignty specifically for the 2023 Census.

2b There is a real voice for members of all communities, especially Pacific peoples and new migrants, in decision-making on data about them, including the use of admin data in the census.

R 3. **Stats NZ should** ensure individual census responses from prisoners are obtained in the 2023 Census.

R 4. Stats NZ should further investigate whether its key methodological assumptions apply at smaller geographic scales.

R 5. Stats NZ should consider the costs and benefits of creating a multiply-imputed data set for use by experts working with microdata, so that researchers using these data can take account of situations where the relatively high level of imputation will tend to distort relationships between variables and cause uncertainty in estimates to be underestimated.

R 6. **Stats NZ should** ensure that all collection instruments (paper and online forms), systems and processes are thoroughly reviewed, tested, and made fit-for-purpose for 2023, including an assessment of the equity implications of all collection instruments (paper and online forms), systems and processes. It is essential that issues with the 2018 collection instruments are addressed for the 2023 Census. In this context, Stats NZ should review the extent to which the way the online forms were administered contributed to missing responses in 2018, with a focus on the differential impacts for different population groups and consider whether changes are needed for the 2023 Census.

2 Summary of quality ratings for variables

The tables below summarise the assessments by the panel of the quality of the responses to each question answered in the 2018 Census. The first table covers variables about people (derived from the individual form); the second table covers those variables collected at the dwelling level (derived from the dwelling form). Variables which are included in this report are shaded grey. Detailed assessments are contained in the Assessment of Variables Report.

The panel has not been able to assess all variables that Stats NZ will be releasing. Attention has been focussed on those whose quality was rated by Stats NZ as **very high**, **high**, or **moderate**. It should be noted that the Stats NZ ratings listed in Tables 2.1 and 2.2 below are the ratings as determined at July 2019. Some of these are currently being reviewed by Stats NZ and could change.

In addition, the usual residence five years ago variable was assessed as, uniquely, this was always designed to be generated from comparing 2013 and 2018 Census data rather than, as previously, asking a specific question in the Census. One variable rated as **very poor** (Absentees) has been assessed, partly as a matter of completeness and to be able to explain the drivers for this rating.

Table 2.1: Summary assessments – variables about people

Variable name	Priority level	Where covered	EDQP Quality rating	Stats NZ Quality rating	Page no.
Absentees	1	Assessment of Variables Report	Very poor	Very poor	
Activity limitations	3	Not assessed	N/A	Poor	
Age	1	Initial Report	Very high	Very high	
Census night population count	1	Initial Report	Moderate	Moderate	
Census usually resident population count	1	Initial Report	Very high	Very high	
Cigarette smoking behaviour	3	Assessment of Variables Report	Moderate /Poor	Moderate	
Birthplace	2	Assessment of Variables Report	High	High	

Variable name	Priority level	Where covered	EDQP Quality rating	Stats NZ Quality rating	Page no.
Educational institution address	2	Not assessed	N/A	Moderate	
Ethnicity	1	Initial report and Section 4 this report	Moderate	High	45
Families and households: extended family type	2	Section 6 in this report	Very poor	Very poor	65
Families and households: family type	2	Section 6 in this report	Very poor	Very poor	65
Families and households: household composition	2	Section 6 in this report	Very poor	Very poor	65
Hours worked in employment per week	2	Assessment of Variables Report	Moderate /Poor	Moderate	
Individual home ownership	3	Not assessed	N/A	Poor	
Industry	3	Assessment of Variables Report	High/ Moderate	High	
Iwi	2	Initial report	Very poor	Very poor	
Languages spoken	3	Assessment of Variables Report and Section 5 in this report	Very high to poor, depending on the language	High	60
Main means of travel to education	2	Assessment of Variables Report	Moderate	Moderate	
Main means of travel to work	2	Assessment of Variables Report	Poor	Moderate	
Māori descent – output	1	Initial report	High	High	
Māori descent – electoral	1	Initial report	High	High	
Number of children born	3	Not assessed	N/A	Moderate	
Occupation	3	Assessment of Variables Report	Poor	Moderate	

Variable name	Priority level	Where covered	EDQP Quality rating	Stats NZ Quality rating	Page no.
Qualifications: highest qualification	2	Assessment of Variables Report	Moderate /Poor	Moderate	
Qualifications: highest secondary school qualification	2	Assessment of Variables Report	Moderate /Poor	Moderate	
Qualifications: post-school qualification level of attainment	2	Assessment of Variables Report	Moderate /Poor	Moderate	
Qualifications: post-school qualification field of study	2	Not assessed	N/A	Poor	
Relationship status: Legally registered relationship status, and partnership status in current relationship	2	Not assessed	N/A	Poor	
Religious affiliation	3	Assessment of Variables Report	High	High	
Sector of ownership	3	Not assessed	N/A	Moderate	
Sex	1	Initial report	Very high	Very high	
Sources of personal income	2	Assessment of Variables Report	High	High	
Status in employment	2	Assessment of Variables Report	Moderate	Moderate	
Study participation	2	Assessment of Variables Report	Moderate /Poor	High	
Total personal income	2	Assessment of Variables Report	High	High	
Unpaid activities	3	Not assessed	N/A	Poor	
Usual residence address	1	Initial report	High	High	
Usual residence one year ago	2	Not assessed	N/A	Poor	

Variable name	Priority level	Where covered	EDQP Quality rating	Stats NZ Quality rating	Page no.
Usual residence five years ago	2	Assessments of Variables Report	Poor	Poor	
Work and labour force status	2	Assessment of Variables Report	Moderate	Moderate	
Workplace address	2	Not assessed	N/A	Moderate	
Years at usual residence	3	Not assessed	N/A	Poor	
Years since arrival in New Zealand	3	Assessment of Variables Report	Moderate	Moderate	

Table 2.2: Summary assessments – variables for dwellings

Variable name	Priority level	Where covered	EDQP Quality rating	Stats NZ Quality rating	Page no.
Access to telecommunication systems	3	Assessment of Variables Report	Moderate	Moderate	
Census night address	1	Initial report	Moderate	Moderate	
Counts of dwellings	1	Assessment of Variables Report	High	High	
Dwelling occupancy status	N/A	N/A	N/A	N/A	
Dwelling type	2	Assessment of Variables Report	Poor	Moderate	
Housing quality: access to basic amenities	3	Assessment of Variables Report	Moderate	Moderate	
Housing quality: dwelling dampness indicator	3	Assessment of Variables Report	Moderate	Moderate	
Housing quality: dwelling mould indicator	3	Assessment of Variables Report	Moderate	Moderate	

Variable name	Priority level	Where covered	EDQP Quality rating	Stats NZ Quality rating	Page no.
Main types of heating and fuel types used to heat dwellings	3	Assessment of Variables Report	Moderate	Moderate	
Number of bedrooms	3	Assessment of Variables Report	High	High	
Number of rooms	3	Assessment of Variables Report	Poor	Poor	
Number of motor vehicles	3	Assessment of Variables Report	Moderate	Moderate	
Sector of landlord	2	Not assessed	N/A	High	
Tenure of household	2	Assessment of Variables Report	Moderate	Moderate	
Weekly rent paid by household	2	Not assessed	N/A	Moderate	

2.1 Recommendation arising from the Assessment of Variables Report

R 7. Stats NZ should undertake further analysis of the impact for comparisons with 2013 Census data of the changes in the wording or response options of questions in the individual and dwelling forms. Examples include: Tenure of Household, Main means of travel to work, Main types of heating and fuel types used to heat dwellings. A comparison between responses to 'Main means of travel to work' and the Ministry of Transport Household Travel Survey data should also be undertaken.

3 Data quality considerations and some guidelines for users of the 2018 Census

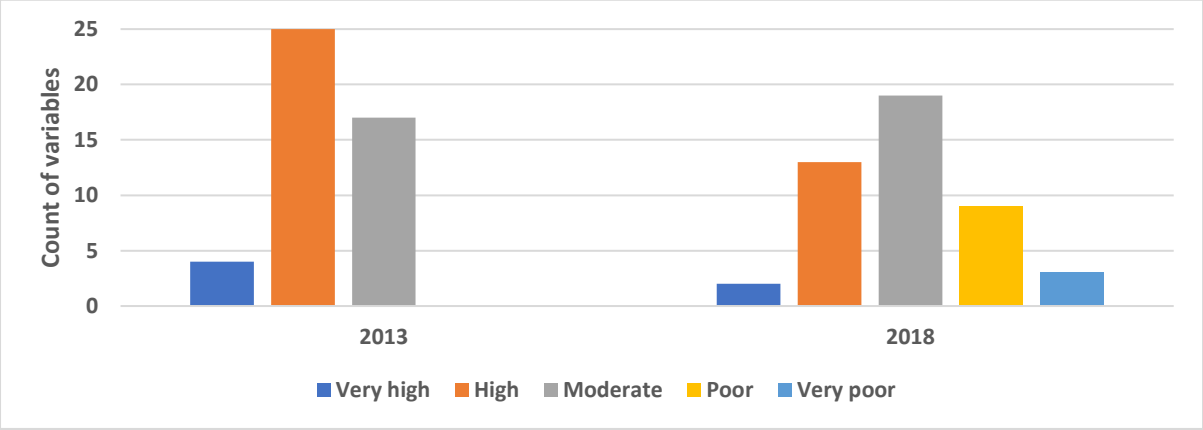
This section describes various issues with the quality of the 2018 Census about which users should be aware. It covers issues such as quality of levels versus change, improved coverage of key variables in 2018 compared to previous censuses, the impact on small areas and ethnicity, and the impact on key analytic units such as households and families.

3.1 Fitness for use of census characteristics

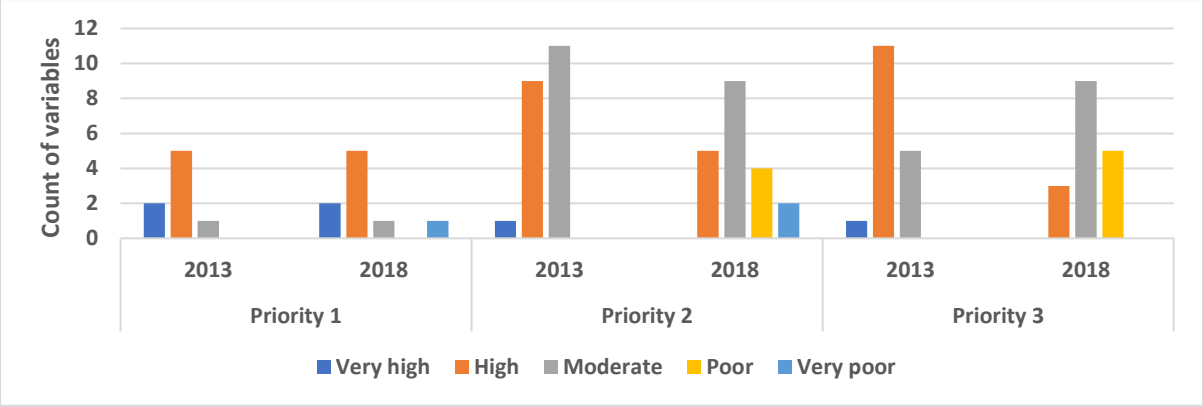
Few of the statistics produced from the 2018 Census have been prepared in the same manner as they were prepared in earlier censuses. Those that were produced in the same way are generally of lower quality compared with earlier censuses. Figure 3.1 below shows that 2018 Census variables are, on average, lower quality than for 2013.[6]

Figure 3.1: Stats NZ's quality ratings for variables – 2013 and 2018 Census, as at July 2019

A. All variables combined



B. Variables grouped by priority level



[6] Based on data from https://www.stats.govt.nz/reports/overview-of-data-quality-ratings-interim-coverage-and-response-rates-and-data-sources-for-2018-census

For example, panel A of Figure 3.1 shows that all variables from the 2013 Census were rated moderate quality or higher, whereas 12 of 46 variables from the 2018 Census were rated poor or very poor. Panel B of Figure 3.1 shows the quality rating breakdown by priority level, where Priority 1 includes variables such as age, sex, census night population count and iwi; Priority 2 includes variables such as birthplace, highest qualifications, and total personal income; and Priority 3 includes variables such as cigarette smoking, occupation, and religious affiliation. Notably, 11 of 12 variables from the 2018 Census rated as poor or very poor were Priority 2 and Priority 3 variables – the only Priority 1 variable rated as poor or very poor was absentees.

Understanding the quality of well-used measures has become very important when using statistics from the 2018 Census. There is now a much wider range of information about data quality than for previous censuses and users will need to become familiar with new approaches to response rates, the scale and sources of substituted or imputed information, as well as the expert assessments of quality made by both Stats NZ and the External Data Quality Panel.

In order to offset some of the loss in quality due to the low overall completion rate for census questionnaires compared to earlier censuses, Stats NZ has had to put in place a new methodology for producing the 2018 Census dataset. We noted in Section 3 of our initial report that we found good reason to have confidence in the methods put in place to ameliorate the situation, given the inadequacy of the census enumeration itself.

The fitness for use of information relating to characteristics that can be obtained only from census questionnaires cannot be presumed as in the past – it needs to be evaluated for specific applications. Supporting this need, Stats NZ has made available a wide range of quality measures about each characteristic such as in the DataInfo+ pages for each variable – see here . Stats NZ (2019a) has also published a paper entitled 'Data sources, editing and imputation in the 2018 Census' which is available on their website here.

The panel have provided their information on quality of the 2018 Census data in this report, their initial report, and in in their Assessment of Variables Report

3.1.1 Assessing quality

In its assessment of the quality of data collected in the 2018 Census, the panel had several sources to draw on.

- Stats NZ prepared a 'warrant of fitness' (WoF) for each variable, based on the completeness with which a census characteristic was able to be measured from either the census itself or another source, a comparison of the distribution of the variable categories or values with the 2013 Census or another relevant source, and the credibility of the change at a high level between the 2013 and 2018 Censuses. These detailed assessments were made available to the panel.

- For some specific major uses of the 2018 Census, in particular for the input to the Representation Commission and the updating of the NZ Deprivation Index, the panel engaged quite extensively with Stats NZ departmental experts.
- The Report of the Independent Review of New Zealand's 2018 Census was released in August 2019. The comprehensiveness with which it identified the sources of failure gave the panel confidence that there was a low risk that there would be any significant quality issues that it would not uncover, although there will be some specialised users of the censuses who may yet uncover some.
- The panel members were selected because they all had longstanding expertise and experience in the production, development and application of population statistics and the population census, and the impact of quality on fitness for use in a wide range of applications. The panel's collective knowledge and experience proved to be invaluable in assessing the wide range of issues it had to take into account, and the need for considerable analysis on our part. Several members of the panel completed analyses of the 2018 Census records which have extended the scope of our findings. Apart from peer review of our report, there were no formal arrangements with other experts, although all panel members would have engaged with colleagues on aspects of their work.
- The panel has drawn on an excellent framework for assessing the quality of statistical practices and measures published by Statistics Canada (2017). Non-response, the tension between consistency and relevance, conceptual limitations and methodological and operational errors are all known to impede any statistical source in some way. The panel chose to give considerable weight to several specific attributes that would be important to test in any census. At times these attributes were given a different weight by the panel compared with the Stats NZ assessments. Those the panel has focussed on are:
    i) **Granularity** – ensuring that quality is assessed at the level of detail in the classification that is most relevant to users.
    ii) **Inter-censal change** – has the methodology resulted in less precision or more bias in the measurement of change from the previous censuses – even if the 2018 Census results were of higher quality or coverage?
    iii) **Required inter-relationships between census variables** – are there particular variables that are not directly measured but are derived from relating two or more variables obtained in either questionnaires of the 2018 Census? This includes family measures, dwelling quality and crowding.
    iv) **Specific uses** – will those known uses that require measures to be derived from the 2018 Census be based on fit for use information? For example, electoral populations as specified in the Electoral Act, Deprivation Indexes published by the University of Otago.

v) **Uniqueness of the census as a source for some statistics** – is there any other source of the information? For example, iwi counts, characteristics of diverse ethnic groups, occupation, dwelling quality.

Users of the 2018 Census data will need to take account of assessments and measures of quality, as the quality that has usually been previously believed to have been provided cannot be assumed for 2018. This is particularly important for any analyses of the 2018 Census that is intended to influence public and community welfare in any way, including use for allocating resources between areas or specific sub-groups of the population. It should be a normal part of analytical work for users to apply the quality information available from Stats NZ to validate their fitness for use, and hence assess the degree of confidence in the analyses and its impact on decision-making.

R 8. Stats NZ should continue to investigate and report on issues linked with quality of the 2018 Census data, including up-dating their assessments of variables in the light of analyses of data quality at lower levels in coding classifications.

The panel is of the view that all reports on 2018 Census data quality – from both the panel itself and from Stats NZ – need to be readily accessible to users of 2018 Census data, and to remain visible until at least the publication of the results from the 2023 Census. On the recommendation of the panel Stats NZ has created a page on its website dedicated to information relating to data quality for 2018 Census.

R 9. Stats NZ should consult with key users of census data to ensure the information they require on data quality is readily available on the Stats NZ website. Information relating to data quality in the 2018 Census should remain visible on the website until at least the results of the 2023 Census are published.

3.2 Key elements of quality that affect fitness for use of data

In this report the panel has explicitly considered several key elements of quality of each variable. Its conclusions on specific quality issues are summarised in this section.

3.2.1 Maintaining the quality of census time-series

The 2018 Census has adopted a number of fundamental changes to methodology which, whilst probably delivering more accurate counts of a number of variables, have come at the cost of breaks in the time series with previous censuses.

The 2013 Census count includes 4.9 percent of 'substitute' records, a form of unit imputation that improves the census counts, but does not provide information about characteristics. Only age and sex were imputed for these substitute records, and all other information is missing. In addition, the 2013 Post-enumeration survey estimated a net

census undercount of 2.4 percent, and an overall response rate of 92.9 percent – the proportion of the true population who submitted census individual forms.[7]

The responses to individual questions would also have contained item non-response. This happens where people have filled in the census form but have either not responded to some questions or have given answers which cannot be coded to a valid response. For instance, the non-response rate in 2013 for total personal income was 9.7 percent, of which 4.9 were substitute records.[8]

Given the low response to the 2018 Census (an interim national collection response rate of 87.5 percent), Stats NZ had to design new methods for the 2018 Census. The results of the changed methodology have fundamentally changed the relationship of the results of the 2018 Census to those of the past. The use of admin data to count non-responding people has resulted in an estimate of the 2018 population that is missing only 1.4 percent, compared to the 2.4 percent missing from the 2013 Census.

Through the use of admin records in 2018, rather than the unit imputation 'substitute' records of 2013, the information about those who did not respond to the field collection is more truly representative, and some variables are available from alternative sources. The 2018 Census is therefore more representative than the 2013 Census (98.6 percent vs 92.9 percent, with more complete information about core variables). Other things being equal, this change in itself would mean that we can expect that the rates of change between 2013 and 2018 will be overstated if directly comparing the two censuses. This overstatement will vary among population groups for any variable in the census.

The use of admin data has had the biggest impact on the counts of young adults (who are typically less well enumerated in censuses worldwide), Māori and Pacific peoples, and certain regions (Northland, Gisborne, and Bay of Plenty in particular). These groups are almost certainly better counted in the 2018 Census datasets than in previous censuses, but at the cost of breaks in the time series.

In addition to the breaks caused by addition of people who completed the 2013 Census as well as people identified in a range of administrative sources, there has been a fundamental change in the use of imputation for the 2018 Census. For more than half of the variables, imputation has been used to enable 100 percent (or close to 100 percent) coverage of the population. For instance, for industry 71.6 percent of the data came from 2018 Census, 20.8 percent from admin data, and 7.7 percent from imputation. This means that great care needs to be taken when comparing census counts between the 2013 and 2018 Censuses.

R 10. Stats NZ should investigate the possibility of recalculating 2013 Census results using (as far as possible) 2018 Census methods. The retrospective use of government administrative records from around the time of the 2013 Census could contribute responses

[7] Stats NZ (2014b). Post-enumeration survey: 2013. Available from www.stats.govt.nz.
[8] http://archive.stats.govt.nz/Census/2013-census/info-about-2013-census-data/information-by-variable/total-income-personal-family-combined-parental-extended-family-and-household.aspx

where non-response has been high and enable more consistent measurement of changes that have occurred in comparable census data between 2013 and 2018 than is currently possible.

3.2.2 Stronger representation of key variables - age, sex, place, and ethnicity

The key effect of the post-census integration with government administrative records has been to improve the overall coverage of the four key measures which define New Zealand's distinct communities. These are basic counts of individuals by age, sex, ethnicity, and place.

Despite the unprecedented problems from response at the time of enumeration for the 2018 Census, and after the remedial action by Stats NZ, the capability exists now for assessing these attributes for a near complete estimate of the population, compared to what was possible in 2013 and earlier censuses. This is due to the quality and completeness of address and person frames that are now available to compare against the enumerated population and identify substitute responses on a scale that has not before been practical.

There has been an unstated presumption that the error structure inherent but not explicit in a traditional census would not change dramatically from one census to the next. This is because it is mainly the same groups that have the higher non-response rates, while the core topics generally evolve rather than change radically at each census. The core methodological elements changed little in the censuses before 2018. This assumption will **not** hold for the 2018 Census, because of the use of admin data to create records for missing people, and the use of 2013 Census, admin data and imputation to replace item non-response for all or most variables.

R 11. Stats NZ should report on data quality at the small area (SA2) level to support analysts and policy makers with an interest in small area analyses and build quality rating calculations by Level 1 ethnicity for every variable relating to individuals into their quality assurance and evaluation plans for the 2023 Census.

3.2.3 Admin data gives some 2018 Census variables higher coverage than earlier censuses.

The administrative records of agencies such as Inland Revenue (total personal income, income sources, industry and sector of ownership), Education (participation in education, qualifications), Immigration New Zealand (birthplaces, year of arrival in NZ), Internal Affairs birth registrations (birthplace, Māori descent, ethnicity, number of children born), and Health (ethnicity) have resulted in a more complete coverage of the population for some variables than is possible through the usual form of census enumeration alone.

While the part played by the use of administrative records in the 2018 Census was invaluable, its extent was unprecedented and unheralded in advance of the census. The public response to the extensive sourcing of information from a wide range of administrative databases without prior consultation about the uses of such data for census

coverage purposes has yet to be seen. In this context, a recommendation arising from the panel's initial report that is listed in Section 1, merits repeating.

R 2. Stats NZ should prioritise engagement and investment to ensure:

2a There is genuine partnership with Māori communities, organisations and iwi to develop and implement decision-making and governance mechanisms, to ensure meaningful involvement of Māori in future censuses. This includes Stats NZ actively addressing the acceptability of the extensive use of administrative data in future censuses and issues of social license and Māori data sovereignty specifically for the 2023 Census.

2b There is a real voice for members of all communities, especially Pacific peoples and new migrants, in decision-making on data about them, including the use of admin data in the census.

For the measurement of 'Years since arrival in New Zealand', the government administrative records only exist since 1997, so that imputation for missing information using administrative records was more comprehensive for younger people. Country of birth is of a higher quality in 2018 than earlier censuses, although the analysis of changes in birthplace populations over time will be subject to the impact of higher non-response rates amongst some small birthplace and ethnic groups in earlier censuses. The continuing use of administrative records will require greater commonality in standard classifications used in government administrative records.

R 12. Stats NZ should systematically investigate the impact of the use of alternative data sources (previous census data, data from a range of admin sources, imputed data) on the quality of data across variables. Analyses should focus not just on whether population distributions are in line with expectations, but also impacts on estimates of inter-censal change, the impact on the sizes of ethnic groups and small areas (e.g. SA2s), and the impact on bivariate associations between variables.

3.2.4 Reduced precision in measuring location and ethnicity compared to earlier censuses.

Although the coverage of the population by place is more complete than in earlier censuses, around 8 percent of the population cannot be placed at a specific address, although it has been possible to establish the meshblock (small area containing around 30–60 households) where they would usually be resident. Ethnicity has been obtained from a mix of sources, and the collection methods and classification structure used have not been uniform – for more details see Section 5 (ethnicity) of the panel's initial report and Section 4 (ethnicity and birthplace) in this report.

Response rates to the 2018 Census varied widely in New Zealand, with the lowest rates in regions such as Northland, Bay of Plenty and Gisborne, which have high Māori and Pacific populations. There are national response rates of around 70 percent for these ethnic

groups. Section 7 of this report (on small area statistics) shows that the proportion of census individual forms in the census dataset at the SA2 level can be as low as 46.9 percent (Wiri West - Auckland). Low response in such areas affects not only the population count, but many variable response rates. 62.2 percent of census residents in Wiri West have no information for legally registered relationship status, and 60.3 percent of occupation data is imputed in this SA2.

The widely varying coverage of those characteristics that can be measured only in the census, and the under-coverage of specific communities, have affected the quality of ethnic group and place statistics. These difficulties will not be fully offset by the strengthened representativeness of the frequently used measures (population counts, age and sex) that define distinct communities, and the high coverage of some characteristics (e.g. personal income) in administrative records.

3.2.5 The use of imputation is at a much higher rate than earlier censuses.

For most of the variables relating to people, a complete, or near complete, data set was created using 2018 Census returns, 2013 Census data, a range of admin data sources, and imputation. In previous censuses only four variables – age, sex, usual residence and labour force status in 2013 (Stats NZ, 2014a) – had missing values imputed. The 2018 Census dataset is thus a much more complete dataset than for the 2013 Census – this method change alone will impact on measures of change. For instance, in the 2018 Census for religious affiliation, the use of 2013 Census data and imputation means that there is no missing data, whereas in 2013 8.2 percent of records were 'not elsewhere included'.

There are some key variables where the imputation rates exceed ten percent. These include: occupation (20.3 percent), main means of travel to work (19.0 percent), work and labour force status (18.9 percent), hours worked in employment per week (18.7 percent), status in employment (17.9), and main means of travel to education (15.5 percent). See here for details on data sources by variable.

3.2.6 Breaks in the time series between 2013 and 2018

Each Census of Population and Dwellings is part of a very long-term statistical series, measuring change across time and place for levels, rates and shares among a wide array of population groups. Field enumeration problems in 2018 and different methods for creating the 2018 Census data file add variability to measures of change between 2013 and 2018, both for measures based on census characteristics as well as alternatively sourced data.

In any Census of Population and Dwellings, there is always a tension between modifying concepts and questions to maintain their relevance and bringing inter-censal consistency by keeping questions the same. Given that the coverage of the 2013 Census of Population and Dwellings is known to be variable across major ethnic groups in the population (see the 2013 Post-enumeration survey and the panel's initial report), the differences in non-

response rates in 2018, and the methods used to adjust for these, are likely to overshadow management of an ongoing tension between consistency and relevance of questions.

The greatly increased quality in measurement of income through the linking to tax records provides opportunities for improved intra-group analysis of incomes. However, comparisons with data on incomes in earlier census data will be less accurate than in the past.

For some variables (e.g. tenure of household, main means of travel to work, and main types of heating and fuel types used to heat dwellings), changes to the 2018 Census question will also impact on measures of change. Variables particularly affected include study participation, dwelling tenure, activity limitations, main means of travel to work, highest qualification, and usual residence five years ago. For instance, for main means of travel to work the reference period for the question changed from census day, to the 'way you usually travel' – meaning that comparisons to 2013 Census should exclude those who worked from home on census day 2013.

Given that in 2023 Stats NZ will certainly not wish a repeat of enumeration problems on the scale experienced in 2018, the same concerns could occur when comparing 2023 Census statistics with those from 2018, unless changes in methods are explicitly managed in the census design.

R 13. Stats NZ should ensure that the methodology to be adopted for the 2023 Census makes explicit provision for high quality measures of intercensal change between 2018 and 2023. The following actions are recommended to avoid further breaks in time-series for census data:

- A high-quality Address Register should form the basis of a management information system for the field and online enumeration to support a quality management strategy that would allow early intervention when things go wrong or not as planned.
- Because the method of field collection of responses will remain a critical part of the next census the enumeration model for 2023 must be developed by building on that which worked in 2013 and earlier rather than that which failed in 2018 and led to unprecedented quality problems.
- Continue to impute for missing item non-response.
- Assess any new changes to methodology very carefully against whether they would lead to further disruption to the census time-series.
- Review the content of the census forms to ensure the information needed to assess the quality and integrity of data is present. An example is reinstating the question on the total count of people in a household/dwelling.
- With appropriate public consultation and attention to privacy and data justice impacts, make use of access to the relevant government administrative records in advance of

the next census to maximise response rates by, amongst other things, targeting field operations, the distribution of paper forms, the number and location of field staff etc.

- Should reliance on administrative records grow, reassess how far statistical surveys (Household Labour Force Survey, New Zealand General Social Survey, and Household Economic Survey), may be better placed to obtain some of the information traditional gathered by the five-yearly census, such as household and family statistics.
- Provide where possible for measures of quality of the enumeration to be an integral part of the census collection and estimation stages as they proceed.
- Consider the role of the PES in 2023, given the changes in census methodology introduced for the 2018 Census.
- Consider producing household and families data by ethnicity.

3.2.7 Reduction in responses which form statistical units for analysis with other census variables.

Some descriptive variables also form statistical units or subject populations for analysis with other census variables. In particular this includes households and families, occupations, and work and labour force. Where the share of responses at the highest level of granularity has fallen in the 2018 Census compared to earlier censuses, some studies such as monitoring the gender, age and ethnic mix of persons in specific occupations will be less reliable than in the past. That around 8 percent of the population cannot be placed in households and families has limited the value of households and families, occupations, work and labour force analyses.

Not surprisingly given the scale of non-response, the post-census integration with administrative records has not been able to resolve fundamental difficulties in the preparation of statistics on households and families that result from missing information in census questionnaires. This is consistent with experiences of other countries in the early years of their use of administrative registers. The matching of people to dwellings for measures such as crowding excludes just under 8 percent of the population.

Dwellings are the third statistical unit that census provides information for. Response to the dwelling form in the 2018 Census was better than for individuals, with a dwelling form received for 92.7 percent of all private occupied dwellings, compared with 96 percent in 2013. For dwelling variables with no other sources of information, the percent missing is typically about 2 percent higher in 2018 than in 2013, and time series data are generally more comparable.

3.2.8 Considerable variability in coverage of variables only reported in the 2018 Census

Across many distinct communities that can be defined by age, sex, age, ethnicity or place, information from the 2018 Census is available for half or less of the population (e.g. L1 Pacific ethnicities by occupation in South Auckland TALBs).

The graphs published separately on the Stats NZ website highlight the considerable extent to which coverage can vary among the subgroups produced in any two or three dimensional tabular analysis, see 2018 Census external data quality panel: Data sources for key 2018 Census variables. Although missing responses to work and labour force status have been imputed, for example, the rate of imputation is such that its significance and effect on quality will vary greatly across population groups (see also Section 7 – small area considerations).

3.2.9 Data for special interest groups is less comparable with previous censuses

In making comparisons with earlier censuses, the high non-response in the 2018 Census compared to the non-response in earlier censuses will affect variables for which there is no alternative source of comparable reliability as the census. This will affect established analyses that have direct policy implications including smoking, disability and employment equity programs.

Particular comparisons that will be of lower quality in analytical studies are those at the more granular levels of variable classifications. This includes information that can be provided from the census questions about different ethnic groups, particularly employment status, travel origin and destination analyses, iwi membership, age group studies, activity limitations, religious group membership, and smoker characteristics.

Where the responses to one question inform responses to another the information loss will be greater. An example is the way that language use, residence 5 years ago, religion and whether born overseas are important in analysing ethnic groups.

For some variables up to around 8 percent of the 2018 Census dataset is sourced from the 2013 Census responses. For ethnicity it was 8.2 percent, for religious affiliation 8.2 percent, and for smoking, 6.8 and 7.8 percent of responses for the two input variables (ever smoked and regular smoker) despite a downward trend in smoking rates between 2013 and 2018. For smoking, comparisons of the impact on different age and ethnic groups, as well as on men and women, are likely to be further obscured by the high level of imputation (8.1 percent for both input variables).

3.2.10 Methodology limitations affect several questions.

Some specific variables affected by changes in 2018 Census methodology are identified below.

Activity limitations

For each Census of Population and Dwellings between 1996 and 2013 some form of disability question was used to inform selection of a sample of census respondents for a special post census survey of disability. The sample was selected from both positive and negative responses because on this topic simple questions have been found to generate inaccurate responses from a significant share of the population.

In a review of the 2013 experience on surveying disability in both the census and an interviewer survey, Stats NZ (2015) noted that "These two enquiries provide false positives or type 1 errors of 28 percent, and false negatives or type 2 errors of 15 percent from the population census screening, compared to the post censual survey." Even without these quality-related methodological issues it is likely that the experience from previous censuses and surveys will have some significance in assessing the quality of the results from the activity limitations topics in the 2018 Census. The values of the same variable for different people do not relate to the same time period.

In the income question, the reference period may be lagged by a year as the information given in the census is usually measured for a tax year, rather than cumulatively up to some selected point.

The information on ethnicity has been obtained from several sources. The ethnicity data from education, for instance, may reflect when this information was first captured (e.g. in educational enrolments), rather than a current identification of ethnicity, which may refer to different points in time before the census reference date.

3.2.11 Summary

In any specific application of the 2018 Census data, a range of quality dimensions (coverage, coherence, consistency, timeliness of the data) needs to be considered to produce an overall quality assessment.

Quality measures produced by Stats NZ are available for most of these aspects. It is important to examine the non-response rates and imputation levels of the components of interest at the most detailed level of any classification, as significant variations at that level may be obscured by the way that differences can cancel out at aggregated levels of the classification (e.g. ethnicity and place – see Sections 4 and 7, respectively).

3.3 Guidelines for use of 2018 Census data

After over 12 months of reviewing quality-related issues with the 2018 Census data the panel have compiled some guidelines for users of these very important data.

When using the 2018 Census data the panel suggest:

- Read the individual assessments of variables written by the panel in either the panel's initial report, in the separate Assessment of Variables Report, or in this report. These provide the panel's overall assessments of the variables, provide key background information (including changes to the question or coding) and often contain caveats about the level at which the data can reliably be used.
- Read the relevant Stats NZ DataInfo+ page. These are listed in Appendix 5 and there are links in the Assessment of Variables Report and here. Also read Stats NZ's (2019a) report on data sources, editing and imputation in the 2018 Census here

- Consider whether the use relies on 2018 Census data alone (i.e. cross-sectional analyses) or uses changes in one or more variables between censuses. In the latter case be aware of the changes in methods in the 2018 Census and their impact on variables.
- Users wishing to understand specific communities and small areas should consult the data sources graphs on 2018 Census external data quality panel: Data sources for key 2018 Census variables to get an indication of data quality for the variables in which they are interested.
- Some variables contain high levels of imputed data (e.g. occupation 20.3 percent; main means of travel to work 19.0 percent; hours worked in employment 18.7 percent). CANCEIS imputations are designed to be unbiased but they do increase uncertainty. Data users who are using individual level micro-data (e.g. in the IDI) will face different problems and options compared to users who are using published tabulations.
- CANCEIS imputations will get many individual responses wrong, but these errors will be more-or-less offsetting so that the overall results are acceptable at a level of aggregation that is high enough for the errors to cancel out. Tabulations using variables with high levels of imputation will therefore be reasonably accurate for tabulations with large counts, but not for small cell counts (e.g. in small areas).
- Analyses at the individual level, such as statistical models fitted to individual-level data, may be very badly affected if these involve variables with high levels of CANCEIS imputation. Anyone planning to do individual-level analyses on data with substantial non-response needs to address these issues directly – e.g. by restricting the analyses to subpopulations with low levels of imputation, or by not using the CANCEIS values and using alternative missing-data techniques such as multiple imputation.
- If cross-tabulating variables, or carrying out complex analyses, consider the quality of each 2018 Census variable in isolation and then when taken together. For example, while data on age is of very high quality, variables analysed by age cross-tabulated by ethnicity may not be. The quality of such analyses will critically depend on which level of the ethnicity classification is used, and which specific ethnicities are being analysed.

R 14. Stats NZ should collate queries from users about data quality issues and the advice Stats NZ provide, and make this information available to all users, to give users comprehensive information about the quality of 2018 Census data.

4 Ethnicity and birthplace, further considerations

New Zealand's Census of Population and Dwellings is the only source of data on most of the ethnic groups that are identified in Stats NZ's (2005) *Ethnicity New Zealand Standard Classification V2.0*[9]. The classification has four levels and in the panel's initial report (Initial report of the 2018 Census External Data Quality Panel) there is a detailed discussion of the classification at levels 1 (L1) and 2 (L2) in section 5.

This section of the final report focuses on issues relating to the data at the lowest level of the ethnicity classification – level 4 (L4). It is at this level that members of the groups who self-identify with particular ethnicities can find information relating to themselves in the census. There is a lot of interest amongst members of these groups, as well as amongst analysts and policy makers in central and local government and researchers in the public and private sectors, in the census data on characteristics of specific ethnicities.

For some groups, applying the census measures of ethnicity to distinguish between ethnic communities has needed other characteristics also measured in the census. Prominent amongst these in New Zealand is birthplace, but increasingly language use and religion are also important distinguishing characteristics. Where the responses to one question informs responses to another in these situations, as noted in Section 3, the information loss will be greater because the rate of response when combining characteristics will be at best that of the characteristic with the lowest response rate.

The census variable that has been used most frequently in association with ethnicity data at all of the classification levels is birthplace. The birthplace variable, which is reviewed in the panel's Assessment of Variables Report, allows users to identify the New Zealand-born (NZ-born) and the overseas-born components of the specific ethnic populations. This enables users to examine characteristics of the two main components of all of New Zealand's ethnic groups – the ones who migrated to New Zealand at some stage in the past, and those who were born in the country. The census data are the only data available in New Zealand for this sort of analysis of migrant and non-migrant components of the population for most of the country's specified ethnicities that are identified at L4 in the classification.

The quality of the ethnicity and birthplace data has been reviewed by the panel in its initial report (ethnicity) and its Assessment of Variables Report (birthplace). In the case of ethnicity (at the L1 and L2 levels), the panel concluded that the ethnicity data at these higher levels of aggregation merited a **moderate** quality rating, rather than the **high** rating

[9] The Ethnicity New Zealand Standard Classification, which is currently under review, can be found at: http://aria.stats.govt.nz/aria/#ClassificationView:uri=http://stats.govt.nz/cms/ClassificationVersion/l36xYpbxsRh7IW1p

that Stats NZ gave the data (EDQP, 2019a). In the case of birthplace, the panel concurred with Stats NZ's **high** rating for the quality of data for this variable (EDQP, 2019b).

With regard to the panel's assessment of the quality of the L1 and L2 ethnicity data, it should be noted that the approach adopted by the panel to the ethnicity variable differs from that used by Stats NZ. Stats NZ's quality ratings for ethnicity at all levels measure how well the ethnicities are counted. The panel, in its initial report, measures the overall *quality* of the data about the ethnic categories, especially the overall quality of the data for those specifying Māori ethnicity.

In this report a different approach is taken again, this time focusing more on the consistency in sources of data for the diverse ethnicities that can be defined at level 4 of the ethnicity classification. The reason for focusing on consistency in data sources is that the Stats NZ measure of the quality of L4 ethnicity data, which is outlined in section 4.3, gives quite different weightings to responses that have been sourced from the 2018 Census, the 2013 Census and from other administrative data sources, or have been derived by imputation methods. The mix of sources for the specific ethnicities, as well as for the various "not further defined" or "not elsewhere classified" ethnicity categories can vary quite considerably. This variability needs to be taken into consideration when considering overall quality ratings for those L1 and L2 ethnic groups that include a large number of specific ethnic groups. Māori are the only specific ethnic group that applies at all four levels in the ethnicity classification.

In the first part of this section we summarise some of the differences between the ethnicity and birthplace data in the 2018 Census and the ethnicity and birthplace data in earlier censuses. This is an important quality-related issue because a frequent use of L4 ethnicity and birthplace data is to examine changes in population groups over time. This is followed by consideration of a recent unpublished assessment by Stats NZ of quality ratings for L4 ethnicities.[10] The final section contains a preliminary assessment of the sources of data relating to the NZ-born and overseas-born components of selected ethnic categories.

4.1 Differences between 2018 Census and earlier censuses

In this section the level 4 classifications of ethnicity and birthplace in the 2018 Census are compared briefly with the classifications used in the 2013 Census. Any differences in the respective classifications have implications for analysis of changes between 2013 and 2018 in the sizes of the populations for specific ethnic and birthplace groups.

4.1.1 Ethnicity

There is a smaller number of ethnic categories at level 4 in the 2018 Census (180) than there were in the previous two censuses (233). The rationalisation of specific categories in 2018 mainly covered very small numbers of people who identified with ethnicities linked with

[10] Stats NZ (2019b) 'Deriving a quality rating for level 4 ethnicity', unpublished report, November 2019

particular islands in countries like the Cook Islands and Solomons. Given that confidentiality requirements for data released by Stats NZ necessitate aggregating very small numbers for specific localities or groups before release of the data this grouping of island-specific information at L4 makes sense. The L4 ethnicity data for the 2013 and 2006 censuses, for example, can easily be regrouped by users to match the 2018 Census level 4 ethnicity categories.

In the 2018 Census every person in the usually resident population has been assigned an ethnic category either "specified" (153 categories) or in one of the following categories: "not further defined" (nfd) or "not elsewhere classified" (nec) (27 categories). There are no "don't know", "refused to answer", response unidentifiable", "response outside scope" or "not stated" categories in the 2018 Census usually resident population.

In the 2006 Census there were 167,784 people in the "don't know", "refused to answer", response unidentifiable", "response outside scope" or "not stated" categories – the equivalent of 4.2 percent of the usually resident population. In the 2013 Census there were 230,649 people in the 2018 Census not specified ethnic categories – 5.4 percent of the population. When comparing numbers in the 180 ethnic categories in 2018 Census with numbers for equivalent categories in 2013 and 2006 users need to keep in mind that the 2018 numbers include people who did not respond to the 2018 Census and whose data has been obtained from other sources. Any change in the size of specified ethnic categories in 2018 is thus a combination of real growth/decline in numbers as well as inclusion of an unknown number of people who did not specify an ethnicity.

The data for the specified ethnic categories in 2018 are not directly comparable with those from earlier censuses because of differences in the sources of the data on ethnicity. This has been discussed at some length in the panel's initial report, but it can be noted here that in the 2018 Census dataset 84.1 percent of the ethnicity responses came from the 2018 Census, 8.3 percent from the 2013 Census, 6.3 percent from administrative (admin) data sources and 1.2 percent were imputed responses. In total, just under 16 percent of the population had their ethnicities at L4 derived from a source other than 2018 Census. In previous censuses all the data on ethnicity came from the responses to the ethnic group question in the respective census.

In the 2018 Census there was considerable variation across specified ethnicities in the shares that came from the different sources. Some summary information illustrates this. Six ethnicities with small populations (under 1,000 each) had 100 percent of their responses from the 2018 Census. By contrast, only 67.7 percent of the 422,241 responses for people from the 18 Pacific ethnicities (excluding Indigenous Australians but including Pacific nfd and nec cases) and 70.9 percent of the 775,836 responses for people specifying Māori ethnicity were sourced from the 2018 Census. In terms of the other sources, percentages of responses ranged between 0 and 26.4 percent for the 2013 Census, 0 and 95.1 percent (Other ethnicity nec) for admin sources, and 0 and 4.2 percent for imputed data.

4.1.2 Birthplace

There were no substantive changes to the classification of birthplaces for the population in 2018 like the ones mentioned above to the ethnicity classification. In the 2018 Census there were 274 countries/regions specified as birthplaces, 241 specific countries/areas and 33 areas/regions with birthplace "nfd" or "nec". Included in the 241 named birthplaces were 7 with no people from them present in the 2018 Census (Vatican City State, St Pierre and Miquelon, Guinea Bissau, Anguilla, Sao Tome and Principe, Mayotte and Antarctica).

In the 2018 Census 57,858 people (1.2 percent of the usually resident population) were people with "birthplaces not elsewhere included" (i.e. not stated or not adequately described) compared with 259,434 (6.1 percent) and 188,187 (4.7 percent) people in this category in the 2013 and 2006 Censuses, respectively. As is the case for ethnicity data, the redistribution of people who did not specify a birthplace on a census form in 2018, across the various birthplace categories using data from other sources, means that direct comparisons of numbers for each birthplace in 2018 with numbers for the same birthplaces at earlier censuses is problematic.

In the 2018 Census dataset 83.8 percent of the birthplace data for the usually resident population came from this Census. The 2013 Census provided 8.6 percent of the responses and administrative sources 6.4 percent. There were no imputed birthplaces. As noted above, the remaining 1.2 percent of the population had missing data for this variable. We do not have data by source for all of the individual birthplaces, but for people included in the Pacific Peoples ethnic category (including indigenous Australians), for example, who were born in New Zealand, their shares from the three sources were 67.8 percent (2018 Census), 16.3 percent (2013 Census) and 15.9 percent (admin data). For the overseas-born component of the Pacific ethnic population, higher shares came from the 2018 Census (69.6 percent) and the 2013 Census (18.9 percent) and a lower share from admin sources (11.6 percent).

4.1.3 Summary

The 2018 Census ethnicity and birthplace data differ significantly from earlier census data for these variables because the large "not elsewhere included" categories have been removed entirely for ethnicity, and almost completely for birthplace. This has been achieved by using a mix of sources to populate the missing cases in the 2018 Census. An important impact of this on data quality is that it is not possible to compare directly the numbers of people in each specific ethnicity or with specific birthplaces across recent censuses. This needs to be kept clearly in mind by users of the ethnicity and birthplace data, especially data at the lower levels in the classification, like L4 for ethnicity.

The mix of sources for ethnicity and birthplace data for 2018 Census has improved coverage of the population overall in terms of these variables and, in the case of the birthplace variable there has also been a commensurate improvement in the quality of the data. This is

because birthplace is a variable that does not change over time, so timeliness of the sourcing of someone's birthplace is not a critical quality issue.

In the case of ethnicity, improved coverage does not necessarily equate to improved quality of data, as the panel has shown with regard to the L1 ethnic groups in its initial report, especially with reference to the Māori population. Ethnic identity is not a stable variable like birthplace – a person's ethnic identity can change over time. There is also not the same level of consistency across different admin data sources in ethnicity classifications as there is for birthplace. In the 2018 Census, 29 percent or more of the ethnicity data for Māori and Pacific ethnic populations came from other sources. The Māori and Pacific ethnicity data is not of the same quality as the data for the NZ European ethnic population, for example, which had only 11.5 percent of their responses from these other sources.

4.2 Assessing the quality of L4 ethnic data

Stats NZ have prepared an assessment of the quality of the L4 ethnic groups on Metric 1 of the three metrics they use for this purpose (see Appendix 3). Metric 1 is "data sources and coverage" and it is measured by a score determined by the quality of the data sources used to produce responses for a census variable. Metric 2, "consistency and coherence" and Metric 3, "data quality", are not included in an unpublished rating of the quality of L4 ethnicity data.[11]

The method employed by Stats NZ to derive quality ratings for L4 ethnic groups has to take account of the numerous nfd and nec categories in the database. These groups also appear in earlier censuses when presenting data for ethnic groups, so there is nothing inconsistent about their presence in the ethnicity database for 2018 Census. Some examples using 2018 and 2013 Census data are provided in Table 4.1.

It is apparent from Table 4.1 that there was not necessarily an increase compared to the 2013 Census in the overall (Total excluding NZ European) nfd/nec ethnic categories in 2018 linked to lower levels of response to the census. However, there is very considerable variability between some of the major ethnic groups in the shares of their ethnicity responses that are nfd or nec. Of the 2,407,470 responses in the total column in Table 4.1, just over 4 percent were in the nfd/nec categories, a slightly lower share than in the 2013 Census. In the case of the Pacific ethnic responses in both censuses, less than 1 percent were in the nfd/nec categories.

[11] Stats NZ (2019b) 'Deriving a quality rating for level 4 ethnicity', unpublished report, November 2019

Table 4.1: Significance of the "nfd" and "nec" categories in MELAA and Pacific ethnicities

2018 Census usually resident population

Ethnic category	MELAA ethnicities			Pacific ethnicities	Total excl. NZ Euro.**
	Middle East	Latin America	Africa*		
2018 Census					
Specified ethnicities	20,037	15,924	9,234	419,976	2,303,175
Not further defined (nfd)	8,268	9,798	7,221	2,724	93,330
Not elsewhere classified (nec)	321	423	1,005	336	10,965
Total ethnicity responses	28,626	26,145	17,460	423,036	2,407,470
Percent specified ethnicities	70.0	60.9	52.9	99.3	95.7
Percent nfd plus nec	**30.0**	**39.1**	**47.1**	**0.7**	**4.3**
2013 Census					
Total ethnicity responses (excl. residuals)	20,679	13,260	14,340	324,015	1,811,979
Percent specified ethnicities	78.3	56.9	52.4	99.7	89.8
Percent nfd plus nec	21.7	43.1	47.6	0.3	10.2

* Including Mauritian, Seychellois and South African Coloured

** Chinese nfd, Cypriot nfd, Indian nfd and Sri Lankan nfd are included in the specified ethnicities. The very large NZ European category (over 3 million ethnicity responses) is not included in the figures in this column. This category comprises 55.6 percent of the 5,420,931 ethnicity responses and 64.1 percent of the total usually resident population (4,699,755) in 2018. The NZ European ethnic category is not considered further in this discussion of L4 ethnic groups.

The ethnicities linked with the Middle East, Latin America and Africa, on the other hand, had much higher shares of responses in the nfd/nec categories in both censuses. In the case of ethnicities linked with the broad grouping Africa, for example, around half were in these two categories in both censuses. This probably reflects two things. Firstly, there is considerable ethnic diversity in Africa's populations that is not reflected in the Stats NZ classification and only small numbers of usual residents in New Zealand identify with many of the ethnicities that are associated with that continent. Secondly, except for migration

from South Africa and, for a short time, refugee flows from the Horn of Africa, attracting migrants from this continent has never been a priority in New Zealand's immigration policy. Many of the small numbers of migrants from different African countries have been aggregated into the nfd/nec categories in 2018 Census.

4.2.1 Dealing with the nfd responses in L4 quality assessment

In the assessment of the quality of 2018 Census data on ethnicity at level 4 the nfd parts of the classification play a much more critical role than at the higher levels of the classification. Stats NZ point out in their unpublished paper that at L1 and L2 an nfd ethnicity response in 2018 Census was considered to be a valid response in the calculation of the quality rating. At L4 this is not the case and most of the area-specific nfd categories (like Middle East nfd) were excluded from making comparisons with other data sources like the 2013 Census and admin sources. Responses in these categories were considered to be "missing" and given a score rating of zero.

Table 4.2: Ratings for L4 ethnicities by L1 ethnic group

2018 Census usually resident population

L1 Ethnic group	Ethnicities*	Population**	Metric 1 L4 quality score	Stats NZ Metric 1 quality rating ***	Panel internal consistency rating
European	62	3,297,864	0.98	Very high	High
a) NZ Euro	1	3,013,440			Very high
b) Other Euro	61	343,620			High
Māori	1	775,836	0.97	High	Moderate
Pacific****	19	381,642	0.96	High	Moderate
Asian	47	707,598	0.95	High	High/moderate
MELAA	45	70,332	0.62	Very poor	Very poor
Other	6	58,053	0.86	Poor	Poor

* The number of L4 ethnicities in each L1 group (including nfd and nec categories)
** The populations include everyone specifying an ethnicity in the L1 group. The total sums to more than the total population because of multiple ethnicities across L1 groups. In the case of the "other European" population the size depends on which level of the classification is being used. The population cited in the table is the L2 "other European" population which includes "Other European nfd.
*** The quality ratings are the ones specified for the bands of quality scores for Metric 1.
**** Including Indigenous Australians

Stats NZ has calculated a L4 quality rating for each of the L1 ethnic groups in the 2018 Census. The rating is for Metric 1, Data sources and coverage, and reflects the quality of the identification of ethnicity based on the different sources used, and the amount of missing

data. This metric reflects the quality of the counts by ethnicity. They have not derived quality categories for each specific L4 ethnicity (Table 4.2).

Stat NZ observe in this regard that "the calculation of quality ratings separately for each of the small level 4 ethnicities is affected by considerable noise, and results would be misleading in ways that are hard to predict". Stats NZ have also pointed out that the downside to having quality ratings for groups of ethnicities is that the rating can be very much influenced by particular ethnicities if these are very dominant (like NZ Europeans in the L1 European ethnic category).

Using their method for scoring data on the basis of its source, Stats NZ derived quality ratings for the clusters of specific L4 ethnicities, that are included in the six L1 ethnic groups. Their ratings ranged from **very high** for Europeans to **very poor** for the Middle East, Latin America, Africa (MELAA) cluster (Table 4.2).

The panel is rating ethnicity for groups other than NZ European and Māori on the consistency, within groups, of the overall quality of the census information about these groups, not the Metric one score specifically for the aggregates that comprise the groups. Because of the very large shares of the L4 MELAA and 'Other ethnicities' clusters that are in the ethnicity nfd/nec categories, and the low shares of responses in these categories that come from the 2018 Census, the panel agrees with the **very poor** and **poor** quality ratings Stats NZ have given the L4 MELAA and 'Other ethnicities' clusters.

The **very high** rating for L1 European ethnic group is heavily influenced by just one of the 62 specified ethnicities in this group – the NZ Europeans (91.4 percent of the European total). The panel supports a **very high** rating for NZ Europeans because 89.1 percent of their responses were obtained from the 2018 Census and under 1 percent were imputed responses (Table 4.3). Stats NZ have not derived a quality rating for the data relating to the other 61 European specific ethnicities grouped under the label "Other Europeans".

4.2.2 Approach to assessing coherence and consistency in L4 ethnicity clusters

The panel's assessment of coherence and consistency in data sources for clusters of L4 ethnic groups is based on the percentages of responses from the four sources that were recorded for each specified Other European, Asian and Pacific ethnicity[12]. In this analysis equal weight is given to the percentages for each country. This is because, as Stats NZ (2019b**)** point out, there are significant variations in numbers of people identifying with the specific L4 ethnicities. Where contributions to L1 ethnic groups are weighted by population

[12] The method that has been used here to assess coherence and consistency in L4 ethnic groups is not as rigorous as the method used in the panel's Initial Report to assess quality of the L1 and L2 ethnic groups. In the case of the latter the multiple ethnicity responses given by individuals were assessed separately when drawing data from the 2013 Census and other sources of administrative data. This method could be applied to all ethnic groups but for the purposes of the brief analysis in this section a method focusing on total numbers specifying a particular ethnicity (irrespective of their other ethnic affiliations if they have these) was used.

size the variability in populations found in the small ethnic groups is masked. This is illustrated in Table 4.3.

Table 4.3: Measures of consistency in data sources, Pacific ethnicities (exc. nfd, nec) and summary data for European and Asian ethnicities, 2018 Census usually resident population

	Percentages				
Specified ethnicities Level 4	2018 Census	2013 Census	Admin data	Imputation	Population
Samoan	68.7	16.2	13.8	1.4	182,721
Cook Islands Māori	69.0	16.0	13.8	1.2	80,532
Tongan	**63.4**	18.7	16.5	1.4	82,389
Niuean	70.5	15.5	12.6	1.3	30,867
Tokelauan	69.1	17.2	12.7	1.1	8,676
Fijian	66.1	12.7	**19.6**	1.5	19,722
Hawaiian	79.0	13.3	5.6	1.4	429
Kiribati	71.7	18.0	8.2	2.1	3,225
Nauruan	82.2	11.1	4.4	**0.0**	135
Papua New Guinean	85.7	**8.8**	3.7	1.3	1,131
Pitcairn Islander	**91.7**	8.3	1.4	**0.0**	216
Rotuman	87.5	9.2	2.8	0.9	981
Tahitian	74.1	15.5	8.6	1.7	1,737
Solomon Islander	85.7	11.2	1.5	1.9	777
Tuvaluan	64.4	**22.2**	11.7	1.6	4,653
Ni Vanuatu	81.8	11.8	**0.0**	**6.4**	990
Average (Pacific)	75.7	14.1	8.6	1.6	
Standard deviation (Pacific)	**8.8**	**3.8**	**5.8**	**1.4**	
NZ European (single ethnicity)	89.1	6.9	3.2	0.9	
Average (Other Specified European)	88.9	7.2	1.9	2.3	
Standard deviation (Other Specified Euro)	**3.9**	**2.9**	**1.4**	**2.3**	
Average (Specified Asian)	89.6	6.4	2.2	1.6	
Standard deviation (Asian)	**8.4**	**5.4**	**3.3**	**1.0**	

* Numbers highlighted in red indicate the highest and lowest percentages within each source category. Numbers highlighted in blue indicate the standard deviations for the Pacific, Other European and Asian ethnicity clusters.

Table 4.3 lists the percentages of responses obtained from the four major sources of data in the 2018 Census for each of the specified Pacific ethnicities. The numbers of people identifying with each of the ethnicities are given in the final column in the table. These

ethnic populations range in size from 182,721 Samoans to 135 Nauruans. The Samoans account for almost half of the total Pacific ethnicity responses (381,642 – Table 4.2) and their percentage shares from each data source play a major role in determining the averages for each of the data sources shown in the table. But it is clear from the percentages from each source for each ethnicity that there is quite a bit of variability in the percentage shares from the 2018 Census and administrative data sources in particular. A measure of this variability is given by the standard deviation which is discussed below.

Given that this is the first time ethnicity responses have been obtained from sources other than the census in question, and that Stats NZ have given the different sources of responses in 2018 different quality ratings, it is important that users of L4 ethnicity data appreciate the considerable within-L1 group variability in shares of data from different sources. This variability is indicated in the relationship between the average percentages of responses from each source for the different specific ethnicities in a particular L1 group and their associated standard deviations. The ethnic groups that this examination of intra-group variability in data sources has been applied to are the Pacific Peoples, the Other Europeans and Asian clusters. The averages and standard deviations for the percentages of ethnicity responses from each data source are shown in the lower part of Table 4.3.

While within-group variation in sources of ethnicity data in the 2018 Census is not a measure of quality per se, it is a useful test of the consistency of the data sources for responses recorded for specific ethnicities grouped in the L1 ethnic categories. The critical measure of this consistency is the standard deviation, which is a measure of the amount of variation in a given distribution of values – a low standard deviation indicates that most of the values tend to be close to the average, while a high standard deviation indicates that the values are quite dispersed around the average. In the case of the L4 ethnicity data for Pacific Peoples, Other Europeans and Asians, the standard deviation measures the variation in percentages of responses from a particular source of data for each ethnicity within these broad groups.

Ratings for Other Europeans and Asians

In the case of the 54 specified ethnicities (excluding nfd and nec cases) in the Other European group, an average of 88.9 percent of the ethnicity responses came from the 2018 Census (Table 4.3). There was a standard deviation (SD) of ± 3.9 percent for this source of data amongst the 54 Other European ethnicities. This means that around two-thirds of the 54 specified ethnicities for Other Europeans had between 85.0 and 92.8 percent of their responses from the 2018 Census. The lowest percentage (75.0 percent) from this source was for the small Flemish population (60 in total).

On average, 2.3 percent of the ethnicity responses for Other Europeans were imputed (SD ± 2.3) (Table 4.3). The highest percentage from this source was for the 7,677 French – 10.4 percent of their ethnicity responses. While this is the highest average percentage of ethnicities derived by imputation methods for the three groups being discussed, the

consistently high percentages of responses from the 2018 Census suggest that a **high** quality rating is appropriate for the Other European ethnicities.

The panel considers a **high/moderate** rating for the 40 specified Asian ethnicities is more appropriate than a rating that is equivalent to the one for Other Europeans. Although the specified Asian ethnicities have an average of 89.6 percent of responses from the 2018 Census, the standard deviation (SD ± 8.4 percent) for this group is considerably greater than that for the 54 Other European ethnic groups (SD ± 3.9 percent) indicating a greater spread of values around the average.

Asian ethnic groups also had higher standard deviations for the percentages of data sourced from the 2013 Census and from other administrative sources than the Other Europeans. Only for the percentages of ethnic responses derived by imputation did the Asians have a smaller SD than the Other Europeans (Table 4.3). These statistics indicate a more variable degree of consistency in the shares of data from the main sources of high-quality ethnicity data for the Asians (Table 4.3). This is consistent with a **high/moderate** quality rating.

The case of Pacific ethnicities

On the basis of an assessment of consistency and coherence in the sources of data for L4 Pacific ethnicities the panel supports a **moderate** quality rating on the basis of within-group variations in the sources of ethnic data for the 16 Pacific ethnicities are shown in Table 4.3. The standard deviations for percentages of responses sourced from the 2018 Census and admin data indicate more variability amongst the Pacific ethnicities than is the case for the Other European and Asian ethnicities.

Pacific ethnicities have a much lower average share of responses from the 2018 Census (75.7 percent) than the other two ethnicity clusters. A standard deviation of 8.8 for the data sourced from the 2018 Census indicates that the shares from this source for two-thirds of the Pacific ethnicities fall between 66.9 percent and 84.5 percent. The highest share (91.7 percent) was for the small population of Pitcairn Islanders (216) usually resident in New Zealand, while the lowest share (63.4 percent) was for the second largest group, the 82,389 New Zealand residents with Tongan ethnicity (Table 4.3).

It is the larger, long-established Pacific populations (Samoans, Tongans, Niueans, Cook Island Māori, Fijians) that have smaller shares of their responses from the 2018 Census. This could reflect two things – firstly the availability of data on these populations in the 2013 Census and from a range of admin sources, and secondly their distribution across a range of suburbs in the main urban areas where census enumeration was less complete than in other areas. The smaller resident Pacific populations (Papua New Guineans, Solomon Islanders, Ni Vanuatu, Nauruans, Tahitians), which owe their origins to more recent migration flows had higher shares of responses from the 2018 Census. The smaller shares of their responses from the 2013 Census and admin data sources is an indication of their recent migrant status.

Larger shares of Pacific ethnicity responses were sourced from the 2013 Census and admin sources than was the case for the Other Europeans and the Asians. However, only a small

share of responses came from imputed data (average of 1.6 percent). One country, Vanuatu, had a much larger than average share of imputed ethnicity responses (6.4 percent) and this is rather surprising given that Vanuatu, like Papua New Guinea and the Solomon Islands, is primarily a source of temporary seasonal workers rather than migrants seeking residence.

The low average share of responses obtained by imputation, and a small standard deviation (1.4 percent) for this source, suggest that a **moderate** quality rating for Pacific ethnicity data at the L4 level in the classification is appropriate. As was shown in Table 4.1, there is a very small proportion (less than 1 percent) of the Pacific L1 ethnic group that had ethnicity responses in the nfd or nec categories – the categories that made a major contribution to Stats NZ's **very poor** quality rating for the MELAA ethnicity data. Data for specific Pacific ethnicities is clearly of a higher quality than that for the various MELAA ethnicities.

4.3 Migrant and non-migrant ethnic groups

Users of data for specific ethnic groups often want to identify the migrant and non-migrant components of the relevant population. As noted earlier, a common way of doing this is by separating the NZ-born from the overseas-born in an ethnic group. The shares of the total L1 ethnic categories that are NZ-born and overseas-born differ markedly (Table 4.4).

Table 4.4: NZ-born and overseas-born components of the major ethnic groups, 2018 Census usually resident population

	Percentages		
Ethnic group	NZ-born	Overseas-born	No info.
European	82.8	17.2	1.0
a) NZ European	89.3	10.7	0.9
b) Other European	22.1	77.9	1.8
Māori	98.0	2.0	1.5
Pacific	66.4	33.6	2.2
Asian	23.0	77.0	1.5
MELAA	23.0	77.0	1.6
Other ethnicity	76.4	23.6	1.7
Total population	**71.7**	**27.1**	**1.2**

It is clear from Table 4.4 that the L1 ethnic groups comprise three reasonably distinct groups:

- those with over 75 percent NZ-born (Māori, NZ European, and Other Ethnicity)
- the Pacific ethnic group with just under two-thirds in the NZ-born category; and

- the mainly migrant populations comprising the Other European ethnicities, the Asian ethnicities, and the MELAA ethnicities.

The marked difference in shares of NZ-born and overseas-born for the NZ European and Other European components of the European L1 ethnic group is another reason for suggesting that this broad category is broken into two components when considering data quality issues. There are also marked differences between specific ethnicities in some of the other L1 groups (e.g. Pacific) and these need to be kept in mind by users of the data. There is considerable within-group variability in the shares of NZ-born and overseas-born at the L4 level of the ethnic classification.

4.3.1 The sources of data for the birthplace components of ethnic groups

The final dimension of ethnicity that we examine in this section is the shares of data that come from different sources for these two birthplace components of the major ethnic categories. These are shown in Table 4.5. As noted earlier, none of the data by birthplace was imputed.

In the case of the NZ-born components, the shares of ethnicity responses obtained from the 2018 Census are lower than those for the overseas-born for the European, Māori and Pacific ethnic groups, but slightly higher for the other three ethnic groups. The Pacific and Māori ethnic groups have the highest shares from the 2013 Census, with the Pacific and Other European populations drawing most heavily on admin sources (Table 4.5).

Table 4.5: Sources of birthplace data for the NZ-born and overseas-born components of the major ethnic groups. 2018 Census usually resident population

	Source of data (percentages)					
	NZ-born			***Overseas-born***		
Ethnic group	2018 Census	2013 Census	Admin sources	2018 Census	2013 Census	Admin sources
European	88.8	7.3	3.9	90.2	5.8	4.0
a) NZ European	89.2	7.3	3.5	92.4	5.8	1.8
b) Other European	71.1	8.4	**20.5**	87.8	5.7	6.5
Māori	71.2	15.4	13.3	84.4	11.6	4.1
Pacific	**67.8**	**16.3**	**15.9**	**69.6**	**18.9**	**11.6**
Asian	85.3	6.8	7.9	83.9	6.9	9.1
MELAA	81.7	8.3	10.1	81.1	8.4	10.4
Other ethnicities	86.1	10.7	3.2	78.1	3.9	**18.0**
Total	**84.7**	**9.0**	**6.2**	**85.0**	**7.7**	**7.3**

The overseas-born populations tend to have higher shares from 2018 Census (only Pacific and Other ethnicities have below 80 percent of their responses from this census) than the NZ-born populations. The overseas-born Pacific ethnic group stands out in terms of its much

higher reliance on 2013 Census and admin data to provide ethnicities, while the Other Ethnicities group has the highest share (18 percent) coming from admin sources. The latter group includes 2,664 in the L4 Other ethnicity nec. category, 87.3 percent of whom were derived from admin sources.

Classification of the L1 ethnic groups into two birthplace components has not shown markedly higher percentages in the share that comes from 2018 Census in one of the sub-groups compared with the other which could have indicated an improvement in data quality. The within-L1 ethnic group variation in shares from different data sources for the NZ-born and overseas-born components of the ethnic populations is also less than that found for the ethnic groups as a whole. On this basis, the panel sees no reason to suggest different quality ratings for sub-groups of the ethnicity data classified on the basis of birthplace in New Zealand or overseas.

4.4 A concluding comment

In this section we have reviewed aspects of data quality for the most specific categories coded for a variable which has a multi-level classification for responses. The exercise has been useful because it makes clear that assessments of data quality for information at the highest levels of aggregation do not necessarily hold for data presented at lower levels of aggregation in the classification. The quality of data at lower levels of aggregation, like the one used for ethnicity, is not necessarily higher or lower than that at other levels. What is important to keep in mind is that there is considerable variability between the specific ethnicities within the L1 groups, and this variability only becomes visible with disaggregation of the data. In this section, the full range of quality categories from **very high** to **very poor** apply when assessing ethnicity data at the L4 level of the classification.

This variability in data quality for particular ethnic groups or clusters of specific ethnicities, in turn, has an important equity dimension. The lower quality data are generally for the non-European ethnic groups, many of whom are not very visible in many of New Zealand's statistical databases. Poor quality census data for these groups means that they are further disadvantaged and marginalised in the one source that aims to produce high-quality data on all ethnic groups.

In its Assessment of Variables Report the panel points out that the quality ratings for data for all of the variables assessed at high levels of aggregation are likely to be different, when data for these variables are examined at lower levels of aggregation (such as SA2 units – see section 7 on small area data in this report). Data quality will also be different when ethnicity is used in combination with other variables relating to individuals in cross-tabulations or statistical analyses.

In the light of evidence presented in the panel's three reports, it is very important that users of the 2018 Census data keep in mind that the general assessments of quality that have been produced by Stats NZ apply at high levels of data aggregation. They are not necessarily

appropriate guides to quality at lower levels of the coding classification systems for many variables, or when data are being examined for small areas.

R 15. Stats NZ should ensure that users of 2018 Census data have readily available to them:

15a Metadata on the relative contributions of different data sources for every variable at all levels in their coding classifications for levels of aggregation down to SA2, and for level 4 ethnic groups.

15b The associated quality rating (at least for Metric 1, data sources and coverage) should be provided at all levels of spatial aggregation (SA2 to region) and for all levels in coding classifications that are frequently employed by users (e.g. L4 for language, L3 for religion, L4 for ethnicity, etc.).

5 Languages spoken

The panel has assessed "languages spoken" in its Assessment of Variables Report. Whilst the assessment of this variable at high levels of geography, and across all languages grouped together may justify the Stats NZ quality rating of **high**, this ignores the fact that a critical use of this data is to support Treaty of Waitangi and international human rights obligations around te reo Māori, as well as for language planning purposes.

There are three official languages in New Zealand: English, Māori, and New Zealand Sign Language. The information on data sources provided to the panel in the WoF was at Level 1 of the classification, where only New Zealand sign language is identifiable – this has a quality rating of **high**. Stats NZ has provided the panel with a metric 1 quality rating for English, of **high**, and of te reo Māori, which is rated as of **poor** quality, which reflects a high degree of variability for te reo Māori responses between 2013 and 2018. No other language below Level 1 has been rated.

It is clear that the quality of language data varies by language, and that the appropriate quality ratings need to be based on the Metric 1 quality of categories at Level 4 of the classification**.** These quality ratings range from **high** to **poor**. It is the Panel's view that this range best represents the quality of this dataset and that an overall quality rating at Level 1 of the classification does not make sense.

Note that while the panel only has evidence that one Level 4 language – te reo Māori – is **poor** quality, there are likely to be others, especially given the extensive use of alternative data sources (2013 Census, imputation) for speakers of some languages (e.g. Pacific peoples). Similarly, there are almost certainly other Level 4 languages that should be rated as moderate, high, and very high, but we currently do not know which languages these are. There are unlikely to be any languages rated as very poor, as this would require that less than about 25 percent of data was derived from the 2018 Census, which is highly unlikely.

The WoF notes "Due to data quality issues we caution the interpretation and use of this data at level 4." This is precisely the level at which most of the individual languages, including te reo Māori, are identifiable.

The rest of this section focuses on te reo Māori. The material presented here is a summary of a standalone brief on te reo Māori intended for key users that will be published separately on the Stats NZ website.

5.1 Te Reo Māori

Although Stats NZ regards 'languages spoken' as a Priority 3 variable, te reo Māori is a crucial component of collective Māori identity and is part of what makes Māori as a people,

and Aotearoa New Zealand as a country, unique (Kukutai, Rarere & Pawar, 2015). Te reo Māori is considered a taonga (highly prized object or resource) under Te Tiriti o Waitangi (Waitangi Tribunal, 2011) and has been an official language of Aotearoa NZ since 1987.

While use of te reo Māori is not legislatively required to be collected in the census, high quality, time-series data on te reo is necessary for monitoring the health of the language, and for supporting the work of national bodies such as Te Mātāwai (the independent statutory entity charged with revitalising te reo Māori) and Te Taura Whiri i te reo Māori (Māori Language Commission), as well as government agencies, iwi and Māori communities. As such it is important that the Panel assess the quality of te reo Māori data generated from the Census 2018 dataset.

Given that the vast majority of te reo speakers are Māori, we limit our analysis to individuals whose ethnicity is recorded as Māori, either alone or in combination with at least one other ethnic group (for an assessment of the Māori ethnicity variable, see: Initial Report of the 2018 Census External Data Quality Panel, 2019). We note that Stats NZ developed its quality rating method for data sources (see below) to be used at the variable level (e.g. language spoken) rather than the category level (e.g. te reo). However, as we have already shown with ethnicity, the aggregate quality rating for a single variable masks substantial internal variation and cannot be treated as a reliable gauge of the quality for specific groups.

5.1.1 Data sources and coverage

Table 5.1 shows that only 65.1 percent of te reo data for Māori respondents in the 2018 Census dataset came from individual census forms (paper and online combined). A further 18.6 percent was sourced from 2013 Census data. The remaining 16.3 percent was derived from various forms of imputation.

Table 5.1. Data source quality for te reo Māori, for individuals recording Māori ethnicity

Data source for te reo Māori speakers	Count	Percent	Quality weight (percent consistency with 2018 te reo response)	Score contribution
2018 Census form	103,935	65.1	1.00	65.1
2013 Census	29,688	18.6	0.56	10.4
Within household donor	5,205	3.3	0.70	2.3
Donor's 2018 Census form	16,500	10.3	0.60	6.2
Donor's response sourced from 2013 Census	3,729	2.3	0.34	0.8
Donor's response sourced from within household	585	0.4	0.42	0.2
Total	159,645	100.0		85

In contrast, the Assessment of Variables Report shows that at the national level and for all languages, 83.8 percent of data came from the 2018 Census, 8.2 percent from 2013 Census data, and 8.0 percent from imputation.

Stats NZ computed a quality rating for Māori speakers of te reo. It measures the consistency of individuals' te reo responses in the 2018 Census with their responses in the 2013 Census, or with statistical imputations. For each source, the resulting quality weighting (0.00 to 1.00) is multiplied by its proportional contribution to the output for te reo and summed to derive an overall quality score.

Overall, the quality rating score is 0.85, which leads to a Metric 1 (data sources and coverage) rating of **poor** quality for te reo data for Māori. Stats NZ also computed a quality rating score for te reo for the total population and this was only fractionally higher at 0.86.

Table 5.2 below shows a significant level of inconsistency in te reo Māori reporting for 386,100 Māori (ethnic) respondents whose 2018 and 2013 Census forms could be linked. Of the 74,400 Māori who reported speaking te reo in Census 2018 (and whose record could be linked to their 2013 Census record), only 62 percent had also spoken te reo in the previous census.

Table 5.2: Agreement between te reo Māori response in the 2013 and the 2018 Censuses for individuals recording Māori ethnicity

	2013 Census – te reo response	
2018 Census – te reo response	Yes	No
Yes	46,200	28,200
No	25,000	286,700

Of the 311,700 Māori who did not speak te reo in Census 2018 (and whose record could be linked to 2013), 8 percent had spoken te reo in 2013. The gains in te reo speakers in 2018 were slightly bigger than the losses. We do not know how representative these patterns are, as the linked records only comprise half of all Māori in the final Census 2018 dataset.

We note that the **high** quality rating given by Stats NZ to the overall 'languages spoken' variable here is due to the numerical dominance of English-speaking individuals recorded as NZ European. NZ Europeans had higher response rates to the Census (and thus a lower contribution from alternative data sources), and almost certainly had a much higher level of language consistency between 2013 and 2018.

English is the dominant language in Aotearoa NZ and the vast majority of New Zealanders are monolingual. The **poor** quality rating for te reo Māori confirms that users should not assume that the 'languages spoken' variable is high quality for all languages.

5.1.2 Consistency and coherence

Table 5.3 shows the number and proportion of Māori speakers of te reo in 2018 is much higher than we would expect on the basis of census counts and inter-censal growth. The 27 percent increase in the number of te reo speakers between 2013 and 2018 is a striking comparison to the nearly five percent decrease in speaker numbers recorded between 2006 and 2013. The 2013 Census was the first census for which an absolute decline in te reo speakers was recorded since the introduction of the language question in 1996, but the proportion of te reo speakers had declined with each census.[13]

Table 5.3. Number and proportion of Māori speaking te reo, 2006, 2013, 2018

	Census			2006-2013		2013-2018	
	2006	2013	2018	N	percent change	N	percent change
N	131,613	125,352	159,645	-6,261	-4.8	34,293	27.4
Percent	23.3	20.9	20.6				

Note: The percent of Māori te reo speakers includes in the denominator individuals that did not give a valid response to the languages spoken question.

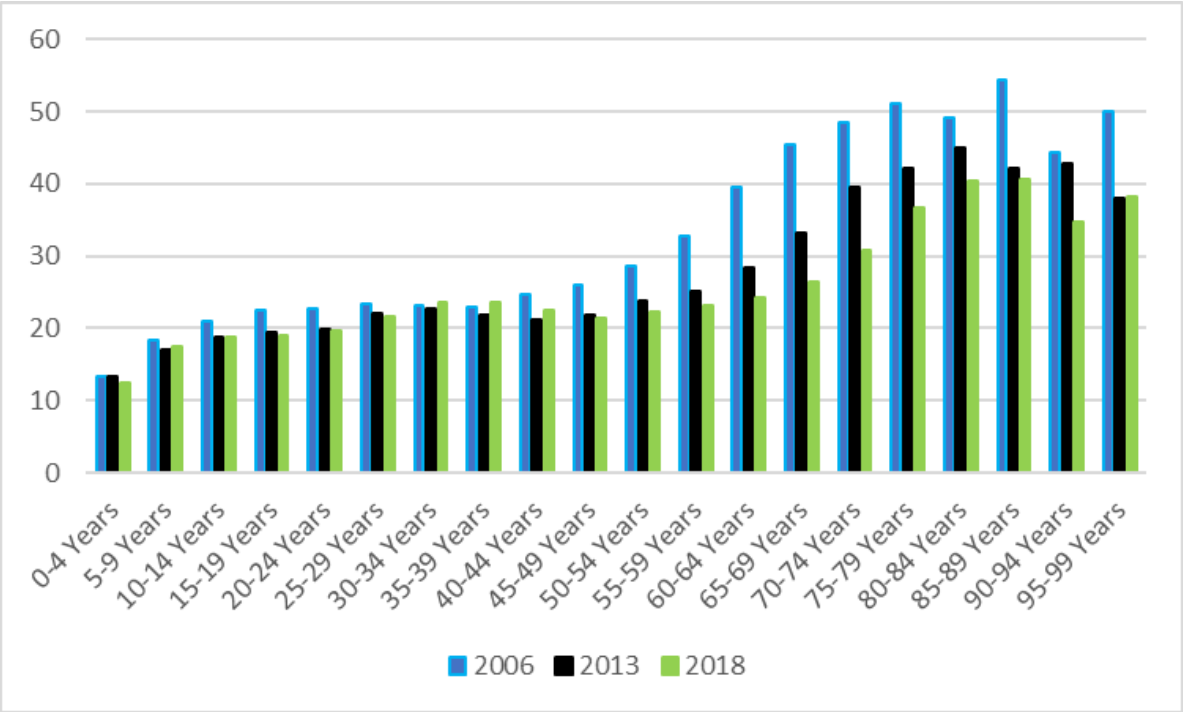
Although not shown in Table 5.2, the number of Māori speakers of te reo only increased by less than 1 per cent between 2001 and 2006 (from 130,482 to 131,613). Thus, the very significant increase recorded in 2018 goes against recent trends and is almost certainly the result of inflated numbers arising from the use of historic census data and imputation in the 2018 Census, rather than a substantial real-world change in the number of Māori able to speak te reo.

This interpretation is supported by Figure 5.1 which shows that the age-specific te reo speaker rates in Census 2018 do not fit well with previous census trends. In particular there were unusual changes in the percentage of te reo speakers at ages 25-44 years, and smaller than expected[14] declines at ages 50 years and older.

[13] We note that in Te Kupenga 2013, the proportion of Māori who reported that they could speak te reo 'very well', 'well' or 'fairly well' was 22.6 per cent, which was fairly close to Census 2013, albeit that Te Kupenga only included adults aged 15 years and older.

[14] Prior research has noted the significant impacts of cohort effects on cross-sectional te reo speaker rates. In short, as older cohorts of Māori pass away they are replaced by cohorts with lower levels of te reo, largely reflecting historical factors including assimilation policies and lack of access to te reo (Kukutai, Rarere & Pawar, 2015).

Figure 5.1: Percentage of te reo Māori speakers in the 2006, 2013 and 2018 Censuses for individuals recording Māori ethnicity



5.1.3 Overall data quality rating

The panel does not judge the Census 2018 te reo data to be of sufficient quality to be used with confidence and rate it of **poor** quality. Certainly, it ought not to be used in conjunction with te reo data from previous censuses to undertake time series analysis. This would lead to very dubious findings of a significant increase in the number of te reo speakers, and only an imperceptible decline in the share of te reo speakers. Neither of these findings is likely to be reflective of the state of te reo Māori.

The poor quality of te reo data from Census 2018 is a significant failing, given the acknowledged importance of te reo to Māori, and to the nation. This shortcoming is amplified given the lack of nationally representative te reo data outside of the census with which to undertake time-series analysis.

R 16. Stats NZ should work with key users such as Te Taura Whiri i te reo Māori and Te Mātāwai to clearly communicate the problems with te reo Māori data from Census 2018, the limitations around its usage, and options to futureproof te reo Māori data moving forward.

6 Families and households

6.1 Introduction

Information about families and households that can be obtained from New Zealand's Census of Population and Dwellings is vital for public policy, for meeting Treaty obligations to Māori[15], for population projections and for the derivation of analytical measures such as household crowding and social deprivation. In a recent report, Stats NZ state that: "Family and household information is used to develop and evaluate government policies, such as income support, social housing, housing affordability, family violence, child poverty, household crowding, household income, household expenditure, and household net worth."[16]

These are clearly very important data, and the question of the quality of the 29 variables or derived variables that comprise the suite of family and household outputs is a very important one. The 29 variables relating specifically to families and households that have been generated from 2018 Census are listed in Appendix 1.

The panel's assessment of families and household data presented here is based on the information available at the time the panel was completing its final report and is consistent with the public position of Stats NZ (e.g. in DataInfo+ pages) as at early December 2019. We have seen some additional analyses carried out by Stats NZ during December 2019. However, due to time constraints and the fact that Stats NZ is still completing its further inquiries into a range of quality-related issues for the families and households data, we have not been able to review what these emerging analyses tell us about the quality of the families and households suite of variables.

In their initial assessment of the quality of the data in the suite of 29 variables relating specifically families and households, Stats NZ gave the entire suite a **very poor** rating. This is the rating that has been in the public domain for some time now through DataInfo+.

In a recent paper shared with the panel on 10 December 2019, Stats NZ's re-evaluation indicates that the suite of families and households variables merit a quality rating of **high-moderate** depending on the variable (Metric 1), or **moderate-very poor** (Metric 3)**,** not their

[15] For example *Stats NZ's Strategic Intentions 2019-2023* notes that Stats NZ, as the leader of the government data system, needs to ensure that decision-makers have the data and information they need to make decisions, and that there is "improved representation of Treaty partners in our data and products". https://www.stats.govt.nz/corporate/stats-nzs-strategic-intentions-201923

[16] From Stats NZ 'The potential for linked administrative data to provide household and family information' https://www.stats.govt.nz/research/the-potential-for-linked-administrative-data-to-provide-household-and-family-information

initial overall rating of **very poor**[17] for the full suite of variables. This is a very significant shift and the panel encourages Stats NZ to publish a paper on the website that explains to users what accounts for this very different assessment of data quality for the suite of variables. As noted above, on the basis of evidence available to the panel up to the beginning of December, the panel considered Stats NZ's initial **very poor** rating to be the appropriate one for these data and this section is based on that assessment.

6.1.1 Use of administrative data

The use of administrative data to fill gaps has improved the quality of the overall dataset, but it has not been able to improve the quality of the families and households datasets. This is because Stats NZ estimate that 357,294 people (from administrative data) were not able to be placed into the dwelling that they would have been in at the time of enumeration and therefore were not able to be given a specific position in a family or household (these are derived at the dwelling level). This means that either whole households and families are missing, or others are incomplete, which impacts the total numbers of families and households along with the family structure and household composition.

These missing households are in meshblocks where response rates to the 2018 Census were low. They include a disproportionate number of meshblocks in areas where Māori and Pacific populations comprise a high proportion of the residents. This means that Māori and Pacific household and family formations are under-represented in the information available from the 2018 Census about families and households.

This is a problem that is not unique to the 2018 Census data. The 2013 Census also had substitute households which had no family or household information. But given the much lower response rates for Māori and Pacific peoples in the 2018 Census (around 70 percent - see Initial Report of the 2018 Census External Data Quality Panel, 2019) the problem is exacerbated because family composition and structure vary considerably amongst different groups in the population.

The loss of robust family and household information is amplified for Māori given their distinct household and family characteristics, and statutory obligations to understand and support whānau wellbeing. Minority ethnic communities will also lose some of their already limited visibility in official statistics as the small sample size of household surveys is generally inadequate for producing representative household and family statistics for smaller populations, especially sub-nationally.

In their report on the potential for linked administrative data alone to provide household and family information, Stats NZ point out that: "Overall, the admin sources investigated show potential for providing household information on an aggregate level, despite some

[17] Stats NZ (2019a) Quantifying data quality issues for household and family information, unpublished paper, 5 December 2019.

limitations. However, the lack of coverage of families in admin data means the potential for producing census-type information on families is currently minimal. Missing family information also affects our ability to replicate the census household composition variable."[18]

In any census, family and household data are subject to the following problems:

- People counted at the wrong address
- Duplicate forms
- Incorrect family coding
- Respondent error

In the case of 2018 Census, the very low quality rating given for family and household data is due in large part to 357,294 admin enumerated individuals being located in meshblocks rather than dwellings, leading to incomplete households. In addition, some known quality problems were exacerbated by a number of operational issues addressed in the *Report of the Independent Review of New Zealand's 2018 Census* (Jack and Graziadei, 2019). These are:

- Missing household data
- Dwellings where there was no response to 2018 Census and thus no individuals or households at the dwellings
- Based off provisional results, 2.3 to 6.6 percent of households which are incomplete (i.e. people are missing).

6.2 Background

By design, censuses of population and dwellings collect information about people and the dwellings they occupy simultaneously so that a connection between the two can be made through the enumeration process. The de facto approach to a census (population present on census night) means that everyone should be captured somewhere. Because of the small size of New Zealand and its population, a good share of the population that is away from its usual residence has traditionally been manually transferred to the latter through judgements made at the processing stage.

The simultaneity of collection in time and place is more difficult to achieve in other ways without a real-time population register, which does not exist in New Zealand. Increasingly, de jure (usual residents) concepts have been included in New Zealand censuses, recognising that the dynamic nature of population groups, and the sensitivity of some family types (e.g.

[18] https://www.stats.govt.nz/research/the-potential-for-linked-administrative-data-to-provide-household-and-family-information

solo parents) to this dynamism is better managed by asking about "usual" practice rather than what was the situation at the date of the census.

After the enumeration stage in March 2018, it became clear that some 357,294 people from admin records (7.6 percent) could be counted in a meshblock (small area) in New Zealand but could not matched to a dwelling. This means their relationship to the occupier of the households or families to which they should belong cannot be established. Such a situation arose because of changes both in the enumeration process and the way data processing was managed, as at both these stages there was a loss in the capacity to match individuals within dwellings to establish households and families.

Although 2018 Census has been essentially post-stratified around two distinct frames, one a person list and the other a dwelling list, there is no frame of households and families. While dwellings are relatively fixed, persons must be located in some place at a particular time. The identification of households and families requires clear-cut connections to be identified between dwellings and persons.

Even when a census enumeration has achieved expected response rates, the capacity to enquire of people about their connections with others in households is quite limited in a population census. The census dwelling form/household set-up form asks about the relation each person has to person 1 and does not have space to ask about relationships between other people in the household. This generally limits the ability of family statistics obtained in the five yearly Census of Population and Dwellings to reflect adequately the contemporary diversity of family forms. The measures continue to be most suited for analysing the construct of the nuclear family.

6.2.1 Families and households data

The Census of Population and Dwellings has always sought to collect information on:

- The number of people
- Their characteristics (e.g. ethnicity, education, etc.)
- Their relationships to others in their dwelling
- Information on the dwelling (e.g. number of rooms, etc.)

By collecting complete (or near complete) information at the same time on both dwellings and the people that occupy them the census has been able to produce integrated outputs about:

- The population (numbers and their characteristics)
- Households and families
- Dwellings/the condition of housing

Amongst other things, this has allowed for analyses of:

- families, (single parent families; same sex couple families; extended families, etc.) and households (e.g. households containing multi-generational families), cross tabulated by characteristics of the family/household
- overcrowding and unmet housing need (by combining information on people and the dwellings in which they live);
- housing attributes and conditions (rooms and facilities, conditions like dampness, etc.).

Data on families and household enable unique perspectives to be obtained, at the local level, on the social structure of communities. What is unique and special about the family and household information from a Census of Population and Dwellings is that statistics are available to analyse changes for small areas (e.g. at the SA1 and SA2 levels of geography), and for small population groups defined by, for example, age or other characteristics (such as income).

Household composition, household affordability, the distribution and sources of household income, housing stock by SA2 (occupied, unoccupied, number of bedrooms, dwelling type), crowding, ownership, tenure, sector of ownership, rent paid and number of residents in non-private buildings are the main housing indicators cited by users. For example, the Ministry of Social Development uses information about families and households that is provided by a census to assist it to determine need for benefits, superannuation and other income support.

The census can provide information on small but important outliers that cannot be made visible through data collected in surveys or data that are available in the IDI. Family and household data are used to consider the economic position of women and children where wage and benefit transfers can often be indirect. This data contributes to planning for health promotion programs, media campaigns, and emergency preparedness.

6.2.2 Household composition and family type

Family and household statistics are derived from the information provided by both the occupier of a household about the dwelling that contains the household, and the information provided by individuals in the household about the relationships between them.

While a non-private dwelling contains people who are not expected to be related, for a private dwelling family relationships are identified from the responses each provide. If a dwelling contains more than one distinct family, then they would be identified separately.

Household composition classifies households according to the relationships between usually resident people. The classification is based on how many and what type(s) of family nuclei were present in a household, and whether or not there were related, or unrelated people present.

Family type classifies family nuclei according to the presence or absence of couples, parents, and children.

An extended family is a group of related people who usually reside together:

- either as a family nucleus with one or more other related people, or
- as two or more related family nuclei, with or without other related people.

Included are people who were absent on census night but who usually live in a particular dwelling, as long as they were reported as being absent by the reference person on the dwelling form / household set-up form.

In the 2018 Census of Population and Dwellings, the variation in response rates across place, ethnicity and age has resulted in a reduced and variable share of private households where information was obtained from all the members of a private household from the same source, at the same time.

6.3 Families and households in the 2018 Census

The response/collection rate for people in 2018 Census is estimated at 83.3 percent (on the method used in 2013) and 87.5 percent on the new basis developed for the census in 2018.[19] Not only were individuals missing, as well as some of the whole households in a dwelling, but so were whole dwellings, and the households and families that they contained. Individuals were also missing from within responding households. This has created an unprecedented challenge for Stats NZ as they work towards repairing the missing households and families' data.

6.3.1 Response challenges for 2018 Census

Non-response was concentrated among particular ethnic groups (Māori and Pacific peoples in particular) and in particular regions (Northland, Gisborne, Bay of Plenty) so the impact of response challenges for 2018 Census family and household statistics will be highest for these groups and regions.

Impact of new methods

Stats NZ has a high quality register of dwellings (the Census Dwelling Frame) and, through the use of admin records in addition to census returns, Stats NZ also has a robust frame for people. However, it has not been possible for Stats NZ to link all the admin records to a dwelling.

Stats NZ were able to use administrative data to (i) identify admin records for complete households that were non-responding households in the census enumeration, and (ii) admin records for individuals missing from households who had returned their census forms.

[19] https://www.stats.govt.nz/reports/2018-census-interim-coverage-rates-collection-response-rates-and-data-sources

Despite the new methods introduced to address the issue of low response to 2018 Census, it was not possible to create the connections between individuals and the families or households in which they live for most of the cases where these connections were not made by census respondents.[20] Whilst allocation to meshblock is sufficient to produce high quality usual resident population counts, it has not been able to improve the quality of the families and households datasets.

The response problems and the people missing from households means that much larger than usual number of relationships between people have not been identified, and that this is likely to have affected those family types that are becoming more prevalent (such as three generation families and the inclusion of unrelated adults), which exist at different levels among different ethnic communities. It also means that even simple measures such as comparisons of household size and crowding are of significantly lower quality than expected or required for policy use.

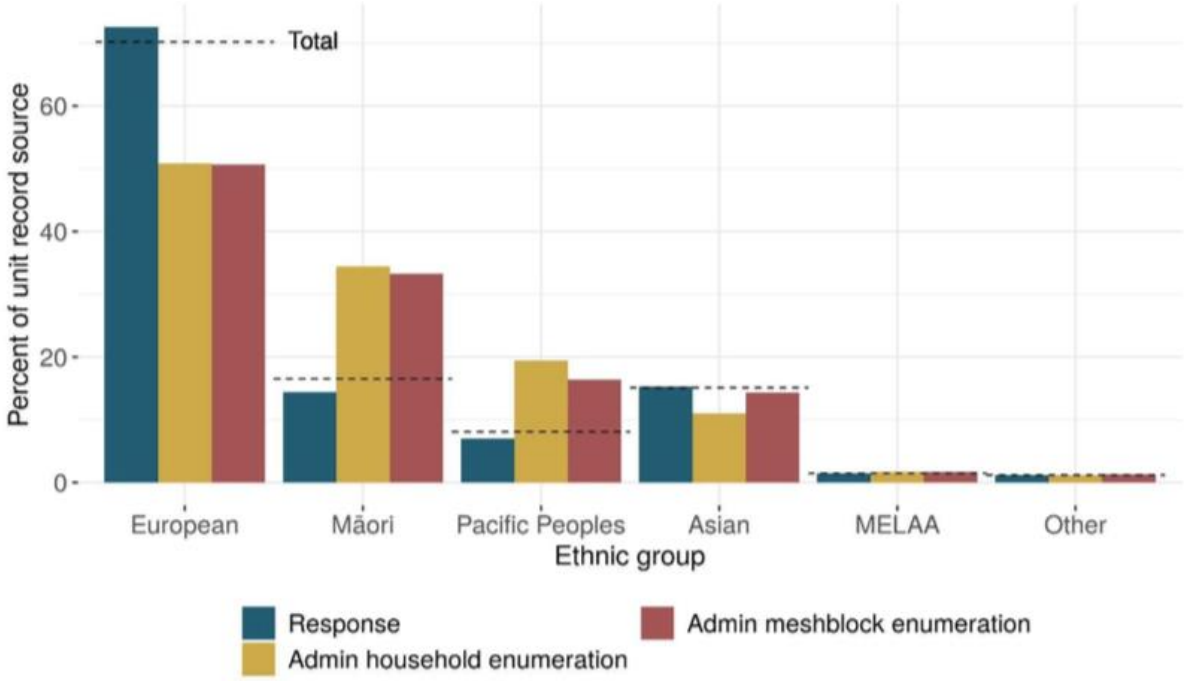
Unpublished Stats NZ analyses seen by the panel at the time of finalising its report shows 3.6 percent of whole households were missing and that provisional results indicate that 2.3 to 6.6 percent of households were incomplete (i.e. had people missing). Panel calculations suggest that up to 10.2 percent of households could be missing or incomplete, which could put the data quality rating for Metric 1 into the Poor category (0.75–< 0.90). This may exaggerate the scale of the problem for some household types, such as where one member of a household of unrelated people is missing. Even the bottom end of this range (5.9 percent missingness) could give a Metric 1 quality rating of Moderate (0.90–< 0.95)

The only information the panel have seen on the distribution of the ethnic group of those records which have not been allocated to households is in the graph below from the Stats NZ paper on adding admin records to the 2018 Census dataset. The panel has not seen information on the socio-economic characteristics of these records.

This graph below (Figure 6.1) shows that Māori and Pacific peoples are disproportionately missing from household and family structures and have had their household data provided by admin records for the whole household and admin records assigned to meshblocks. These two groups have larger family sizes, so the 2018 Census households and families dataset is likely to underestimate family size for these groups, and in the locations where they live.

[20] https://www.stats.govt.nz/methods/overview-of-statistical-methods-for-adding-admin-records-to-the-2018-census-dataset

Figure 6.1: Percent identifying with level 1 ethnic group, by unit record source



Regional and ethnic implications

The allocation of admin records which cannot be located in particular families and households to a meshblock is not evenly spread across New Zealand. Their distribution largely reflecting the non-response profile for the 2018 Census. Such 'meshblock only records' represent 11.2 percent of records in Gisborne Regional Council; 10.8 percent in the Northland region 9.8 percent in Bay of Plenty; down to 5.3 percent in Canterbury.

The disproportionate impact on Māori and Pacific peoples of the non-response problems in 2018 Census have already been mentioned. These are compounded when it comes to analysis at the regional level and smaller spatial units. The poor quality of the families and households data poses a major problem for Stats NZ when it comes to enabling such data to be used in analyses that require an ethnicity identifier – e.g. in customised outputs (rather than standard outputs).

These quality problems with the families and household data, notwithstanding, Stats NZ is encouraged to produce these data by ethnicity, given that household structures vary markedly by ethnicity. Some other census offices produce such data (e.g. based on the ethnicity of person 1); Stats NZ should consider this for the future.

6.3.2 Observed impact of household and family coding issues on data

Stats NZ has carried out comparisons between 2018 Census and 2013 Census distributions of family types. This has identified significant biases in the 2018 households and family data which Stats NZ has not been able to correct. Comparisons of numbers of households by household/family type have not been possible due to the missing households in 2018 Census.

Stats NZ have identified serious problems in how its new processing system handled the complex processing and coding of household and family data, stating in the WoF that “The family coding module has made it difficult to action fixes to records.”

This processing system problem has led to major errors in family relationships, including:

- A large decrease in one-parent families
- Potential undercount of children under 5 years old
- Underage partners in opposite sex couples
- There are some very old “children”, very young “parents”, and very young people living alone
- There is an overcount in same-sex couples

The quality checks identified significant errors which appear to have led to implausible changes in the household/family data:

- A major increase in the number of households composed of a couple and other person(s)
- Implausibly large increases in the age of the older partner in same-sex couples

Correcting these errors would have meant system changes and rerunning of the processing systems – probably leading to changes in the age-sex structure. Due to time constraints to produce outputs in time for use for determining electoral boundaries, Stats NZ were not able to rerun these data to correct the errors.

A Stats NZ paper shared with the panel stated "We had manual operators who were trained in family coding but did not solely focus on it. This was largely due to the lower number of households that went through manual coding – we only sent 2.9 percent of households to manual operators (compared to 18 percent in 2013). This was primarily due to the new family coding process, where higher thresholds were set before a household was sent to manual operators."

Stats NZ state in the WoF that “the lack of a dedicated family coding team meant poorer quality in terms of fixes via manual intervention”. In 2013 staff were dedicated to family coding, but in 2018 staff worked on family coding but were then deployed on other work. “This meant that the quality of the coding was less than ideal” with approximately 79 percent accuracy for manual intervention.

6.4 Conclusion

At the time of writing this report Stats NZ’s stated position (e.g. in the DataInfo+ pages for household and families) is “We are currently investigating whether we can improve the quality of family and household data.”

Given the fundamental problems with the data the panel endorses the Stats NZ quality rating, as at December 2019, of **very low** quality for this suite of 29 variables. Data quality will be lower for Māori and Pacific, and for those regions with low response rates (e.g. Northland, Gisborne, Bay of Plenty). On this basis, the panel do not believe that the data should be made freely available, as there could be a major risk of false conclusions being reached.

There may be particular projects which use a subset of the household and families data which are valid, but such potential uses will need to be considered on a case-by-case basis. This will require procedures similar to those for the IDI where project proposals are considered by Stats NZ and, if approved, researchers are given access to the household and families data in controlled environments (such as the IDI). See also Recommendation 19, pg. 87.

R 17. Stats NZ should support a dedicated team for the 2023 Census to undertake post-processing for families and households data, and other complex variables, and not divert this team to other tasks.

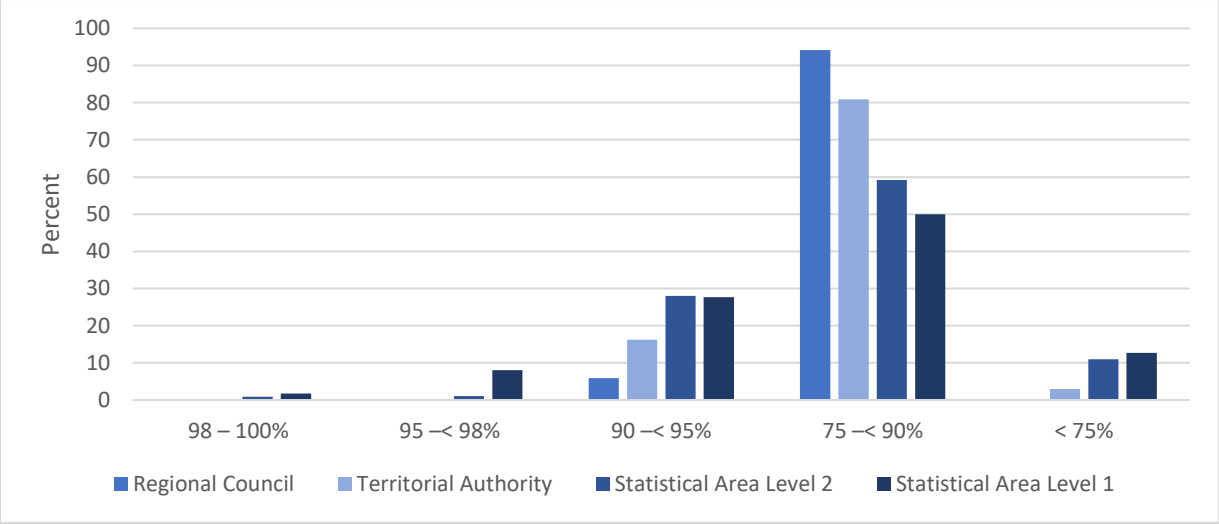
7 Small area considerations

The census uniquely enables data on a range of census variables to be produced at small areas (and also for small population subgroups). Surveys can never produce the micro-data at a small area level that is typically available from the census. Central government, local government, businesses and community groups use highly localised data to allocate resources to local areas, to plan and deliver services (e.g. identifying future demand for care for the elderly), for retail store location decisions, for local advocacy, etc.

Response to the 2018 Census was patchy, with great variation in response across areas of the country. Nationally, 85 percent of the census dataset is made up of people who provided an individual census form, though the percent of census responses differs by variable.

Figure 7.1 shows the distribution of the proportion contributed by census responses for different geographies, using the variable 'usual residence one year ago' as an example. At a national level, 'usual residence one year ago' has 84.6 percent of information sourced from 2018 Census individual form responses, for a quality rating of **poor**. As geographies become smaller, variation increases, with more areas with higher levels of response from census individual forms than the national average, and some with lower levels.

Figure 7.1: Distribution of percent census individual form responses by geographic level, for usual residence one year ago.



In this section the focus is on small areas using the statistical area 2 (SA2) geography. SA2 replaces the former area unit geography and aims to reflect communities that interact together socially and economically. The SA2 classification contains 2,253 SA2s in 2018. SA2s typically have populations of 1000-4000 people.

Some small areas were particularly affected by low response rates, and as such have lower quality data across many variables. For example, there were 24 SA2 areas (one percent of the total) with less than 60 percent census responses from census individual forms for usual residence one year ago (see Table 7.1).

It is notable that 15 of the 24 were in the Auckland region, including the four SA2 areas with the worst census response. Auckland contains only 26 percent of all SA2 areas in New Zealand, yet 63 percent of the SA2 areas with the lowest levels of usual residence one year ago information derived from census individual forms are in this region. South Auckland was particularly badly affected, with 12 SA2 areas with less than 60 percent response found in the South Auckland local boards of Otara-Papatoetoe (6), Manurewa (4), and Mangere-Otahuhu (2). The remaining nine low-responding SA2 areas not in Auckland were from the regions of Northland (4 - all but one from the Far North District), Bay of Plenty (2), Hawke's Bay (2), and Otago (1 - Queenstown Central).

Less extreme (but still worrying), there were an additional 92 SA2 areas with between 60 percent and 70 percent of usual residence one year ago information derived from 2018 Census forms, including:

- 37 from Auckland region, including 27 from local boards in South Auckland (Mangere-Otahuhu, Otara-Papatoetoe, Manurewa)
- 16 from Bay of Plenty region, in the territorial authorities of Rotorua, Tauranga, Whakatane, Opotiki and Kawerau
- 10 from Northland region, including 8 from the Far North District
- 8 from Hawke's Bay region, including 6 from the Hastings District
- 5 in Gisborne region, all in Gisborne City
- only two from the South Island (one from Christchurch City and one from the Queenstown Lakes District)

Most of these areas have high Māori and/or Pacific populations, highlighting the connection between small geographies and low response for Māori (74.3 percent) and Pacific (73.5 percent) in the 2018 Census (see discussion in Section 2.5 of the panel's initial report).

Table 7.1. SA2 areas with < 60 percent of usual residence one year ago information coming from 2018 census forms.

SA2	Percent	Territorial Authority /Local Board	Region
Wiri West	46.9	Manurewa	Auckland
Mount Eden North East	52.3	Albert-Eden/Waitemata	Auckland
Otara Central	54.9	Otara-Papatoetoe	Auckland
Ferguson	55.0	Otara-Papatoetoe	Auckland
Ngapuna	55.5	Rotorua	Bay of Plenty
Ngapuhi	55.6	Far North	Northland
Waima Forest	55.9	Far North	Northland
Otara West	56.5	Otara-Papatoetoe	Auckland
Flaxmere West	56.6	Hastings	Hawke's Bay
Panmure-Glen Innes Industrial	57.0	Orakei/ Maungakiekie-Tamaki	Auckland
Otara South	57.1	Otara-Papatoetoe	Auckland
Harania North	57.3	Mangere-Otahuhu	Auckland
Burbank	58.0	Manurewa	Auckland
Fordlands	58.0	Rotorua	Bay of Plenty
Queenstown Central	58.1	Queenstown-Lakes	Otago
Otangarei	58.5	Whangarei	Northland
Mangere West	58.6	Mangere-Otahuhu	Auckland
Bridge Pa	58.7	Hastings	Hawke's Bay
Otara East	58.9	Otara-Papatoetoe	Auckland
Rowandale West	58.9	Manurewa	Auckland
Hokianga North	58.9	Far North	Northland
Grange	59.1	Otara-Papatoetoe	Auckland
Queen Street	59.2	Waitemata	Auckland
Clendon Park North	59.8	Manurewa	Auckland

From SA2 areas with populations >300 (n=2063).

Only three SA2 listed have populations <1000: Ngapuna, Panmure-Glen Innes Industrial, and Hokianga North.

Based on the most complete Census variable: 'Usual residence one year ago'.

The rest of this section focuses on the impact of low response on the data quality in SA2 areas. Low response resulted in either high levels of 'no information' (missing) for census variables, or 'filling in' data from up to three other sources:

- **2013 Census data**. If an individual was successfully linked to the Integrated Data Infrastructure (IDI), which included the 2013 Census dataset, existing responses to the 2013 Census could be copied across to fill in gaps in some individual variables.
- **Administrative data**. Similarly, if an individual was successfully linked to the IDI, existing administrative (admin) data could be copied across to fill in gaps in some individual variables.
- **Imputation**. The primary form of imputation used was a form of 'nearest neighbour' 'donor' imputation (i.e. find a census respondent who is similar to the census respondent with missing information for a census question, and copy cross the 'donor's response). The specific system used was CANCEIS (CANadian Census Edit and Imputation System), developed by Statistics Canada. This is described in more detail in section 3.2.5 of the panel's initial report and also in Stats NZ (2019a).

While 'filling in' will substantially improve the data compared to leaving large amounts of 'no information', the data will tend to be worse than if higher response rates to the 2018 Census had been achieved.

The quality of data for SA2 areas most affected by low response will be different for different variables, depending on whether and which alternative data sources were used. This is summarised in Table 7.2, which shows the impact on the 24 SA2s from Table 8.1 for four exemplar variables:

- 'Legally registered relationship status', as an exemplar of a variable with high levels of 'no information'; 16.7 percent overall
- 'Ethnicity', as an exemplar of a variable with high use of 2013 Census data; 8.3 percent overall
- 'Total personal income', as an exemplar of a variable with high use of administrative data; 16.5 percent overall
- 'Occupation', as an exemplar of a variable with high use of imputation; 20.3 percent overall

Table 7.2 emphasises that the areas with the lowest levels of information sourced from census forms cannot be characterised very well using the 2018 Census dataset. Relationship status information will not be available for 40-62 percent of adults in these areas.

Table 7.2. Percent of 'no information' and use of alternative data sources for exemplar census variables among SA2 areas with < 60 percent of information derived from 2018 census forms.

SA2	No information: Legally registered relationship status Percent	2013 census data Ethnicity Percent	Admin data Total personal income Percent	CANCEIS imputation Occupation Percent
Wiri West	62.2	23.0	37.8	60.3
Mount Eden North East	52.6	18.7	44.6	40.7
Otara Central	48.0	23.4	43.3	57.5
Ferguson	46.8	22.6	42.7	56.5
Ngapuna	45.1	21.8	44.0	48.1
Ngapuhi	48.7	22.5	41.4	44.8
Waima Forest	45.2	23.0	40.4	43.8
Otara West	46.8	23.6	40.9	54.2
Flaxmere West	42.7	21.2	40.1	52.4
Panmure-Glen Innes Industrial	46.9	13.1	41.8	49.2
Otara South	45.1	23.0	41.7	54.3
Harania North	44.1	22.4	39.4	51.4
Burbank	42.7	21.7	38.2	49.0
Fordlands	42.5	21.8	39.2	46.6
Queenstown Central	45.6	7.4	39.4	49.1
Otangarei	42.1	17.8	40.9	50.0
Mangere West	43.1	23.3	39.2	52.2
Bridge Pa	44.8	18.7	34.4	41.6
Otara East	45.0	21.0	41.7	53.7
Rowandale West	42.4	19.9	39.6	51.7
Hokianga North	41.1	24.2	41.1	44.4
Grange	39.8	18.9	35.4	44.9
Queen Street	43.9	5.1	40.8	49.6
Clendon Park North	40.7	19.2	37.3	46.7

Extensive use of ethnicity data from 2013 Census data will under-estimate ethnic mobility in these areas – which are known to be communities with large Māori and Pacific populations. Interestingly, some small areas with low response (e.g. Queenstown Central; Queen Street, Auckland) relied less on ethnicity data from the 2013 Census. This may be because administrative data was used instead of 2013 Census data to fill gaps in ethnicity data for these small areas, as a high proportion of people in areas such as Queenstown Central and Queen Street may be recent arrivals to New Zealand who would not have completed a 2013 Census questionnaire. A fuller discussion of the impact of using historic data for ethnicity can be found in section 5.5. of our initial report.

Income data will be reliant on administrative sources (IRD tax data) for around 40 percent of adults in most of these poor responding areas. In its assessment of variables report the panel have rated Total Personal Income as High Quality with, nationally, 16.5 percent use of admin data. However, for these poor responding areas the 'data sources and coverage' quality weight of 0.84 for administrative-sourced income data suggests that for areas with more than 32 percent use of admin data the quality of personal income data should be rated as moderate (see information on the quality weight and example of a quality rating calculation here).

Use of imputed (i.e. somebody else's) occupation for around half of the adults in many of the worst-responding areas indicates that it will not be possible to get an indication of the occupation distribution – and how it has changed since 2013 – in these communities. Note that while imputation of occupation is likely to be unbiased in large populations, it is likely to be more volatile in smaller populations (e.g. SA2s), and there is greater risk of a reliance on a small number of 'average' individuals to determine the occupation of non-responders. Note also that the accuracy of imputed occupation (i.e. how often they match actual occupations) is estimated to be only 40 percent for the least granular ('major group') level.

Three final points to note are:

First, a number of other variables have high amounts of 'no information' or use of alternative data sources. So, it is not just data on relationship status, ethnicity, income and occupation that will be affected for SA2's with low response rates to the 2018 Census. Data for many (probably most) variables will be affected and will be low quality in these areas. The extent of 'no information' and use of alternative data sources for selected variables for selected SA2s (as well as by ethnic group and region) is presented in a series of graphs found on 2018 Census external data quality panel: Data sources for key 2018 Census individual variables.

Second, the analyses presented in this section has focussed on a few small areas that have been very adversely affected by poor quality census data. However, a large number of small areas will be at least moderately affected or worse. For example, there are 267 (12.9 percent) SA2 areas with at least 25 percent missing data for 'Legally registered relationship status; 243 (11.8 percent) SA2 areas in which at least 25 percent of data for 'Total personal

income' came from administrative sources; and 436 (21.1 percent) SA2 areas in which at least 25 percent of data for 'Occupation' was imputed.

Third, while this section has focussed on the small areas with the highest proportion of 'no information' or the greatest use of alternative data sources, *variability* between small areas in the extent of no information and use of alternative data sources can also cause problems. For example, estimates of bi-variate associations conducted with small area as the unit of analysis (e.g. associations between small area deprivation and proportion of adults in a small area who are legally married) may be biased by differential amounts of 'no information' or use of alternative data sources. For example, because there is variability across areas in the extent to which 'Legally registered relationship status' had 'no information', and this variability appears to be patterned by deprivation (i.e. poor areas had the most 'no information'), then estimates of area-level associations between deprivation of being 'legally married' may be biased. Users conducting area level associations need to be aware of this potential bias. Figure 7.1 highlights the extent of variability in response at the SA2 level and other levels of geography.

Issues relating to small areas highlighted in this section suggest a number of recommendations for Stats NZ. Two of these appear in Section 3 but they are repeated here because they have particular relevance for data relating to small areas.

R 11. Stats NZ should report on data quality at the small area (SA2) level to support analysts and policy makers with an interest in small area analyses and build quality rating calculations by level 1 ethnicity for every variable relating to individuals into their quality assurance and evaluation plans for the 2023 Census (also in Section 3, pg. 37 and Section 10, pg. 95).

R 12. Stats NZ should systematically investigate the impact of the use of alternative data sources (previous census data, data from a range of admin sources, imputed data) on the quality of data across variables. Analyses should focus not just on whether population distributions are in line with expectations, but also impacts on estimates of inter-censal change, the impact on the sizes of ethnic groups and small areas (e.g.SA2s), and the impact on bivariate associations between variables (also in Section 3, pg. 38).

R 18. Stats NZ should report on 2018 Census SA2 unit source indicator and item source indicator by variable and calculate the metric 1 data quality to support users with an interest in small area analyses, because of the number of small areas experiencing very low coverage and consequently lower data quality across variables.

8 Poor and very poor quality data

In sections 4–8 of this report, reference has been made to data which Stats NZ and/or the panel have rated as **poor** or **very poor.** This section contains a summary of some of the key findings relating to variables that have been rated in these two quality categories.

8.1. Approach to data rated as poor and very poor quality

Stats NZs quality rating framework (see Appendix 3) covers three metrics: data sources and coverage; consistency and coherence; and data quality. Each quality metric is rated separately, and Stats NZ set the overall quality rating as the lowest of the three metric ratings.

It is relatively straightforward to calculate the Metric 1 (data sources and coverage) quality rating for lower levels of classifications or lower levels of geography. The Metric 2 (consistency and coherence) quality measure is perhaps the most important one for users interested in changes over time. However, the calculation of Metrics 2 and 3 for disaggregations is not straightforward.

The most important assessment of the quality of any variable has to be that at the lowest level of the classification for which information is available. This is particularly important for hierarchical classifications including ethnicity, birthplace, religion, language, and occupation.

Stats NZ consider that it is acceptable to generally release data at the appropriate geographic levels and levels of the relevant classification, accompanied by appropriate quality information, that has been rated as having an overall quality rating of **very high, high**, or **moderate.** In its initial report, the panel endorsed this decision.

Stats NZ have already released some data that is of **poor** quality overall (e.g. usual residence one year ago) after it had gone through a process to determine suitability for release. While the panel agrees that it is clearly desirable to release as much data as possible for wider use, this has to be weighed against the risk of uses of the data, perhaps inadvertent, that lead to false conclusions because the data are not of sufficient quality to support the relevant analyses.

The panel believes that caution needs to be exercised by both Stats NZ and users when considering release of data that is rated as being of **poor** quality (at any level of disaggregation of the relevant classification or geography), even where at higher levels of a classification the results meet tests which give quality ratings of **very high, high**, or **moderate**.

Stats NZ have already made the decision that will not be publishing data that is deemed to be of **very poor** quality. One such variable is iwi affiliation which had a response rate of only 71.3 percent. The panel has endorsed this decision in its initial report. Stats NZ are,

however, working with representatives of the Māori community to allow access to the data for analyses to see what use, if any, might be made of the data collected. The panel believes that this is a sensible approach to determine whether any value can be obtained from these data.

In this section we outline characteristics of data rated as poor and very poor quality, options for release of data that is of variable quality at different levels of disaggregation and geography, and give an example where poor quality data from 2018 Census have the potential to mislead users unless they are very careful in the way they take data quality issues into consideration when interpreting their findings.

8.1.1 Data rated as poor quality

The following variables have been rated by Stats NZ as having overall **poor** quality data:

i) Activity limitations;

ii) Individual home ownership;

iii) Number of rooms;

iv) Qualifications: Post-school qualification field of study;

v) Relationship status: Legally registered relationship status, and partnership status in current relationship;

vi) Unpaid activities;

vii) Usual residence one year ago;

viii) Usual residence five years ago;

ix) Years at usual residence.

Data that is rated as poor has one or more of the following:

- a data sources and coverage rating of 75–<90 (see Appendix 3)
- consistency and coherence defined as “Variable data is not consistent overall with expectations across one or more consistency checks. There is an overall difference in the data compared with expectations and benchmarks. Where this difference occurs, this cannot be fully explained through likely real-world change, incorporation of other sources of data, or a change in how the variable has been collected.”
- data quality defined as ”Significant data quality issues emerged during evaluation. Data is considered fit for use but there are limitations on how it can be used and interpreted. There are significant issues with respondent interpretation, coding, and/or classification problems.”

8.1.2 Data rated as very poor quality

There are only two individual/personal variables that have been rated by Stats NZ as having data of overall **very poor** quality. These are iwi affiliation and absentees from the household.

The following household variables are currently rated by Stats NZ as having overall **very poor** quality, although (as noted elsewhere in this report) Stats NZ are reviewing these ratings:

(i) Families and households:

a. Extended family type;

b. Family type;

(ii) Families and households:

a. Household composition;

Data that is rated as very poor has one or more of the following:

- a data sources and coverage rating of <75 (see Appendix 3)
- consistency and coherence defined as "Variable data is highly different from expectations across all consistency checks. There is a large overall difference in the data compared with expectations and benchmarks that cannot be explained through real-world change, incorporation of other sources of data, or change in how the variable has been collected."
- data quality defined as "Major data quality problems exist. Data does not reflect reality due to respondent misinterpretation, coding and/or classification problems."

8.1.3 Options for release of data rated as poor or very poor quality

In releasing the 2018 Census data Stats NZ has to strike a balance between maximising the benefit from the use of the data, whilst minimising the risk of uninformed use that stretches the data further than the data quality can support.

The panel believe that there are a number of options open to Stats NZ when releasing disaggregated data rated as being of **poor** or **very poor** quality (even where the overall quality is **moderate** or higher). The relevance of the options will depend on the uniqueness of the particular census question, the significance of the consequent decisions that will be based on the analysis, and the expertise of those using the information.

The options are:

- Release the disaggregated data as official statistics without additional information – this is not advised

- Release the disaggregated data with well signposted guidance, quality ratings and metadata (including data about the quality of the data) at the level of geography or classification users will use the data – this is recommended
- Release the data through restricted access mechanisms only to users who have been briefed in depth on the quality considerations of their proposed analyses or research – this is recommended for variables rated as poor quality overall
- Do not release the information as official statistics – this is already Stats NZ's approach for variables rated as very poor quality overall. This should not rule out informed investigation of the data (as with iwi data).

An unpublished paper shared with the panel contains the following observations[21]:

- "all poor and very poor quality variables were assessed for suitability for release at a national level in Census totals by topic. All of the poor quality variables were suitable for … release with the addition of a footnote stating the quality rating of the variable and a link to the variable metadata added into the tables when released. Where appropriate, additional information on known issues was added into the 'data quality processes' section of the metadata.
- …this investigation included calculating a metric one rating for each of the variables of interest at each of the main levels of geography (TA, SA2 and SA1). Metric one rates data sources and coverage, and while this is only one aspect of quality, it is a measure that can be calculated at different geographic levels and for cross-tabbed variables. For almost all the poor quality variables metric one was also rated poor quality."

The paper states "Poor quality variables can be released via customised data request once users have been informed of the quality and any specific concerns about the variable….if the information requested includes cross-tabs of more than one poor quality variable, then the request will be shared with Data Quality team so the request can be checked, and quality concerns be communicated to the customer. If the customer is happy to proceed, knowing the quality and issues associated with the variables they have requested, the table they receive will also include footnotes on the quality of the variables, and a link to the information by variable metadata."

The panel do not believe that for variables rated as of poor quality overall it is sufficient for Stats NZ to leave it to the customer to decide whether they are happy to use such data 'knowing the quality and issues'. The panel believe that Stats NZ have a wider duty to protect users from inadvertent use of data rated as of poor quality overall by reviewing proposed projects and proposed use of the data to assess whether the data is fit for such a purpose.

[21] Stats NZ (2019d). Release of data rated as poor and very poor, unpublished report, October 2019.

It might be possible to release some data rated as **poor** or **very poor** at certain levels of aggregation of the relevant classification or geography (e.g. national level data only), and Stats NZ has already done this. However, the panel recommend that any such release should only be done if an assessment of the quality has been carried out at the proposed level of disaggregation and it has been assessed as being of at least **moderate** quality.

8.1.4 The potential to mislead: an example

An example of the potential to mislead is given in section 5 of this report where data relating to Māori who stated they speak te reo in the 2018 Census is examined. The panel states "… the appropriate [Language] quality ratings need to be based on the Metric 1 ratings of categories at Level 4 of the classification. These … range from high to poor. It is the Panel's view that this range best represents the quality of this dataset and that an overall quality rating at Level 1 of the classification does not make sense." Stats NZ and the panel have rated the quality of te reo Māori data as **poor**.

With regard to numbers of Māori speakers of te reo in the 2006, 2013 and 2018 Censuses, it is noted in Section 5 that "…the very significant increase recorded in 2018 goes against recent trends and is almost certainly the result of inflated numbers arising from the use of historic census data and imputation in the 2018 Census, rather than a substantial real-world change in the number of Māori able to speak te reo". Certainly, it ought not to be used in conjunction with te reo data from previous censuses to undertake time series analysis. This would lead to very dubious findings of a significant increase in the number of te reo speakers, and only an imperceptible decline in the share of te reo speakers. Neither of these findings is likely to be reflective of the state of te reo Māori.

8.2 The panel's preference

The panel endorses Stats NZ's release of data rated as **moderate**, **high** or **very high** quality overall. Some of these variables contain disaggregated data (e.g. at small areas, or for small population subgroups) that are of **poor** or **very poor** quality (e.g. see sections 4, 5, and 7).

For data rated as of **moderate** or higher quality overall, the panel's recommendations R 15a and R15b on release of metadata on data sources and quality ratings at small areas and low levels of the relevant classifications strike the right balance between maximising the use of 2018 Census data whilst allowing users to identify and understand any poor or very poor quality data disaggregations – thereby minimising risks of inadvertent misuse.

The panel believe that data rated as being of **poor** quality overall has the potential to mislead and that such data should not be released as official statistics. The panel continue to be of the view that access to data rated as **poor** quality overall should be restricted to accredited individuals working in controlled databases who are able to work closely with Stats NZ to understand the quality characteristics of the data and to determine what value, if any, can be derived from the data.

R 19. Stats NZ should only make data rated as being of **poor** or **very poor** quality overall available where project proposals are considered by Stats NZ on a case-by-case basis, similar to the current procedures for accessing Stats NZ microdata such as the Integrated Data Infrastructure (IDI). This includes data relating to families and households unless Stats NZ determine and advertise a higher quality rating for these data.

9 Outstanding items in the Terms of Reference

As explained in the Introduction, following discussion with Stats NZ it was agreed that of the outstanding items in the Terms of Reference the panel would:

- report on the process for data processing and evaluation reports, and the few most significant problem reports
- report on the data and metadata release schedule as know prior to the panels final report

These are discussed briefly below.

9.1 Data processing and evaluation problem reports

The panel was provided in May 2019 with an overview of the processing and evaluation approach for the 2018 Census, and this has subsequently been described in a paper that Stats NZ have published online.

9.1.1 Data processing and evaluation approach

In broad terms, once initial processing of the data was completed, there was a variable-by-variable assessment of the quality of the resulting data. If data issues were found and there was still time in the processing timetable, then there was the potential to make fixes (either automated or manual) and to rerun the data. This is a normal census process.

During the evaluation process Stats NZ analysed the data and checked the data quality. This included conducting:

- Time series checks – 2006, 2013, 2018
- Checks against expectations – e.g. population estimates and projections
- Checks at lower levels of the classification
- Checks at lower levels of geography (priority 1 variables down to SA2; priority 2 and priority 3 to start at Regional Council level then lower if necessary)
- Consistency checks
- Checks of key cross-tabs, e.g. individual variables by age, sex, ethnicity

Note that:

The **11 priority 1 variables** (e.g. age, sex, Māori descent, ethnicity etc.) were assessed for consistency:

- at level 1 of the classification by territorial authority (TA) compared with the benchmarks
- at the lowest level of classification (if applicable) at a national level.

The remaining priority 2 and 3 variables were assessed for consistency:

- at level 1 of the classification by regional council (RC)
- at the lowest level of classification (if applicable) at a national level.

During the processing and evaluation process 'Problem Reports' might be raised by the analysts assessing a variable. These were considered by the Technical Advisory Group (TAG) which comprised a range of senior census, analyst, and methodology members. They would consider whether the proposed problems were significant enough to warrant change, whether the proposed remedies were likely to be effective, whether there was time and/or resources to make the change, and make a recommendation. TAG met weekly and considered, in total, 210 problem reports.

The problem report and TAG assessment process appeared to the panel to provide a structured and critical assessment of data issues identified and a consistent approach to resolving them where possible/practical.

At the end of this process Stats NZ wrote internal assessments of quality for individual variables – so called 'Warrants of Fitness' (WoFs). These:

- Assigned a quality rating to each variable for each of the three quality metrics – data sources and coverage; consistency and coherence; and data quality (see Appendix 3)
- Provided a breakdown of the data sources, non- response rates, and a data quality rating calculation for Metric 1 (data sources and coverage)
- Contained an outline of the edits, including data edits
- Made recommendations for using the data (including recommendations for the next census)

To complete its evaluation of variables (e.g. in the separate *Assessments of Variables Report*) the panel had access (mostly from July 2019) to 63 separate internal Warrants of Fitness (WoFs).

The amount of information provided in WoFs varied by variable as did the quality of the WoFs, partly dependent on the experience and knowledge of the analysts carrying out the analyses and completing the WoFs.

Relatively few WoFs contained cross-tabulations. The majority of variables were assessed at the National or Regional Council level – few were assessed at TALB or SA2 level. And most assessments were carried out at the highest level of the relevant variable classification.

A number of WoFs contained inconsistencies and what turned out to be mistakes, which were reviewed and corrected by Stats NZ before publication in the panel's initial and final reports. WoFs did not all contain data source information, and none calculated Metric 1 quality ratings by ethnicity or Regional Council.

Whilst the panel was not party to individual problem reports as they were being identified, the broad approach was shared. The information provided to the panel assured us that the quality assessment process appeared robust and that TAG had sensible membership. Example problem reports shared early in the process (May 2019) were reassuringly thorough.

R 20. Stats NZ should continue to undertake WoF assessments for variables in the 2023 Census and should implement a systematic template for WoF content which should include quality ratings for Regional Councils and level 1 ethnicity. Quality control processes should be implemented to ensure assessments are of a consistently high standard.

9.1.2 Problem reports

The panel had hoped to assess examples of the few problem reports which had the most significant impact on the 2018 Census dataset. Whilst the panel were provided with four problem reports, these were not for significant issues – for instance two were for datasets (iwi and same-sex data) that will not be published. The panel was not able to assess examples of significant problem reports.

9.2 Data and metadata release schedule as known in December 2019

9.2.1 Data Release schedule

It goes without saying that the initial 2018 Census data release was delayed by about a year from the original planned publication date. This will clearly have impacted on users plan for use of this data and will have reduced the benefits that could be derived from the data.

The Stats NZ release calendar lists regular releases and includes a table of planned publications. In early December 2019 this included, for the 2018 Census, 18 planned publications from December 2019 to June 2020. See Appendix 2.

9.2.2 Metadata release schedule

The OECD state that "For the ISO standard, metadata is defined as data that defines and describes other data and processes" and statistical metadata as "data about statistical data".

Metadata can be defined as information that is needed to be able to use and interpret statistics. Metadata describe data by giving definitions of populations, objects, variables, the methodology and quality.

- Structural metadata are used to identify, describe or retrieve statistical data, such as dimension names, variable names, dataset technical descriptions

- Reference metadata (sometimes called explanatory metadata) describe the contents and the quality of the statistical data.

On this basis, the 2018 Census metadata should include: descriptions of methods, information on variables and their structure, information on databases etc. as well as information on the contents of datasets, and their quality.

Stats NZ has produced a much wider range of documentation of methods for the 2018 Census than for previous censuses. The panel welcomes this greater range of information which are in the 2018 Census methods and research section of the Stats NZ website.

The DataInfo+ pages for the 2018 Census provide standardised information on each variable, including data sources used, quality rating, and information on any changes to the question or classification.

Information on the data sources used in each variables are provided within the DataInfo+ pages. The relative contribution of data sources to individual variables is summarised here in Table 4. This table dates from July 2019 and may be out of date. It also does **not** provide information on the relative contribution of data sources for dwelling level variables.

As described elsewhere in this report, the relative contributions of different data sources (2018 Census, 2013 Census, admin data, imputation and 'no information') vary significantly by variable, by level of classification for a particular variable (e.g. ethnicity) and by level of geography (e.g. National compared to SA2). As described in the L4 ethnicity and small area statistics section of this report, this impacts on data quality at these levels of disaggregation.

It is therefore important that users have readily available to them metadata on the relative contributions of data sources, and the associated quality rating at least for Metric 1 (data sources and coverage) at these low levels of disaggregation.

10 Towards the 2023 Census

In the lead up to the 2018 Census of Population and Dwellings, Stats NZ frequently observed that the 34th enumeration of New Zealand's population was "a change Census". In their *2018 Census Strategy* (Stats NZ, 2016: 9) reference is made to modernising census operations by "radically alter[ing] the mix of modes used in the current collection model". It was noted in the strategy that a modernised census in 2018 would provide an opportunity "to further test the potential of replacing traditional census collection with administrative data" (Stats NZ, 2016: pg. 6).

It is acknowledged that many of the critical operational tests for the 2018 Census had to be curtailed because of the Kaikoura earthquake and subsequent pressure on the timetable for delivery of the actual enumeration. The absence of a real test of the census enumeration strategy for the 2018 Census cannot be allowed to be repeated during the preparations for the 2023 Census.

One outcome of the poor response to the 2018 Census has been the replacement of missing census responses with data from a range of administrative sources that was not anticipated at the outset of the enumeration. Although the targets for digital responses to the census were achieved at the national level, the very uneven coverage of areas and population groups (especially Māori and Pacific peoples) by the on-line strategy, and the failure to provide alternative ways (paper forms) of enabling responses to the census during the enumeration, meant that reliance on the 2013 Census, administrative data and imputation has been much greater than planned. This has posed a range of challenges for Stats NZ, both in the production of the census data file as well as in assessing the quality of data contained in that file.

There have been two positive outcomes of the protracted post-census process of producing a 2018 Census dataset that seeks to meet the standards of official statistics. The first has been substantial in-house development of methodologies for producing census data files using responses drawn from a mix of sources. The second has been the development of ways of assessing and reporting on the quality of data that comes from a mix of sources for a wide range of variables measuring characteristics of individuals and dwellings, households and families. The 2018 Census has definitely been "a change Census", but not in the way that was envisaged in the *2018 Census Strategy*.

10.1 Some suggestions and recommendations

As an important part of Stats NZ's quality assessment programme for the 2018 Census, the External Data Quality Panel has had the opportunity to reflect on a wide range of data quality related issues. While the panel's Terms of Reference did not specifically request recommendations that might have a bearing on preparations for the 2023 Census, the

following suggestions are offered in the spirit of contributing insights that might be useful as Stats NZ prepares for the 35th enumeration of the country's population.

Some of the recommendations made elsewhere in this report, and in the panel's other reports, are repeated here where they relate specifically to preparations for the 2023 Census.

10.1.1 Rebuilding trust and confidence in the census amongst key stakeholders, especially Māori and Pacific peoples

Delays in producing substantive outputs from the 2018 Census have generated a considerable amount of criticism of the value of the census. A key challenge facing Stats NZ is rebuilding public confidence in the ability of the Department to deliver a modernized census that enumerates a very high share of all population groups in New Zealand and is not compromised by significant under-enumeration, especially of Māori and Pacific peoples.

Regaining the trust and support of the key users of census data and those people who can help support the census, including local government and community groups, will be critical for the 2023 Census. Stats NZ can never know the local circumstances across the whole of New Zealand as well as local authorities and community groups do. If Stats NZ wants to rebuild their trust and enlist them as advocates for the census operation, it will need to give such groups more of a voice in reviewing the operational arrangements. Stats NZ will need to draw on feedback on whether their proposed operational approach (and allocation of resources) will work locally.

Critically important for the 2023 Census is a much better response rate from SA2s that were very poorly enumerated in Northland, Auckland, the Bay of Plenty and Gisborne. Auckland in particular had challenging issues in areas with highly mobile populations, including many immigrants, as well as inner city areas with many access controlled buildings. It could be argued that a 2023 Census cannot be judged a success unless it is a success in Auckland, home to around a third of the country's population.

Regaining the trust and support of Māori will require a demonstrable shift in perception amongst iwi and other Māori organisations that Stats NZ is committed to a meaningful partnership to deliver on its Tiriti o Waitangi obligations that are specified in *Stats NZ's Strategic Intentions 2019–2023.*[22] While comment on operational aspects of the 2018 Census is not part of the panel's brief, the need to improve coverage of the 2023 Census in order to deliver data of acceptable quality on iwi affiliation and use of te reo Māori in

[22] In *Stats NZ's Strategic Intentions 2019-2023* it is recognised that as the leader of the government data system, the Department needs to ensure that decision-makers have the data and information they need to make decisions, and that there is "improved representation of Treaty partners in our data and products". https://www.stats.govt.nz/corporate/stats-nzs-strategic-intentions-201923

particular, is essential. The census is the only nation-wide source of information on these, amongst many other, critically important dimensions of New Zealand society.

In this context, it is appropriate to recall the panel's Term of Reference which specifically state that "data issues that may affect the usefulness of the data for Māori and iwi as Treaty partners" are to be addressed. In Section 2 (Key messages and recommendations from the panel's initial report) the following recommendation relating to data on iwi is included.

R 1. Stats NZ should ensure data collection in future censuses is comprehensive enough to accurately measure iwi affiliation, and the Department should take responsibility, in partnership with iwi, for investigating alternative ways to measure iwi affiliation so that the Census is not the only source of reliable data on these Māori cultural and socio-economic entities (pg. 25)

Building trust and confidence in the 2023 Census will require significant additional investment in the lead up to and the execution of the 2023 Census. To this end the panel recommends that:

R 21. Stats NZ should have an organisational commitment to, and focus on, achieving effective partnership with Māori to develop a census delivery model that will achieve a very high response (>94 percent) from Māori in the 2023 Census.

This focus and investment will be essential to provide the capacity both to plan and deliver on a census that is developed in partnership with Māori. Delivering effectively for Māori in 2023 should also ensure strategies are implemented that will also result in a much better response from Pacific peoples to the census, the other main group that has been compromised by very low response rates in 2018.

To secure a very high response from Pacific residents will require culturally appropriate initiatives that, in themselves will have investment implications. In the panel's view, the 2023 Census cannot be a cost-cutting census – if it is to achieve the sorts of response rates indicated above, there will need to be significantly more investment in this census than there was in the 2018 Census. In order to focus attention and investment on areas and groups that were poorly enumerated in 2018, it would be appropriate to set specific response rate targets for sub-national areas and particular L1 ethnic groups.

R 22. Stats NZ should set response rate targets for particular Territorial Authority and Auckland Local Board areas and ethnic groups that had low response rates in 2018. These targets will drive a focus and resources into areas/groups needed to achieve a much more balanced and complete response profile in 2023 than was achieved in 2018.

Admin data will continue to play a role in the 2023 Census, to fill in (hopefully very much smaller) data gaps. It is therefore also important to meaningfully engage with Māori and Pacific communities about the use of such admin data for them to achieve agreement around social licences and data governance. To this end it is appropriate to repeat a recommendation made earlier in this report:

R 2. Stats NZ should prioritise engagement and investment to ensure:

2a There is genuine partnership with Māori communities, organisations and iwi to develop and implement decision-making and governance mechanisms, to ensure meaningful involvement of Māori in future censuses. This includes Stats NZ actively addressing the acceptability of the extensive use of administrative data in future censuses and issues of social license and Māori data sovereignty specifically for the 2023 Census.

2b There is a real voice for members of all communities, especially Pacific peoples and new migrants, in decision-making on data about them, including the use of admin data in the census. (pg. 25; pg. 38)

10.1.2 Some recommendations relating to preparing for the 2023 Census

A great deal has been learned about the assessment of census data quality during the 18 months the Expert Data Quality Panel has been interacting with Stats NZ on methods for producing and evaluating the data that began to be released to users and the general public from late September 2019. The panel repeats the following advice for Stats NZ as it prepares for the next enumeration of the population in 2023:

R 6. **Stats NZ should** ensure that all collection instruments (paper and online forms), systems and processes are thoroughly reviewed, tested), and made fit-for-purpose for 2023, including an assessment of the equity implications of all collection instruments (paper and online forms), systems and processes. It is essential that issues with the 2018 collection instruments are addressed for the 2023 Census. In this context, Stats NZ should review the extent to which the way the online forms were administered contributed to missing responses in 2018, with a focus on the differential impacts for different population groups, and consider whether changes are needed for the 2023 Census. (pg. 26).

R 11. Stats NZ should report on data quality at the small area (SA2) level to support analysts and policy makers with an interest in small area analyses and build quality rating calculations by level 1 ethnicity for every variable relating to individuals into their quality assurance and evaluation plans for the 2023 Census. (pg. 37; pg. 81)

R 13. Stats NZ should ensure that the methodology to be adopted for the 2023 Census makes explicit provision for high quality measures of intercensal change between 2018 and 2023. The following actions are recommended to avoid further breaks in time-series for census data:

- A high-quality Address Register should form the basis of a management information system for the field and online enumeration to support a quality management strategy that would allow early intervention when things go wrong or not as planned.
- Because the method of field collection of responses will remain a critical part of the next census the enumeration model for 2023 must be developed by building on that

which worked in 2013 and earlier rather than that which failed in 2018 and led to unprecedented quality problems.

- Continue to impute for missing item non-response.
- Assess any new changes to methodology very carefully against whether they would lead to further disruption to the census time-series.
- Review the content of the census forms to ensure the information needed to assess the quality and integrity of data is present. An example is reinstating the question on the total count of people in a household/dwelling.
- With appropriate public consultation and attention to privacy and data justice impacts, make use of access to the relevant government administrative records in advance of the next census to maximise response rates by, amongst other things, targeting field operations, the distribution of paper forms, the number and location of field staff etc.
- Should reliance on administrative records grow, reassess how far statistical surveys (Household Labour Force Survey, New Zealand General Social Survey, and Household Economic Survey), may be better placed to obtain some of the information traditional gathered by the five-yearly census, such as household and family statistics.
- Provide where possible for measures of quality of the enumeration to be an integral part of the census collection and estimation stages as they proceed.
- Consider the role of the PES in 2023, given the changes in census methodology introduced for the 2018 Census.
- Consider producing household and families data by ethnicity. (pg. 40)

In addition to these recommendations relating to census instruments and data sources, the panel has five specific recommendations arising from its extensive work with the various approaches employed by Stats NZ in its assessment of data quality:

R 3. Stats NZ should ensure individual census responses from prisoners are obtained in the 2023 Census. (pg. 26)

R 17. Stats NZ should support a dedicated team for the 2023 Census to undertake post-processing for families and households data, and other complex variables, and not divert this team to other tasks. (pg. 74)

R 20. Stats NZ should continue to undertake WoF assessments for variables in the 2023 Census and should implement a systematic template for WoF content which should include quality ratings for Regional Councils and level 1 ethnicity. Quality control processes should be implemented to ensure WoF assessments are of a consistently high standard. (pg. 90)

R 23. Stats NZ should review the priority ratings that it gives to variables for the 2023 Census so that key statutory duties, including in respect of Māori (e.g. te reo Māori), and

variables used to construct units of analyses (such as households/families), are reflected in higher priority for the associated variables.

R 24. Stats NZ should ensure that external scrutiny, in advance of the 2023 Census, is focussed on methodology, field planning and systems for interaction with the public as well as quality management and quality measures. Ex post reviews, such as that of the External Data Quality Panel, can contribute to improvements in future censuses, but for the 2023 Census it would be more cost effective if some of that external expertise were applied well in advance of the enumeration.

References

Census External Data Quality Panel. (2019) Initial Report of the 2018 Census External Data Quality Panel. Statistics New Zealand: Wellington. Available from: https://www.stats.govt.nz/reports/initial-report-of-the-2018-census-external-data-quality-panel

Jack, M. and Graziadei, C. (2019). *Report of the Independent Review of New Zealand's 2018 Census*. Statistics New Zealand: Wellington. Available from: https://www.stats.govt.nz/reports/report-of-the-independent-review-of-new-zealands-2018-census

Statistics Canada (2017). Statistics Canada's Quality Assurance Framework (Third Edition, release date April 21, 2017). ISBN 978-0-660-08114-4. Available from: https://www150.statcan.gc.ca/n1/pub/12-586-x/12-586-x2017001-eng.pdf

Stats NZ (2005) Ethnicity New Zealand Standard Classification V2.0. Available from: http://aria.stats.govt.nz/aria/#ClassificationView:uri=http://stats.govt.nz/cms/ClassificationVersion/l36xYpbxsRh7IW1p

Stats NZ (2014a). 2013 Post-enumeration survey Available from www.stats.govt.nz.

Stats NZ (2014b). Understanding substitution and imputation in the 2013 Census. Available from www.stats.govt.nz.

Stats NZ (2019a). Data sources, editing, and imputation in the 2018 Census. Retrieved from www.stats.govt.nz.

Stats NZ (2019b). Deriving a quality rating for level 4 ethnicity, unpublished report, November 2019.

Stats NZ (2019c). Quantifying data quality issues for household and family information, unpublished report, 5 December 2019

Stats NZ (2019d). Release of data rated as poor and very poor, unpublished report, October 2019.

Stats NZ (2019e). Stats NZ's Strategic Intentions 2019-2023. Available from: https://www.stats.govt.nz/corporate/stats-nzs-strategic-intentions-201923

Appendix 1 – Executive Summary from the Initial Report

The Executive Summary from the panel's Initial Report is reproduced below. It summarises the panel's key findings relating to: 1) the methodology employed to produce the information on individuals and dwellings contained in the final 2018 Census data file, 2) the quality of the core demographic variables which formed the main part of the Stats NZ data release on 23 September 2019, 3) the ethnicity variable, 4) the Māori descent electoral population, 5) limitations on the quality of the 2018 Census data.

Executive Summary

One in six New Zealand residents did not complete a questionnaire for the 2018 New Zealand Census of Population and Dwellings. This was largely due to operational failures that made it difficult for a significant number of individuals and households to access census questionnaires, and to fulfil their statutory duty to participate.

In response to this unexpectedly high level of non-response, Stats NZ initiated a large-scale census mitigation project that involved the extensive use of alternative government data to fill the gaps. This resulted in a significant delay in the release of results from the Census 2018. While census mitigation has enabled Stats NZ to produce a range of statistical outputs from Census 2018, there are also long-standing key statistics that remain unavailable.

The Census External Data Quality Assurance panel was convened by the Government Statistician in August 2018. The panel provided ongoing advice and guidance to Stats NZ with regard to their mitigation methods and considered the quality of the population statistics that resulted from that work. This report is the first of two to be prepared by the panel.

The timing of this report was determined by Stats NZ's timetable for the first release of Census 2018 statistics. In this report, we assess the methodologies used by Stats NZ to produce the final dataset, as well as the quality of the first release of key statistics. Our work is intended to assist users to make informed judgements about the usability of the data and related statistics produced by Stats NZ and others. We assessed key variables in the first data release using a range of data quality criteria. We were only able to assess quality based on the information provided to us by Stats NZ. Depending on the variable, the level of detail varied significantly. Although the report is technical in nature, we have written it with the broader interests and expectations of the New Zealand public firmly in mind.

The use of new methodologies and alternative government data sources to produce the final Census 2018 dataset marks a significant departure from previous census practice. For example, for the first time, the 2018 Census usual resident population count includes a count of those who did not complete a census questionnaire.

Key findings from the report are summarised below and broadly follow the report structure.

Statistical methods

The panel endorses the statistical approaches used to mitigate non-response.

Stats NZ has undertaken major efforts to augment the census enumeration with data from other sources, using administrative data from the Integrated Data Infrastructure (IDI) as well as data from the 2013 Census. The use of administrative and 2013 Census data has improved the quality of results that we would otherwise have had from the 2018 Census. The addition of administrative records reduced the 2018 Census undercount compared to previous censuses for the population as a whole, and for Māori and Pacific ethnic groups in particular.

However, the unprecedented use of administrative data to augment census data raises questions around ethics, social licence (i.e. tacit approval from the New Zealand public), cultural licence (collective mandate for the trusted use of Māori data), and Māori data sovereignty. While the panel has been advised of the statutory legitimacy of the record linking that has enabled the new methodology to be adopted, we remain unclear about the social and cultural licence to do so. There has not yet been a comprehensive and open public consultation with New Zealanders, including with the groups most affected by the use of alternative data, to gauge the acceptability of the revised census approach.

Key demographic variables

The table below summarises the panel's assessment of quality for the key variables released by Stats NZ in September 2019.

The panel assesses that the linking of government records has improved the coverage and/or accuracy of counts of core demographic elements of a census: age, sex, place of usual residence and ethnicity. Nearly all of the population can be categorised by every one of these four elements.

Variable name	Stats NZ Quality rating	Q/A Panel Quality rating
Age	Very high	Very high – at the national and regional council levels of geography.
Census night address	Moderate	Moderate – at the national and regional council level. There is greater uncertainty at lower levels of geography.
Count of population – census night	Moderate	Moderate – The rating is mostly due to comparability with previous census estimates, particularly for overseas visitors.
Count of population – usually resident	Very high	Very high – at the national and regional council/territorial authority and Auckland Council local board areas (TALB) level There are a small number of meshblocks where NPDs have been allocated to different meshblocks compared to 2013. Users should be careful if they come across such changes, but this will not impact on the quality of data at higher levels of geography.
Dwelling occupancy status	Not rated	Not rated

Ethnicity	High	Moderate – particularly for levels of the ethnicity classification below Level 1
Māori descent – electoral	High	High
Māori descent – output	High	High
Sex	Very high	Very high – down to the SA2 level of geography.
Usual residence address	High	High – at the national and regional council/TALB level.

The Censuses of Population and Dwellings provide a long-term series of demographic, social and economic analyses of the New Zealand population. Because the core demographic elements were measured differently in 2018 than in 2013, measures of change have been distorted by both methodology and response rate variations.

Māori descent electoral

Having reviewed the methodology and examined the sensitivity tests initiated by Stats NZ, the panel is confident that the measure of the Māori descent population for electoral purposes meets the accuracy requirements. We note that the use of administrative records, 2013 Census data, and the estimation methodology has resulted in a larger increase in the Māori descent electoral population from 2013 than occurred between 2006 and 2013.

Ethnicity

Data on ethnicity is a critical census output, and high-quality census data are particularly important for groups that have special status, rights or interests. Measures of ethnicity are critical for planning, development of services and policies, and monitoring for equity. Te Tiriti o Waitangi creates distinct obligations for Māori.

The panel has taken a broader view of the needs of users of ethnicity data than simply the ethnicity variable itself. We rate the quality of ethnicity data as 'moderate', rather than Stats NZ's rating of ethnicity as 'high' quality.

There is significant variability in the quality of ethnicity data by ethnic group. This reflects different patterns of non-response, and the reliance on different alternative data sources, which have different quality characteristics. The quality of ethnicity data generally reduces as the level of ethnic and spatial specificity increases. We find that 2018 Census ethnicity data are of high quality for European; moderate to high quality for Māori; and moderate quality for some Pacific groups. We assessed the ethnicity variable across three dimensions:

Metric 1: Data sources and coverage. Stats NZ provided a high rating for the data sources used to complete the overall ethnicity variable. We assessed ethnicity at Level 2, which is the lowest level for which we had adequate information, and only then at the national level. The metric 1 ratings ranged from very high to moderate at Level 1, and very high to poor at Level 2. The Panel is confident that the quality for this metric will be lower at Levels 3 and 4, with more groups having moderate or below moderate ratings.

Metric 2: Comparability. The Panel's view is that Census 2018 should be treated as a break in the time series, and that comparisons with ethnicity data prior to 2018 should be undertaken with extreme caution, particularly for Māori and Pacific ethnic groups. The

Panel rated this metric as high to moderate, depending on the ethnic group. Stats NZ provided a high rating for the ethnicity variable overall.

Metric 3: Data quality. The data quality metric relates to the data produced from the census forms received and from other data sources. Stats NZ rated the ethnicity variable as high on this metric. The Panel has rated it as moderate to high.

In addition, the Panel notes that for ethnic groups with low census enumeration response rates, there is a high reliance on alternative government data sources for information on characteristics (e.g. education, language, occupation). Where such data are unavailable, or of poorer quality, there will be information gaps. In most instances, the data will produce less reliable analyses of inter-censal change in ethnic group characteristics than earlier censuses.

Limitations on the quality of census 2018 statistics

For Census 2018, there are key limitations on the quality of analyses that have not been recognised as significant in past censuses. These include:

- Household and Families data will generally be of low quality and will not enable comparisons with 2013.
- The analyses of many population groups of importance to government, ethnic communities, local authorities, Māori and service providers will be affected by the lower responses to much of the census questionnaire. Comparative analyses with earlier censuses, or comparisons between groups defined by the categories will be of lower quality. This ranges across the information that can only be obtained from the census questions including about different ethnic groups, occupations, forms of employments status, travel origin and destination analyses, iwi membership, age group studies, activity limitations, religious group membership, languages spoken and smoker characteristics.
- Information that relates people to dwellings will be incomplete as just under ten percent of the population cannot be placed in a specific dwelling, even though they can be located in New Zealand at small area (meshblock) level.
- Stats NZ will not publish iwi data as official statistics due to insufficient data quality. In this regard Stats NZ have not met their Treaty obligations to Māori. Our final report will consider statistical strategies that might be pursued to produce useful estimates for iwi, should iwi wish for that work to be done.

Quality measures

There are many uses to which Census 2018 data will be put. Users of Census 2018 data/statistics will need to explicitly consider the fitness for purpose of the information they wish to use, by consulting the rich array of documentation and quality measures.

The panel will continue to assess the quality of Census 2018 data and will publish a final report covering all the main variables in the census dataset by the end of 2019.

Appendix 2 – Data release schedule

In early December 2019 the Stats NZ release schedule included, for the 2018 Census, 18 planned publications from December 2019 to June 2020.

Month	Working title
Dec 2019	Small areas dataset for 2018 Census
Dec 2019	Selected topics from 2018 Census - NZ.Stat tables
Dec 2019	2018 Census data in the Data Lab - for research
Dec 2019	Applying 2018 Census data to electorate allocation
Feb 2020	Place summaries – 2018 Census data about NZ regions, cities, districts, and small areas
Feb 2020	2018 Census data in the IDI - for research
30 Mar 2020	2018 Post-enumeration survey and 2018-base estimated resident population
30 Mar 2020	Estimated resident population 2018: Data sources and methods – report
Mar 2020	2018 Census data – population and migration
Mar 2020	2018 Census data – housing
Mar 2020	2018 Census data – ethnicity, culture, and identity
May 2020	2018 Census data – Māori ethnicity and descent
May 2020	2018 Census data – travel to work and education
May 2020	2018 Census data – activity limitations and cigarette smoking behaviour
Jun 2020	Ethnic group profiles – 2018 Census data about ethnic groups
Jun 2020	2018 Census data – education and training
Jun 2020	2018 Census data – work, income, and unpaid activities
Jun 2020	2018 Census data – population (other topics)

Appendix 3 – Stats NZ data quality assurance definitions for the 2018 Census

Stats NZ's Data quality assurance for 2018 Census outlines the quality assurance framework and quality rating scale used by Stats NZ to assess the quality of data from the 2018 Census to determine whether it is fit for purpose and suitable for release. The following are excerpts from this report.

The 2018 quality rating scale is made up of three metrics:

- metric 1 – data sources and coverage
- metric 2 – consistency and coherence
- metric 3 – data quality.

An overall variable rating was assigned to each by taking the lowest score that variable has received from the three metrics, across the range.

Metric 1: Data sources and coverage

This metric calculates a score by rating the overall quality of the data sources used for a census output of a variable. This aims to:

- give customers clarity around what sources have gone into the combined output for a census variable
- show how the rating given to a source (which is based on the quality of the source) will then impact the total score (and quality) of a variable
- calculate an approximation of 'missingness' and uncertainty of output values for a census variable.

To calculate a score for a variable, each source that contributes to the output for that variable is rated and multiplied by the proportion it contributes to the total output.

The rating for a valid census response is defined as 1.00. Ratings for other sources are the best estimates available of their quality relative to a census response.

We calculated the ratings for admin data sources by comparing the 2018 Census received responses with the data from admin source, with a value being derived from the match rate between the two sources.

Bands for data sources and coverage ratings

The bands used for metric 1 are similar to those used in the 2013 Census metric for non-response:

Very high	0.98–1.00
High	0.95–< 0.98
Moderate	0.90–< 0.95

Poor	0.75–< 0.90
Very poor	0.00–< 0.75

Metric 2: Consistency and coherence

Stats NZ rated the level of consistency and coherence in the data on:

- comparability with the expected trends
- comparability with other sources
- contribution of other sources to the census data for this variable.

The ratings account for changes occurring for variables in the 2018 Census as a whole, including the use of admin data and, in some cases, the change in question or concept. In some cases, 2018 Census data may be moving away from expected time series trends, due to methodological changes that have brought the data closer to the 'real world' situation, by addressing historic issues, or biases within census coverage.

For new or changed variables where there is no previous census data for comparison, we used other data sources and expectation reports as the primary source of comparison. These may only be comparable at a national level.

Explainable change (see 'moderate' ratings below) could be the result of real-world change, incorporation of other sources of data, or a change in how the variable has been collected.

Priority 1 variables were assessed for consistency:

- at level 1 of the classification by territorial authority (TA) compared with the benchmarks
- at the lowest level of classification (if applicable) at a national level.

Priority 2 and 3 variables were assessed for consistency:

- at level 1 of the classification by regional council (RC)
- at the lowest level of classification (if applicable) at a national level.

Five detailed descriptions guided their assessment and categorisation of variables for this metric:

Very high	Variable data is highly consistent with expectations across all consistency checks.
High	Variable data is consistent with expectations across nearly all consistency checks, with some minor variation from expectations or benchmarks that makes sense due to real-world change, incorporation of other sources of data, or a change in how the variable has been collected.
Moderate	Variable data is mostly consistent with expectations across consistency checks. There is an overall difference in the data compared with expectations and benchmarks that can be explained through a combination of real-world change, incorporation of other sources of data, or a change in how the variable has been collected.

Poor	Variable data is not consistent overall with expectations across one or more consistency checks. There is an overall difference in the data compared with expectations and benchmarks. Where this difference occurs, this cannot be fully explained through likely real-world change, incorporation of other sources of data, or a change in how the variable has been collected.
Very poor	Variable data is highly different from expectations across all consistency checks. There is a large overall difference in the data compared with expectations and benchmarks that cannot be explained through real-world change, incorporation of other sources of data, or change in how the variable has been collected.

Metric 3: Data quality

This metric relates to the data produced from the census forms received and from other data sources. This includes aspects such as coding, level of detail/classification, accuracy of responses, and any other specific quality issues that may have been identified in problem reports. Stats NZ used the same overall approach that was used in 2013 for this metric. The ratings are:

Very high	Data has no data quality issues that have an observable effect on the data. The quality of coding is very high. Other data sources used do not create any quality impacts for this variable. Any issues with the variable appear in a very low number of cases (typically less than a hundred).
High	Data has only minor data quality issues. The quality of coding and responses within classification categories is high. Any impact of other data sources used is minor. Any issues with the variable appear in a low number of cases (typically in the low hundreds).
Moderate	Data has various data quality issues involving several categories or aspects of the data, or an entire level of a hierarchical classification. The data quality issues could include problems with the classification or coding of data, such as vague responses resulting in coding issues, or responses that cannot be coded to a specific (non-residual) category, thereby reducing the amount of useful, meaningful data available for analysis. The use of other data sources may be contributing to these issues.
Poor	Significant data quality issues emerged during evaluation. Data is considered fit for use but there are limitations on how it can be used and interpreted. There are significant issues with respondent interpretation, coding, and/or classification problems.
Very poor	Major data quality problems exist. Data does not reflect reality due to respondent misinterpretation, coding and/or classification problems.

Appendix 4 – Families and Households variables

There are 29 variables within the Families and Households suite of questions and derived variables. They are:

- Number of People in Family
- Number of Children in Family;
- Number of Usual Residents in Household
- Number of Usual Residents Aged 15 and Over in Household
- Number of Usual Residents Aged Under 15 in Household
- Identification of Individual's Family Nucleus
- Individual's Role in Family Nucleus
- Dependent Child Under 18
- Dependent Young Person Indicator
- Number of Dependent Children in Family
- Number of Adult Children in Family
- Age of Youngest Child in Family
- Age of Youngest Dependent Child in Family
- Family Type
- Family Type with Type of Couple
- Family Type by Number of Children
- Extended Family Type
- Family Type by Child Dependency Status
- Household Composition
- Number of Dependent Children in Household
- Age of Youngest Child in Household
- Age of Youngest Dependent Child in Household
- Household Composition by Child Dependency Status
- Type of Couple
- Age of Male Partner in Opposite-Sex Couple
- Age of Female Partner in Opposite-Sex Couple
- Age of Older Partner in Same-Sex Couple
- Age of Younger Partner in Same-Sex Couple
- Sex of Sole Parent

Appendix 5 – Links to Stats NZ DataInfo+ pages

Click on the variable name below to go to the relevant Stats NZ Data.Info page.

Variable name
Absentees
Access to telecommunication systems
Activity limitations
Age
Birthplace
Census night address
Cigarette smoking behaviour
Count of dwellings
Census night population count
Census usually resident population count
Dwelling occupancy status[1]
Dwelling type
Educational institution address
Ethnicity
Families and households: extended family type
Families and households: family type
Families and households: household composition
Hours worked in employment per week

Housing quality: access to basic amenities
Housing quality: dwelling dampness indicator
Housing quality: dwelling mould indicator
Individual home ownership
Industry
Iwi
Languages spoken
Main means of travel to education
Main means of travel to work
Main types of heating and fuel types used to heat dwellings
Māori descent – output
Māori descent – electoral
Number of bedrooms
Number of children born
Number of motor vehicles
Number of rooms
Occupation
Qualifications: highest qualification
Qualifications: highest secondary school qualification
Qualifications: post-school qualification level of attainment
Qualifications: post-school qualification field of study
Relationship status: Legally registered relationship status, and partnership status in current relationship

Religious affiliation
Sector of landlord
Sector of ownership
Sex
Sources of personal income
Status in employment
Study participation
Tenure of household
Total personal income
Unpaid activities
Usual residence one year ago
Usual residence five years ago
Usual residence address
Weekly rent paid by household
Work and labour force status
Workplace address
Years at usual residence
Years since arrival in New Zealand

Appendix 6 – Glossary

2013 Census	Census of Population and Dwellings undertaken on 5 March 2013. For some 2018 Census topics, responses from the 2013 Census were used to fill in missing data.
Absentee	A person who is identified on the census dwelling form as usually living in a particular dwelling but who did not complete a census individual form at that dwelling because they were elsewhere in New Zealand or overseas at the time of the census.
Administrative (admin) data	Data collected by government or other organisations for non-statistical reasons, such as births, tax, health, and education records. These are typically records describing events or interactions with government agencies and have been obtained in the course of some statutory obligation or service provided by a government agency.
Administrative (admin) enumeration	The use of administrative data to add people to the usually resident census population when a census response has not been received.
Auckland Local Board	Statutory community-level governance districts within Auckland Council. There are 21 local boards: Albert-Eden, Devonport-Takapuna, Franklin, Great Barrier, Henderson-Massey, Hibiscus and Bays, Howick, Kaipātiki, Māngere-Ōtāhuhu, Manurewa, Maungakiekie-Tāmaki, Ōrākei, Ōtara-Papatoetoe, Papakura, Puketāpapa, Rodney, Upper Harbour, Waiheke, Waitākere Ranges, Waitematā, Whau.
CANCEIS	Canadian Census Edit and Imputation System. A method for 'imputing' (filling-in) data for missing responses/respondents. Used by a number of national statistical institutes for census imputation.
Census Post-enumeration Survey (PES)	A household sample survey run soon after census day, to measure coverage achieved in the census. The census undercount and overcount as measured by the PES are used to produce the official census coverage and response rates. The 2018 PES went to 15,000 households throughout New Zealand during April–July 2018.
Census usual resident population count	A count of all people who usually live in New Zealand and were present somewhere in New Zealand on census night.
Classification	System of categorising the responses to questions that are not values. Many census variables use standard classifications systems (e.g. country of birth, ethnicity, occupation). The

	classifications used for census variables may differ from the classifications used for the equivalent administrative variable.
Coverage rate	The census usual resident population count expressed as a percentage of the New Zealand estimated resident population (ERP)
Donor imputation	Method of imputation which uses data from similar individuals or households to 'impute' (fill-in) data for missing responses/respondents
Dwelling	A building or structure using for habitation, e.g. houses, motels, hotels, prisons, rest-homes
Dwelling form	Census questionnaire with information on the dwelling. For paper forms this includes a listing of people within the dwelling and their relationship to the person completing the dwelling form. See household summary form.
Electorate	Geographic area contributing one seat to the New Zealand parliament. Under the Electoral Act 1993, the number of electorates in the South Island is fixed at 16, and the South Island quota (the South Island General Electoral Population divided by 16) determines the number of General electorates in the North Island and Māori electorates.
Estimated Resident Population (ERP)	An estimate of all people who usually live in New Zealand at a given date (e.g. Census day). Population estimates are produced using data from the most recent Census of Population and Dwellings, updated for estimates of the components of demographic change (births, deaths and net migration) since that last census. This is not the same as the Census Usual Resident Population Count which typically slightly undercounts the New Zealand ERP. Population estimates based on the ERP include adjustments for net census undercount and residents temporarily overseas. See coverage rate; undercount.
Ethnicity	A measure of cultural affiliation. It is not a measure of race, ancestry, nationality, or citizenship. Ethnicity is self-perceived and people can belong to more than one ethnic group. Stats NZ uses a hierarchical classification system for ethnicity, with • 6 categories at 'Level 1': European; Māori; Pacific; Asian; Middle Eastern, Latin American and African (MELAA); Other; • 21 categories at 'Level 2', including New Zealand European; Samoan; Chinese; Middle Eastern; • 36 categories at 'Level 3', including South Slav; Filipino; • 180 categories at 'Level 4', including Serbian; Tahitian; Malay; Kenyan; Indigenous American.

Family	A couple, with or without child(ren), or one parent with child(ren), usually living together in a household. Related people, such as siblings, who are not in a couple or parent-child relationship, are therefore excluded from this definition.
Household	One person who usually resides alone, or two or more people who usually reside together and share facilities (such as eating facilities, cooking facilities, bathroom and toilet facilities, and a living area), in a private dwelling.
Household summary form	Online census form containing a listing of people within the household and their relationship to the person completing the household summary form.
Integrated Data Infrastructure (IDI)	A large database maintained by Stats NZ. It contains de-identified data about people and households sourced from government agencies (i.e. administrative data), 2013 Census, Stats NZ surveys, and non-government organisations (NGOs). Data from different sources are linked together, typically at the individual (person) level.
IDI-ERP	The New Zealand Estimated Resident Population (ERP) derived from linked records in the IDI.
IDI-ERP_Sure	The IDI-ERP restricted to exclude people who are less likely to have been New Zealand residents at the time of the census (i.e. the IDI-ERP_Sure includes only people who are 'sure' to belong to the New Zealand resident population at the time of the census.
Imputation	The process of replacing missing data with estimated values through statistical methods. For the 2018 Census, the method for estimating values was nearest-neighbour imputation methodology (NIM), which finds similar respondents with a response to the variable in question. The processing system then finds the closest match to the respondent with missing or unidentifiable data and imputes the donor respondent's response. See CANCEIS.
Individual form (or questionnaire)	Census questionnaire to be completed by each person in a dwelling. This includes questions about ethnicity, education, income, etc. pertaining to the individual.
Individual response	Where an individual form was received for a respondent by Stats NZ.
IRD	Inland Revenue Department
Iwi	Māori tribe or extended kinship group, often descended from a common ancestor and/or associated with a distinct territory.
Level 1 (2,3,4) Ethnicity	See Ethnicity

Māori descent electoral count	The number of people in the census usual resident population determined to be of Māori descent for electoral purposes. The Māori descent census question asks, “Are you descended from a Māori (that is, did you have a Māori birth parents, grandparent, or great grandparent, etc.)?” For electoral purposes only “Yes” and “No” answers are considered (i.e. “Don’t know”, not stated and unidentifiable responses are not considered). For 2018, data from other sources were used when a response other than “Yes” or “No” was given. Cf. Māori descent output.
Māori descent output (variable)	Census variable that assesses the Māori descent population in New Zealand. For 2018, valid responses were “Yes”, “No”, and “Don’t know”. For 2018, data from other sources were used when a response other than “Yes”, “No” or “Don’t know” was given. Cf. Māori descent electoral count.
MELAA	Middle Eastern, Latin American and African: A grouping at Level 1 of the ethnicity classification
Meshblock	The smallest geographic units for which statistical data are reported. These vary in size from part of a city block to a large area of rural land, with an ideal size range of 30–60 dwellings (around 60–120 residents).
No information	Where data could not be sourced (from a response to the 2018 Census, 2013 Census data, administrative data or from statistical imputation) for units in the subject population of a variable. For example, where the number of children born could not be sourced for a female in the census usually resident population aged 15 years and over.
Non-private dwelling (NPD)	A dwelling providing communal or transitory type accommodation (e.g. hotel, campground, prison, defence barrack, rest home, university hall of residence).
Not elsewhere classified (nec)	A residual category for responses that have no appropriate category, because they are infrequent or unanticipated. These categories never appear within classifications as stand-alone descriptors, but are combined with descriptors, often taken from a higher level in the classification. For example, for Qualifications, BSc Environmental Biology would go to Biological Sciences nec.
Not elsewhere included (nei)	Used in some outputs for a combination of residuals, such as ‘not stated’, ‘response outside scope’, ‘response unidentifiable’, ‘refused to answer’, and ‘don’t know’. This item should have a footnote indicating its composition.
Not further defined (nfd)	A residual category used in hierarchical classifications for responses containing insufficient detail to be classified to the

	most detailed level of a classification, but which can be classified to a less detailed category further up the hierarchy.
Not stated	A category used when a person gave no response to a question relevant to them or when there was no alternative data source for that information, such as a 2013 Census response, administrative data or statistical imputation
Partial response	Where an individual was listed on a dwelling form (paper) or household summary form (online) but no individual form was received by Stats NZ.
Private dwelling	A dwelling accommodating one or more people who usually live independently within the community (e.g. a house or flat)
Refused to answer	A category used only when it is known that a person has purposefully chosen not to respond to the question.
Region	The first tier of local government. There are 16 regions: Northland, Auckland, Waikato, Bay of Plenty, Gisborne, Hawke's Bay, Taranaki, Manawatu-Wanganui, Wellington, Tasman, Nelson, Marlborough, West Coast, Canterbury, Otago, Southland
Response	Completion of some or all items on a census form. In line with international practice, a census 'response' in 2018 was achieved when the minimum information to count a person was received. Thus, the listing of an individual on a dwelling form was considered a response, even if no individual form was received for that individual.
Response outside scope	A category applied if the meaning and intent of the response are clear ('positively identified') but clearly fall outside the scope of the classification/topic as defined.
Response rate	Number of census responses expressed as a percentage of the New Zealand Estimated Resident Population (ERP). In the report, 'total response rate' considers both individual and partial responses when calculating response rate (see 'response' above); 'individual response rate' considers just individual responses when calculating response rate; 'partial response rate' considers just partial responses when calculating response rate
Response unidentifiable	A response given that is: • illegible • unclear regarding its meaning or intent. This most commonly occurs when the response being classified contains insufficient detail, is ambiguous, vague or contradictory (for example, when the tick boxes 'yes' and 'no' have both been ticked)

	• clear and seemingly within the scope of the classification, yet it cannot be coded as a suitable existing option in the classification or code file (such as 'not elsewhere classified' or 'not further defined').
SA1	Statistical Area 1: A geographic unit built by joining meshblocks, with an ideal size range of 100–200 residents, and a maximum population of about 500.
SA2	Statistical Area 2: A geographic unit which aims to reflect communities that interact together socially and economically. In major urban areas, an SA2 often approximates a single suburb, generally with a population of 2,000–4,000 residents. SA2s in district council areas generally have a population of 1,000–3,000 residents. In rural areas, SA2s may have fewer than 1,000 residents if they cover large areas that have sparse populations.
Social Licence	Permission or mandate or societal acceptance that an agent may act or behave in a certain way. E.g. the permission for Stats NZ to make decisions about management and use of the public's data. Recognises the distinction between statutory legitimacy and political legitimacy.
Statistical geography	Classification of places in New Zealand into different levels of geography. The current classification system (SSGA18) provides a range of geographic units from 'meshblock', the smallest geographic unit (roughly 30-60 dwellings) to 'region', the largest geographic unit and top tier of Local Government (e.g. Northland region, Auckland region).
Territorial Authority (TA)	The second tier of local government, below regions. There are 67 territorial authorities: 13 city councils (**Auckland**, Hamilton City, Tauranga City, Napier City, Palmerston North City, Porirua City, Upper Hutt City, Lower Hutt City, Wellington City, **Nelson City**, Christchurch City, Dunedin City, Invercargill City); 53 district councils (Far North, Whangarei, Kaipara, Thames-Coromandel, Hauraki, Waikato, Matamata-Piako, Waipa, Otorohanga, South Waikato, Waitomo, Taupo, Western Bay of Plenty, Rotorua, Whakatane, Kawerau, Opotiki, **Gisborne**, Wairoa, Hastings, Central Hawke's Bay, New Plymouth, Stratford, South Taranaki, Ruapehu, Whanganui, Rangitikei, Manawatu, Tararua, Horowhenua, Kapiti Coast, Masterton, Carterton, South Wairarapa, **Tasman, Marlborough**, Buller, Grey, Westland, Kaikoura, Hurunui, Waimakariri, Selwyn, Ashburton, Timaru, Mackenzie, Waimate, Waitaki, Central Otago, Queenstown-Lakes, Clutha, Southland, Gore); and the **Chatham Islands Council**.

	Six territorial authorities (bolded) are also regions and therefore Unitary Councils.
Undercount (net)	Extent to which the census usual resident population count undercounts the New Zealand estimated resident population. Taken as the difference between 'gross undercount' (the number of people who were supposed to be counted by the census, but who were not counted) and 'gross overcount' (the number of people counted more than once, and people who were counted in the census who should not have been counted. Expressed as a percentage (e.g. 2 percent undercount).