Initial 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 this report. We also thank members of the Census Independent Review Panel who met twice with us, and an independent reviewer who provided useful comments on this report.

We were supported throughout the review by a small secretariat provided by Stats NZ. They have been vital to the conduct of the review and we thank them for their efforts.

Introduction to the 2018 Census External Data Quality Panel

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.

As set out in the Terms of Reference, the objectives of the panel are 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.

Of the nine focus areas listed in the Terms of Reference, the following five are covered in this report:

- 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 will be covered in our second report:

- 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.

Originally, the panel was to convene from August 2018 until April 2019, and to produce a report upon 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 necessitated deferral of reporting. The panel will convene through to November 2019.

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 (also refer to section 6).

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 - See section 5
Māori descent – electoral	High	High – See section 4
Māori descent – output	High	High - See section 4
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 intercensal 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 eight 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.

Table of Contents

1 Introduction

1.1 Background

The New Zealand Census of Population and Dwellings provides an official count of how many people and dwellings there are in the country at a set point in time. It also provides detailed social, cultural and socio-economic information about the total New Zealand population and key groups in the population. The New Zealand population is counted by their age, sex, and ethnicity, and in their regions and communities. Censuses in New Zealand have been undertaken since 1851, and have been conducted every five years since 1881, with exceptions being: during the Great Depression (1931); during the Second World War (1941); in 2011 as a result of the Christchurch earthquakes; and, when a Census was held in 1945 rather than 1946. The most recent census was undertaken on March 6, 2018.

The 2018 Census failed to obtain information from many New Zealanders. This has raised concerns about the quality and usability of census data.

This report is the first of two reports produced by the 2018 Census External Data Quality Panel ('the panel'). It describes and evaluates the methodological efforts undertaken by Stats NZ to improve the quality and usability of the 2018 Census data.

This first report, timed to coincide with the first release of data by Stats NZ on 23 September 2019, describes and evaluates the methods that Stats NZ has employed to generate a 2018 Census data file. This report focuses on the methods used to generate estimates of the total population, methods used to estimate the Māori descent population for electoral purposes, and an assessment of the validity of electoral population counts and electorate numbers (both Māori and General). It also comments on the quality of some of the key data released as part of the first release, including ethnicity data.

The second report, scheduled for December 2019, will give a more in-depth assessment of the quality of each of the 2018 Census data variables that Stats NZ will release.

The panel hopes that the views and advice contained in these reports will help both casual and experienced users of census data to make informed choices as to how they can, and cannot, use 2018 Census data, and questions they should ask themselves when using 2018 Census data.

1.2 This report

This report is in seven sections.

This section (**Section 1**) explains the purpose of the report and describes the role the panel has had in evaluating the methodological approaches undertaken by Stats NZ to improve data quality.

Section 2 discusses the importance of the census. We reflect on key uses and value of the census, and the impact of the change in census operations in 2018 from previous censuses.

Response rates for the 2018 Census are presented, which highlight some of the data quality problems for which methodological solutions are required.

Section 3 describes and evaluates the methodology employed by Stats NZ to generate the 2018 census data file. It describes and evaluates the methods used to estimate the New Zealand population, the methods used to add individuals not counted by the census, and the methods used to impute characteristics for individuals for whom data on characteristics were not available. The statistical, legal, and ethical appropriateness of the methodological approaches undertaken are also discussed.

Section 4 comments on the Māori descent and sub-national populations that have been generated for the Representation Commission. Implications of the different collection methods used in Census 2018 compared to previous censuses for the Māori descent variable are discussed. The impact of the Māori descent electoral variable on number of Māori and general electorates, and on boundaries for electorates is also discussed.

Section 5 comments of the quality of ethnicity data, with a particular focus on ethnicity at the Level 1 classification, and on ethnic data for Māori and Pacific populations.

Section 6 comments on the quality of the Census data in the first release of data on 23 September 2018, specifically on the usual resident count and census night count of the total New Zealand population, the regional and sub-regional counts of the population by age and sex, plus the count of dwellings in New Zealand.

Section 7 concludes the report and offers some guidelines about how to judge when Census statistics can be trusted and when Census statistics must be used with caution.

2 Importance of the Census

Census taking in New Zealand is part of an endeavour led by the United Nations (UN), with censuses carried out in virtually all UN nations at least every ten years[1]. New Zealand, along, with Australia, Canada, and Ireland, carries out a census every five years.

2.1 Uses of the Census

The census provides unique information on very small areas (the meshblock is the smallest area on which results are published in New Zealand; meshblocks typically contain 60–120 individuals) and for small population subgroups (e.g., the Tuvaluan community).

The census has always sought to collect information on:

- The number of people in New Zealand
- Their characteristics (e.g., ethnic identity, occupation)
- Their relationships to others in their dwelling
- Information on the dwelling (e.g., number of rooms).

By collecting complete (or near complete) information at the same time on both housing and people, previous censuses have been able to produce integrated outputs about:

- The different ethnic communities that make up the New Zealand population (numbers and their characteristics)
- Households and families
- Housing/dwellings
- Occupations, workplaces, and employment.

Past censuses have also provided analyses of:

- Family type (e.g., single parent families; same sex couples) and household (e.g., households containing multi-generational families), cross tabulated by characteristics of the family/household (ethnicity, for instance).
- Overcrowding and unmet housing need (by combining information on people and the dwellings in which they live).

As in other countries, census data is used in New Zealand for the following broad purposes:

- **Electorates and electoral boundaries**. While the census is important for creating electoral boundaries in most countries[2], in New Zealand the Statistics and Electoral legislation is more prescriptive than seen in other places, particularly in the calculation of Māori and General electoral seat numbers and electoral populations.
- **Central and local government policy making and monitoring**, especially for Māori and population groups that are exposed to greater risks and/or structurally

[1] Of the world's 241 countries and territories, only five percent did not undertake a census in the 2010 census round (2005-2014) (Kukutai, et al., 2015).

[2] For example, in the United States the results of the decennial census determines the apportionment of seats in the House of Representatives for the subsequent decade.

marginalised. There is a clear drive across Government for policy initiatives to be evidence-based and to achieve desired outcomes cost effectively. The Child Poverty and Wellbeing strategies of the government are examples of this. In many cases this requires robust identification and analysis of smaller groups within the overall population.

- **Allocating resources from central government to local areas** (e.g., for health[3] and education) based on the relative needs of the local population. For resource allocation purposes, it is crucial that population counts (both total counts and by key characteristics) are accurate, consistent, and comparable over the area that the resources are to be allocated. This is becoming increasingly important as Government seeks to maximise the value for money from expenditure by better targeting.
- **Investment planning**. For both government and the private sector there are significant capital investments where timing, location and scale are affected by the geographical patterns of movement, demographics, and anticipated levels of population. Retailers make extensive use of census data in store location decisions.
- **Academic and market research**. The ability to produce multivariate statistics for small areas is vital for many research uses. Basic population counts and counts by characteristic are also required.
- **Statistical benchmarks** (e.g., intercensal population estimates which use rebased census data and adjust for census undercount, births, deaths, residents temporarily overseas on census night, net migration etc.). Census data are used to improve the quality of many other statistics, which may be used for the above categories. Many of Stats NZ's statistics are benchmarked or grossed up using census data. As such the census is integral to the operation of sample surveys and to the appropriate use of administrative data.
- **A data frame to select samples for social surveys**, such as the Household Labour Force Survey, the Household Economic Survey, the NZ General Social Survey, and the Māori Social Survey Te Kupenga.

[3] For example, the Population-Based Funding Formula (PBFF) used to allocate District Health Board funding requires a set of DHB populations, a set of service-based cost weights for each age, sex, ethnicity and socio-economic groups as defined by NZ index of Deprivation (NZ Dep) quintile group, plus rural and overseas and refugee adjustments. The population projections have as their base the estimated resident DHB populations that, in turn are based on the census usual resident count adjusted for census undercount, census ethnicity non-response, residents temporarily overseas on census night, births, deaths and net migration.

2.2 Obligations and legal requirements

The census is also relevant to obligations that Government has in relation to domestic legislation and Te Tiriti o Waitangi, as well as in terms of international conventions and declarations. These are discussed briefly below.

2.2.1 Te Tiriti o Waitangi obligations to Māori

As a Crown agency, Stats NZ has obligations under Te Tiriti o Waitangi relating to the production of official statistics. Stats NZ have acknowledged the 'special relationship' that the agency has with Māori in relation to Te Tiriti o Waitangi, and the responsibilities that flow from this relationship with regards to official information (Stats NZ, 2018).

In their report *Enduring census information requirements for and about Māori,* Gleisner, Downey and McNally note that Stats NZ identify responsibilities they have to support Māori well-being and development 'on their own terms' and 'to have equity as citizens' (2015, p.7).

Stats NZ's strategic priorities for the period 2017/18 – 2020/21 also note a number of obligations to Māori in relation to stewardship of data, and the need for active partnerships (Stats NZ, 2018b)[4]. Iwi census data have been used in Treaty settlements by enabling the history of population change to be taken account of in settlement arrangements (see, for example, Treaty of Waitangi Fisheries Commission, 2003).

2.2.2 Obligations under international conventions and declarations

Various other conventions, including UN conventions, outline rights and obligations that have relevance to government agencies, including Stats NZ.

The United Nations Declaration on the Rights of Indigenous Peoples (UNDRIP)[5] reaffirms the rights of Māori, as indigenous peoples, to be fully included in decision-making on matters that may impact them. Specifically, Article 19 notes that:

> *States shall consult and cooperate in good faith with the indigenous peoples concerned through their own representative institutions in order to obtain their free, prior and informed consent before adopting and implementing legislative or administrative measures that may affect them. (UNDRIP, 2007).*

This principle is important in the context of Census 2018 and Stats NZ's decision to draw extensively on alternative government data sources without a formal mechanism to engage with Māori as Treaty partners. Māori data sovereignty[6] and indigenous data sovereignty

[4] The 2019 Stats NZ document 'Empowering agencies to use data more effectively' states that an enhanced approach to data requires: "Treaty partnership that is fully enabled. Māori co-design and engage with the data system. Data is accessible for Māori". Retrieved from: stats.govt.nz/about-us/data-leadership#achieve

[5] https://www.un.org/development/desa/indigenouspeoples/wp-content/uploads/sites/19/2018/11/UNDRIP_E_web.pdf

[6] Māori data sovereignty principles published by the Māori data sovereignty network Te Mana Raraunga can be found here: https://www.temanararaunga.maori.nz/new-page-2

principles also highlight the importance of free, prior, and informed consent, and of indigenous peoples having rights to govern data that are about or from them[7].

2.2.3 Equity obligations

The UN, in their work and recommendations on official statistics, note the importance of statistics that provide information about specific population groups. The *Statistical Commission* report states that

> *... population and housing censuses are designed to generate valuable statistics and indicators for assessing the situation of various special population groups, such as women, children, youth, the elderly, persons with disabilities, migrants, refugees and stateless persons, and changes therein (UN Economic and Social Council, 2015).*

Similarly, in outlining the 'essential roles' of the census the UN *Principles and Recommendations for Population and Housing Censuses, Revision 3* states that:

> *The basic feature of the census is to generate statistics on small areas and small population groups with no or minimum sampling errors. (United Nations 2017, p. 2).*

Thus, there is an implicit assumption that the census should provide information equitably, that is, to the same level of depth and detail for specific population groups. The granularity of ethnicity and spatial data collected in the census makes it a unique data source that can provide valuable insights into the nature and causes of ethnic, socio-economic, and spatial inequities, and changes in inequities over time. It is also the only data source where data on iwi (tribe), an institutional form unique to Te Ao Māori, are collected.

2.2.4 Legal obligations under the Statistics Act 1975

Under the Statistics Act 1975, there are some specific types of information that are legally required to be collected by Stats NZ as part of the census. Section 24 notes:

> (1) At every census of population and dwellings particulars relating to all of the following matters shall be obtained from every occupier or person in charge of a dwelling:
>
> (a) the name and address, sex, age, and ethnic origin of every occupant of the dwelling
>
> (b) particulars of the dwelling as to location, number of rooms, ownership, and number of occupants on census night.

2.2.5 Electoral Act obligations

Under the Electoral Act 1993, census counts are used as part of the statutory formula to determine electoral population and the number of General and Māori electorates (described further in Section 4).

[7] The CARE principles for Indigenous data governance published by the Global Indigenous Data Alliance (GIDA) can be found here: https://www.gida-global.org/care

2.3 Value of the Census

In the 2014 report *Valuing the census*, Bakker provides an estimate of the financial value of the benefits to New Zealand from the use of census and population information. He states that:

> *The census provides information on people in New Zealand: it has surveyed the entire population every five years since 1881. As such it provides both a comprehensive picture and a linked time series dataset that has no direct comparators (Bakker, 2014, p. 5).*

For uses of census data where it was possible to quantify the value of the benefit, the report concluded that "*every dollar invested in the census generates a net benefit of five dollars in the economy."* Such benefits quantified in the report included:

- The benefits from more accurate health funding allocations as funding is delivered more accurately to more needy areas;
- Reductions in the costs associated with underutilised fixed capital investments, in both the public and private sectors, because of better information on their timing and location (infrastructure funded by Central and local government, aged care, retail);
- Benefits from improved precision and insight in policy making in a range of government agencies, especially for Māori and marginalised groups;
- Improvements in the value added by a range of firms which use census data in a wide variety of analyses provided to government and private sector firms; and
- Gains from improved survey accuracy and reductions in sample size for private sector market research companies, and Stats NZ in respect of a range of other non-census products.

However, for many of the most important uses of census data the benefits could not be quantified. These included:

- Maintaining the integrity of New Zealand's regular process for determining the electoral boundaries for General and Māori seats;
- The basis for the NZ deprivation index (widely used in a range of research and policy work aimed at helping New Zealand's most vulnerable people);
- Underlying work on the Long-Term Fiscal model which informs tax and expenditure policy choices affecting the next 10–50 years.

The report also identified unquantified indirect uses such as: robust demographic data to underpin economic models; use of census data as part of modelling work underlying the calculation of sustainable pathways for regional councils; and ecological modelling used to estimate potential future environmental loads and impacts.

2.4 Operation of the 2018 Census

2.4.1 Census enumeration prior to 2018

The New Zealand census adopts a 'population present' (or 'de facto') enumeration base. That is, it seeks to count everyone present in a household on census night, with additional information on those visiting or away from home, so that outputs on a 'usual residence' basis can be produced. Additional procedures are adopted to count people in non-private dwellings (NPDs) – such as old people's homes, prisons, boarding schools, as well as in hotels and cruise ships, etc.

The traditional New Zealand census model was to hand deliver census questionnaires to households – consisting of a dwelling questionnaire (with information on the dwelling plus a listing of people within the dwelling and their relationships) and separate individual questionnaires for each person within the household. A field officer would seek to make contact at delivery and ask about the number of people requiring an individual questionnaire.

The 2006 New Zealand Census introduced an online option for the first time. The field officer gave those householders requesting the online option an access code to use and 7 percent of households made use of the online option. In 2013, every household was given an Internet Access Code envelope by default, rather than on request. Collectors were also trained and scripted to promote the online option on the doorstep. Around one-third (34%) of households completed their 2013 Census online.

2.4.2 Impacts of changes to 2018 Census

The 2018 Census introduced major design changes to the census operation.

Stats NZ adopted a 'digital first' strategy encouraging online completion of census questionnaires. Paper questionnaires were made available primarily as a back-up and generally upon request. Thus, rather than hand delivering questionnaires as had been the model in previous censuses, there was a post-out of letters containing an Internet Access Code. Almost every household received a letter with an internet access code (either through the post or delivered by field staff). This prompted a shift of the field force from a delivery/collection role to one solely focused on following up non-response.

The 2018 Census design and operational changes are outlined in greater detail in the Report of the Independent Review of New Zealand's 2018 Census (Jack and Graziadei, 2019). That review found the following contributory factors to low response rates in the 2018 census:

- The size of the field workforce was insufficient. Given the uncertainty of how the [new] model would work in New Zealand, the initial targets were too aggressive and further reductions affected field operations.
- The short-term nature of the contracts, a cumbersome onboarding process, and a lengthy delay in provisioning staff with equipment resulted in significant workforce attrition.
- The importance of paper forms in this model was underestimated.

- For the online version, the same access code was required for each individual within the household to complete their respective form.
- Without a [online census] save option, members of the household were forced to complete their full questionnaire and submit. If the form was not completed in a session the respondent had to start over again.
- For certain parts of the country, general communication and/or approaches were not sufficient to trigger public cooperation.
- Some respondents, and sometimes entire communities, waited a long time to receive paper, some indicating they never did. For those who ultimately did receive paper, many thought it was then too late to respond.
- The limited number of bilingual packs (only 4,800 printed) diminished the capacity for field staff to effectively engage with Māori respondents.
- Due to a lack of field resources combined with the decisions to remove paper and contact from list-leave operations, targeted populations and dwellings were not properly equipped to fully participate.
- The strategies developed to increase participation, particularly for Māori, were not effective and in some cases not well executed. The field workforce was understaffed and did not have the tools required to be productive or engage in a meaningful way. This included a lack of coordination between the targeted field operations and the community liaison engagement teams.
- The address list was delivered too late, and the number of duplicate addresses caused challenges during field operations.
- The strategies for collection in non-private dwellings and secure-access dwellings was also an issue. Problems began with the quality of the address frame that supported these operations. Incorrect identification of the dwellings resulted in insufficient workforce and materials to adequately support the operations.
- A lack of resources (both workforce and insufficient paper packs) led to several NPDs not enabled by census night.
- A lack of knowledge and expertise in how to use data from the management information systems meant a less than effective response to mitigate issues as they arose during field operations.
- It was decided not to follow up on partial responses during Non Response Follow Up. This resulted in an increase in partial responses for both online and paper compared to earlier censuses.

2.5 Response rate of the 2018 Census

The response rate to the 2018 Census of Population and Dwellings was substantially lower than previously achieved.

Table 2.1 compares response rates in 2018 against previous censuses for the total population and for Māori, Asian, and Pacific ethnic groupings, and those aged between 15–29 years.[8] The 'Individual form' response rates capture all individuals from whom an

[8] The panel notes that 2018 results should be considered 'interim', based on the estimated New Zealand population at 6 March 2018 of 4,768,600 (Stats NZ 2019a) and that final results will be determined by the

individual form was collected – either online or on paper. The 'partial form' response rates capture all those for whom a household summary form (online) or dwelling form (paper) listing an individual (e.g., their age and sex) was received but the individual form was absent. The total response rates represent the sum of the individual and partial response rates.

Total response rates in 2018 (87.5%) were 5.7 percentage points worse than in 2013 (93.2%). Individual form response rates in 2018 (83.3%) were 8.9 percentage points worse than in 2013 (92.2%), as a result of more partial responses in 2018 (4.2%) than in 2013 (1.0%).

Māori (74.3%) and Pacific (73.5%) and those aged 15–29 (81.1%) were disproportionately missed, and also had disproportionately more partial responses (Māori: 6.1%; Pacific: 8.4%; those aged 15–29 years: 6.1%). This is a notable change from 2006 when there was >90 percent response for each of these groups, and partial responses did not exceed 1.5 percent.

Table 2.1: 2018 Census response rates (%)

	2006			2013			2018 (Interim)		
	Individual form	Partial form	Total	Individual form	Partial form	Total	Individual form	Partial form	Total
National population response rates	94.5	0.6	95.1	92.2	1.0	93.2	83.3	4.2	87.5
Sub-group response rates[a]									
– Māori	93.1	0.6	93.7	88.5	1.2	89.7	68.2	6.1	74.3
– Asian	91.0	1.1	92.1	91.7	1.6	93.3	81.7	6.1	87.8
– Pacific	92.4	1.5	93.9	88.3	2.5	90.8	65.1	8.4	73.5
– 15–29-year-olds	91.9	0.9	92.8	88.5	1.8	90.3	75.0	6.1	81.1

Source: Modified from Jack and Graziadei, 2019

[a]Ethnicity was not imputed in 2006 or 2013 Census data. These response rates are estimates based on imputation of level 1 ethnicity indicators for population estimates as published in NZ.Stat. These assume that the ethnic imputation is applied at the same rate for all records requiring imputation.

These response rates show that many people were missed by the Census in 2018 – far more than the previous two censuses – and that response rates varied by ethnicity and age. Consequently, in order to achieve coverage of the population comparable to 2013 and 2006, the addition of records from sources other than the 2018 census was required. This included a greater proportion of Māori, Pacific ethnic groups, and those aged 15–29 years. Note that this represents a change in practice from previous censuses: the 2018 Census is the first for

results of the Post Enumeration Survey, which is expected in March 2020. However, these interim response rates are likely to differ only very slightly from the final response rates.

which the usual resident population count includes a count of those who did not complete a census questionnaire.

Table 2.2 shows how the population for census outputs (the 'Census usual resident population') has been built from census data and other sources. A large proportion (84.5%) was obtained by completed individual forms. A further 4.3 percent was obtained by 'partial respondents', that is, those listed on a household set-up form or dwelling form but who did not complete an individual form, and a small number of rough sleepers (enumerated on an individual form without an accompanying dwelling). Thus, 88.8 percent of the census usual resident population was obtained directly from completed Census forms.

Table 2.2: 2018 Census usual resident population count, by unit record source

2018 Census usual resident population count By unit record source				
Unit record source	Count		Percent	
Individual form	3,971,892		84.5	
Individual from household listing	202,914		4.3	
Field enumerated rough sleeper	99		<0.1	
Census responses		*4,174,902*		*88.8*
Admin household enumerations – occupied	99,159		2.1	
Admin household enumerations – unoccupied	42,252		0.9	
Admin household enumerations		*141,411*		*3.0*
Admin enumeration at responding private dwelling	20,643		0.4	
Admin enumeration at prison or penal institution	4,707		0.1	
Admin enumeration at defence establishment	798		<0.1	
Other dwelling-based admin enumeration		*26,148*		*0.6*
Admin meshblock enumerations		*357,294*		*7.6*
Total		*4,699,755*		*100*
Source: Stats NZ. (2019a). All numbers rounded to the nearest 100.				

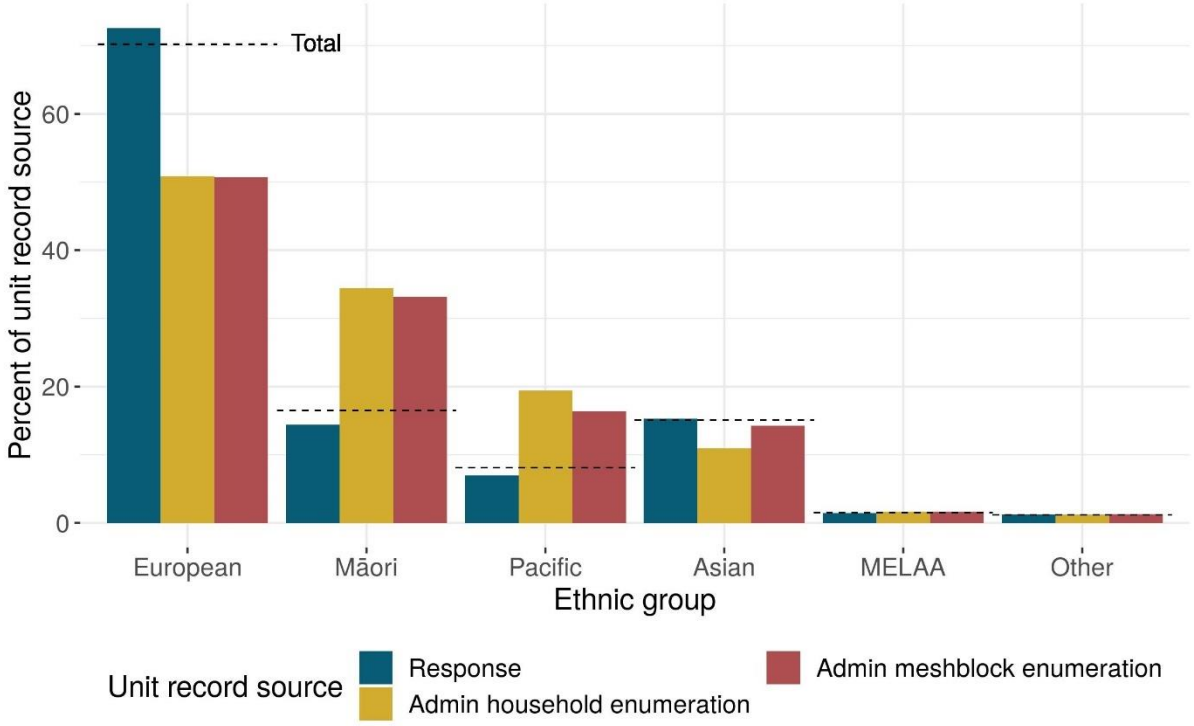
The remainder were obtained from administrative sources (i.e., the Integrated Data Infrastructure (IDI)[9]] dataset maintained by Stats NZ, as well as prison and defence establishments). Administrative enumerations either added individuals into households (3.6%), or – where households could not be accurately determined for individuals – into meshblocks (7.6%). The methods for obtaining these additional records from administrative sources are described and discussed in section 3.

[9] The Integrated Data Infrastructure (IDI) is a large database containing de-identified data about people and households. Data are sourced from New Zealand 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.

Note that the final Census usual resident population of 4,699,800 is estimated to cover 98.6 percent of the estimated New Zealand population at 6 March 2018 of 4,768,600 (Stats NZ 2019a; this is an interim estimate, as noted above). The under-count of 68,800 represents 1.4 percent of the estimated New Zealand population, compared to 2.4 percent in 2013 and 2.0 percent in 2006. However, the 2018 result is obtained only **after** 524,900 were added to the Census dataset from the administrative data.

Figure 2.1 shows that records added from administrative sources were far more likely to be for Māori and Pacific New Zealanders. The impact of adding administrative records is shown by the dotted lines, and indicates that had administrative records not been added, the Māori and Pacific share of the population would be **underestimated**, and the European share of the population would be **overestimated**.

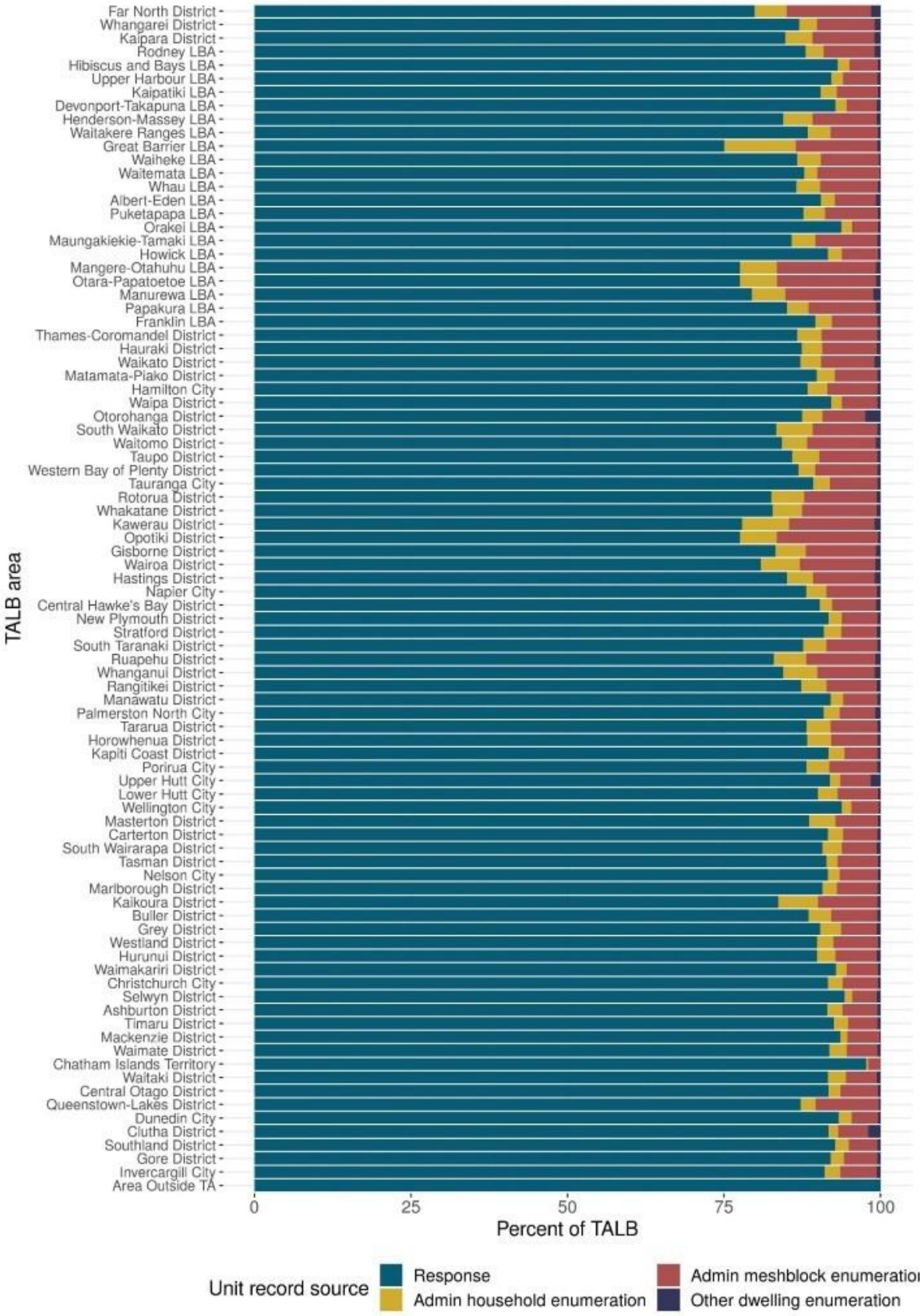
Figure 2.1: Percent identifying with level 1 ethnic groups, by unit record source



MELAA; Middle Eastern, Latin American and African
Source: Stats NZ (2019a).

Figure 2.2 shows the areas of the country (territorial authorities and Auckland local boards) where Census responses were lowest, and therefore records were more likely to be obtained from administrative sources. Areas most affected by Census low response (>20%) were the Far North District, Auckland local boards including Great Barrier, Mangere-Otahuhu, Otara-Papatoetoe, and Manurewa, as well as Kawerau and Ōpōtiki. These are all areas with relatively high Māori and/or Pacific populations.

Figure 2.2: Percent of population from each unit record source, by Territorial authority and Auckland Local Board area



Source: Stats NZ (2019a)

These results highlight a systematic under-count of Māori and Pacific and the young as part of the Census operation, particularly in the Far North, parts of Auckland, and the Eastern Bay of Plenty; and, as a consequence, a greater reliance on administrative sources for these groups. Without some attempt to improve the data quality, these population subgroups would be particularly poorly described by the Census data, and overall descriptions of the New Zealand population would be biased.

2.6 Fitness for use of census characteristics

In this report, we have explicitly considered several key elements of quality, and our conclusions are summarised in this section.

In order to offset some of the loss in quality because of the low overall completion rate for census questionnaires compared to earlier censuses, Stats NZ had to put in place a new methodology during the last twelve months. One outcome is that for those characteristics obtained only from the dwelling or individual census questionnaires, their fitness for use needs to be evaluated at each application, rather than presumed as in the past. Supporting this need, Stats NZ has made available a wide range of quality measures about each characteristic.

The key effect of the post-census integration with administrative records has been to strengthen the **representativeness** of the usually resident population of New Zealand in the most frequently used measures of age, sex, place, and ethnicity that define New Zealand's distinct communities. After the remedial action by Stats NZ for the 2018 Census of Population and Dwellings, we have the capability of measuring these attributes for a larger share of the population than occurred in 2013, and probably before that.

Although the coverage of the population is more complete, the **specificity** or precision with which place and ethnicity have been measured is less than in earlier censuses. Just under eight percent of the population cannot be placed at an address, although it has been possible to establish the meshblock where they would usually be resident. Ethnicity has been obtained from a mix of sources, and the collection methods and classification systems used have not been the same.

Each Census of Population and Dwellings is part of a very long-term statistical series, measuring change, levels, rates and shares across population groups, time and place. The **comparability of measures of intercensal change** has been severely disrupted between 2013, 2018 and 2023. The different methodology and field enumeration problems add variability to measures of change between 2013 and 2018 for measures based on census characteristics as well as alternatively sourced data. Given that in 2023 Stats NZ will certainly not wish a repeat of enumeration problems on the scale experienced in 2018, measures of intercensal change will again be disrupted by changes both in the non-response rate differences, and from any further methodology changes.

In any Census of Population and Dwellings, there is always a tension between modifying concepts and questions to maintain their relevance and bringing intercensal consistency by keeping questions the same. Given the uncertainty that now exists about the coverage of the 2013 Census of Population and Dwellings, the differences in methodology and non-

response rates are likely to overshadow how such ongoing tensions between consistency and relevance are managed.

The share of the relevant population **represented** by key derived units of analysis used in census analyses has fallen compared to earlier censuses. The representativeness of analyses will need to be validated in accord with the intended use. Of some of the more frequently used units of analysis:

- Just under eight percent of the population cannot be placed in households and families. Not surprisingly given the scale of the non-response, the post-census integration with administrative records has not been able to resolve fundamental difficulties in the preparation of household and family statistics 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 eight percent of the population.

Across the distinct communities defined in some way by age, sex, age or ethnicity, there is not **consistency in the coverage of characteristics** that were only reported in the Census at enumeration time. For some communities defined in this way, we have this census-only information for fewer than half of the population. The graphs attached to this report highlight how much coverage can vary.

The **high coverage of government departmental administrative records** obtained by statute from Inland Revenue, Corrections, education, and health sectors have resulted in a more complete coverage of the population for these characteristics than possible through the usual form of census enumeration.

In making comparisons with earlier censuses, the high non-response in the 2018 Census for census only questions will mean **analyses of special interest groups** (defined by e.g., occupation, disability, smoking) will be inconsistent with that of earlier periods for which comparisons are made. For groups defined in this way, comparative analyses with earlier censuses, or comparisons between groups defined by the categories, will be of lower quality. This includes information that can be provided from the census questions about different ethnic groups, forms of employments status, travel origin and destination analyses, iwi membership, age group studies, religious group membership, and smoker characteristics. As discussed above, where the responses to one question informs responses to another, such as the way that language use and residence 5 years ago are important in analysing ethnic groups, the information loss will be greater.

Known methodology limitations exist for several questions (sex, disability, family). 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. The sample was selected from both positive and negative responses because on this topic simple questions have been found to generate wrong 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 censal survey."* It is likely that this experience will have some significance in assessing the

quality of the results from the activity limitations topics in the 2018 Census of Population and Dwellings.

Reductions in quality of varying severity from the loss of specificity in place and ethnicity, reduction in key analytical units, widely varying coverage of census-only characteristics, and under-coverage of special interest groups will offset the strengthened representativeness of the frequently used measures that define distinct communities, and the high coverage of some characteristics in administrative records. In any specific application, each of these quality dimensions needs to be considered to determine how they balance out. Quality measures produced by Stats NZ are available for most of these aspects.

3 Statistical Methods

3.1 Introduction

The New Zealand Census has traditionally been based on straightforward enumeration of the population, households, and dwellings, with relatively little statistical adjustment after the fact. For a census of this type, the individual and household response rates and the completeness of each variable provide a good guide to the quality of the data.

Even in the past, the Census had to consider non-response. A small number of variables – age, sex, usual residence (for those away on census night), and labour force status – were previously imputed where missing. Māori descent was imputed only for Electoral purposes. Values for occupied but non-responding households were taken from field staff fieldbooks and nearby 'donor' households. The Post-Enumeration Survey was used to provide independent data on population size which, while not used to adjust the census counts, was used in constructing population estimates.

The 2018 Census had a very low response rate. Stats NZ has undertaken major efforts to augment the census enumeration with data from other sources. These mitigation strategies have produced a census file with high completeness for many variables, particularly those that categorise the population by ethnicity, age, sex, and location. In addition, obtaining income from tax records has resulted in a higher completion of responses than ever before for this topic. It is important to recognise that neither the low response rate nor the high final completeness is a reliable indicator of data quality. The data are not as bad as an 89 percent household response rate would indicate, nor are they as good as the final data completeness might suggest.

There were three types of incompleteness in the census enumeration of people:

- Complete non-response: no census forms have data for this person (estimated at 593,700 people)
- Household form only: the person is mentioned on the household form but not on an individual form (202,900 people)
- Individual-variable non-response: the person filled in an individual form, but some questions were not answered (varies by question).

Stats NZ added data from other sources to mitigate all three of these types of incompleteness.

This section on statistical methods aims to:

- Summarise how additional data were added
- Review the appropriateness of these mitigation methods
- Describe some of the ways the mitigation methods affect data quality and fitness for particular uses.

3.2 Mitigation methods

The scale of non-response was a surprise to Stats NZ, and the original plans for handling low-level non-response needed to be updated. From the panel's first meeting in August 2018 until February 2019, our main role was providing comment to Stats NZ methodologists as they presented their options. This section is based on presentations made to us during this period.

3.2.1 Options considered

The original plan was to use administrative data to fill in specific missing variables in census responses, to impute non-responding households using a near-neighbour approach, and to use the post-enumeration survey to adjust the population estimates. The planned use of administrative data, in a limited way, was essentially new in Census 2018

Two additional approaches were initially considered. The first approach was to treat the Census Dwelling Frame as an accurate census of dwellings and using administrative records to add households to non-responding dwellings. However, linkage of non-responding dwellings to the IDI spine (without any individual data) proved to be insufficiently accurate as a base for the whole mitigation strategy, as did the census data on which non-responding dwellings had usual residents. It was not possible to simply start with dwellings, and place people in them.

The second approach was to use Dual System Estimation (capture-recapture modelling) to combine the census and administrative data files and create an estimated Census population. In contrast to the usual uses of Dual System Estimation, the goal was not just to estimate the number of missing people, but to actually create records for them, using a sophisticated Bayesian model. Stats NZ decided this procedure would not have been computationally feasible with available time and resources, and it would not have provided household and family information.

Modified forms of both approaches have been used in the final Census data, as described in detail below. The addition of administrative records uses addresses from the administrative data and is not restricted to placing people into a Census dwelling. Dual System Estimation has been used to provide benchmark counts for the population as a whole and subpopulations rather than to create a census record file.

3.2.2 Filling in variables from administrative data

Census 2018 records were probabilistically linked using variables including name, date of birth and meshblock to the IDI spine, the set of cross-database links that is used to create the Integrated Data Infrastructure (IDI). Linkage to the IDI spine provides a unique identifier allowing census data to be linked to other IDI databases, which have had names removed (Stats NZ 2019a). These other databases include the 2013 Census dataset, birth, death and immigration data, education enrolment data, benefit data and tax data.

Stats NZ report that 97.7 percent of Census 2018 records were linked to the IDI spine (Stats NZ, 2019d). As they estimated 1.2 percent of links were missed, this suggests that very few people were actually not on the IDI spine. They estimate that less than one percent of links

are incorrectly joining two different people. These linkage rates are higher than has generally been the case for the IDI, which Stats NZ attribute to the higher quality of census data.

When an individual was identified in the Census 2018 data and successfully linked through the IDI spine to the IDI, existing administrative data could be copied across to fill in gaps in individual variables. For example, birthplace was taken from immigration or birth records; smoking and usual 2013 residence from the 2013 Census; and ethnicity mostly from birth, education and health data, with a very small proportion from Corrections and defence.

Linkage of a specific Census record to administrative data was particularly valuable for individuals who were named on the household form but for whom an individual form was not received. These individuals would otherwise be represented only by age, sex, and relationship to the person who filled out the household form. However, linkage was less successful for such people – only 83.9 percent were able to be linked to the IDI spine.

3.2.3 Addition of individuals from administrative records

For many New Zealanders (an estimated 12.4% of the population), the 2018 Census field operation obtained no individual data at all.

The IDI is intended to contain records for everyone who has ever lived in New Zealand, so it should contain records for nearly all the current New Zealand residents who did not respond to the Census (there are some exceptions, such as citizens of Australia, Niue, Tokelau, or the Cook Islands who have entered without needing a visa and who have not yet obtained an IRD number). The IDI over-covers the New Zealand resident population for two reasons. First, it includes people who are no longer New Zealand residents. Second, there will be a very small number of people who have more than one record on the IDI Spine.

The file used to augment the Census was called the IDI-ERP, standing for “Estimated Residential Population constructed from the IDI”. The IDI-ERP was intended to include everyone who had lived in New Zealand in the two-year period prior to the census. It was constructed by including people who have tax, health, education, or ACC data within the two years prior to Census day, plus permanent or long-term arrivals within the previous two years, plus children under five with birth or (long-term) visa records. Records with linked death or emigration data indicating they were not resident at the Census date were removed. Records clearly identified as duplicates were removed.

3.2.4 Households and dwellings from administrative records

Creating a census record from administrative data requires an address. For each individual, a set of sources considered sufficiently reliable were scanned (including the 2013 Census, motor vehicle registrations, address of residence reported for medical care, and contact addresses from Inland Revenue). The most recent address from all these sources was used.

The address information was used in a four-level process.

First, people could be **added to households who responded** to the Census, in a narrow set of circumstances. If responders said they lived with someone for whom there was no census form, and there was a person in the administrative data with the correct address and

relationship to the responder, they would be added. In total, 20,600 people were added to 13,500 dwellings, with 9,300 dwellings receiving just one added individual.

Second, records from the IDI-ERP were used to **create full households** populating some non-responding private dwellings. It is straightforward to find sets of people in the IDI-ERP who have the same recorded address, but a valid household record requires more that.

Following an approach from the US Census Bureau (Keller et al, 2018), Stats NZ built two statistical models using census and administrative data from census responders. One model was for the probability that an individual's administrative address was correct, i.e., that it would have been the address given on a census form if one had been received. The second was the probability that a household had the composition that the administrative data suggested it did.

An administrative household was created if the probability of having the correct address was high enough for every individual in the household, and the probability of having the correct household composition was high enough. Details are in Matheson-Dunning & Lin (2019). Of 89,400 potential administrative households, 57,600 were considered sufficiently reliable to add to the census data.

Third, after the previous steps there were still a substantial number of individuals who were present in administrative data but had not been assigned to the Census. A more restrictive administrative file (called **IDI-ERP Sure**) was constructed by requiring that people (other than infants under 1 year) have at least two sources of administrative data – tax or health data, plus something else. Next, a small number of records were randomly removed to compensate for double-counting due to missed links. This third step removed 181,000 records.

In the **fourth** step, the remaining 422,100 administrative records were regarded as genuinely missing from the Census count but could not be assigned to a responding household or a non-responding dwelling. A further statistical model (based on Census responders) was used to predict whether their address was likely to be in the correct meshblock. If it was, they were **assigned to that meshblock but not to a dwelling;** if it was not, they were omitted. This step added an additional 357,300 individuals to the census count.

The Census undercount will consist of the remaining 64,800 individuals from the IDI-ERP Sure (who are known not to be counted) plus a small number of Census non-responders not present in the IDI-ERP Sure. The estimated undercount of 68,800 quoted in Section 2 above, which comes from Dual System Estimation, is consistent with these figures.

Table 2.2 in Section 1 shows all these contributions to the Census total.

A special case of the first two steps occurs for two specific types of Non-Private Dwellings: **Prisons** and **Defence Establishments**. The Statistics Act places an obligation on the person in charge of the dwelling to ensure census response, but the response rate in both cases was low.

Many prisoners were apparently not given an opportunity to fill in the Census. Individuals in prison will be in the administrative data, but these data will not specify which prison they are in. Upon request, the Department of Corrections supplied Stats NZ with unit-record data

files for every prisoner with details (where available) of age, date of birth, sex, ethnicity, iwi and location. This allowed data for Census non-responders identified from administrative records to be placed in the correct prison locations.

While the data from Corrections did not include names, Corrections provided the Justice ID, enabling linkage to the IDI Spine. Nearly half (4,700 out of 9,700) of the records for the prison population come from administrative data. The quality of iwi data was too poor to use for census mitigation and ethnicity data was subsequently drawn from other administrative datasets, with only a very small proportion sourced directly from Corrections.

There is no Ministry of Defence data in the IDI, but the Ministry supplied similar data to that from Corrections. Census responders were matched and removed, and a list of date of birth, sex, and dwelling of census non-responders determined. These variables determine a candidate list of people in the administrative data; the list was randomly sampled to give the required number of records. About a quarter (800 out of 3,200) of the records for Defence establishments come from data supplied directly by the Ministry of Defence.

3.2.5 Imputation of values from other individuals

Many missing Census variables are also available in administrative data or the previous Census, and so could be filled in when individuals were added to the Census file in the previous section, improving the data completeness substantially. However, not all variables can be filled in this way. One reason to include a question in the census is precisely that the information is not available from alternative sources – this is especially true of the new questions in 2018. Some questions asked were new for 2018 (e.g., different housing quality and disability variables) and so were not recorded in 2013. Some variables will not by their nature be stable over the period since 2013. These include iwi affiliation (where Stats NZ has already decided not to release census results) and smoking.

Missing data for these variables can only be filled in by imputation from someone else's data. Imputation is also needed for other variables when an individual happens not to have had the information recorded in administrative databases.

Statistically, imputation methods exist on a spectrum from fully deterministic to highly stochastic. A deterministic system uses the most likely value of a missing variable for an individual; a stochastic system has a high probability of picking the most likely value, but it might instead pick less likely values with correspondingly lower probabilities.

Deterministic imputations will perform better when the data are used to estimate complex relationships between variables, because the imputed value is more likely to be the correct value. Stochastic imputations will perform better when the data are used to summarise the population distributions of single variables, because the imputed values represent the uncertainty in imputation more accurately.

Stats NZ used nearly deterministic imputation, where a missing value is copied from one of a small set of similar people or households. Specifically, they use a system called CANCEIS (CANadian Census Edit and Imputation System), developed by Statistics Canada.

The CANCEIS system searches records that are near neighbours on a master list to find potential donors who are good matches on a set of specified matching variables. Once the

good matches are found, the system can either pick randomly from them or simply choose the closest match; the latter mode of operation was used.

The list was sorted geographically, so the near neighbours will have been geographically close and will tend to be similar in more ways than just the matching variables. The search of near neighbours initially examines 5000 records. If no satisfactory match is found, the search is progressively expanded.

In CANCEIS, imputation is carried out using a set of sequential modules so that, for example, information about usual residence is imputed before trying to impute ethnicity. Ethnicity, in turn, is imputed before income, smoking, and employment variables. Dwelling and family information is imputed last. This sequential imputation procedure makes it more likely that when multiple variables are imputed for an individual, they are imputed together and maintain at least some of the relationships between variables.

The CANCEIS procedure ran separately for different members of a household: information on disability or smoking or occupation for two household members will not typically have come from members of the same donor household and need not have any particular relationship.

A few variables were imputed in other ways. Sex was imputed from given name in a very small fraction of cases. The net impact of this on Census data will be very small, but it is an undesirable practice from the viewpoint of sex and gender diversity. Stats NZ have already signalled their plans to handle sex and gender differently in the future. Date of birth was imputed from age.

Several variables related to ethnic and religious identity were imputed using a random 'donor' record from the individual's household, in preference to considering people outside the household. This strategy has the benefit of reducing the under-representation of small minority groups, since members of the household are more likely than neighbours to share ethnicity or religion. As a trade-off, it will reduce the estimated within-household ethnic or religious diversity.

Māori descent for electoral purposes had a more complicated imputation approach; we deal with this variable in detail in section 4.

3.2.6 Dual System Estimation

Separately from the augmentation of the census data by administrative data, Stats NZ attempted to estimate the size of the usually resident population, both as a whole and broken down geographically and by age, sex, and ethnicity, using Dual System Estimation. This technique is widely used for population size estimation in official statistics and in ecology (where it is known as capture-recapture estimation), but some features of the application to the census are unusual.

The basic idea of Dual System Estimation is that we have two lists, from the Census itself and from the IDI-ERP Sure, that each attempt to capture the whole population. We know how comprehensive the administrative list is for Census responders: we know what fraction of them were missed by the IDI-ERP Sure. We also know how comprehensive the Census is for people in IDI-ERP Sure: we know what fraction of them were Census non-responders. If being missed by the administrative list is independent of being a Census non-responder and

is the same for everyone, it is simple to estimate the number of people missed from both lists, allowing the true population total to be estimated. In typical applications of the method there are concerns that the two lists have similar strengths and weakness and will tend to miss the same hard-to-find subsets of the population. In the Census this is less of a concern because the lists are constructed in very different ways.

At a whole population level, the assumption of independent and homogenous recapture probabilities would be relatively implausible: some groups, such as recent immigrants, will reliably be in administrative data, and their rate of Census non-response need not be the same as for the rest of the population. The assumption may be plausible, however, for smaller and more homogeneous population subgroups. Dual System Estimation can be used to estimate the size of each of these population subgroups, and the estimates can be added up to give a population size estimate, as well as population size for larger subgroups such as all Māori or all people from the Southland region. The DSE models used by Stats NZ assumed homogeneity/independence for people within a Territorial Area/Local Board (there are 87 of these: 66 territorial authorities and Auckland territorial authorities divided into its 21 local boards). Without the further stratification by age and ethnicity that Stats NZ used, this would be a concern. Maps of census response rates did show smaller-scale geographic variation, but in ways that could reasonably be explained by ethnic and age-group differences. Stats NZ did not have time to explore the homogeneity assumption in more detail, but the panel regards it as reasonable.

There were other technical issues in running dual system estimation. Two key ones in this setting are linkage error and timeliness of data. They are more important here than in typical applications of the method because the two population lists both contain a large fraction of the population and because very high accuracy is needed. Typically, one or both of the lists is from a relatively small survey, and the resulting sampling uncertainty is large enough to make small biases in estimation relatively unimportant. For Census 2018 the sampling uncertainty is very small, and even small biases matter.

Failing to link a Census responder who is actually present in the administrative file will be rare, but the competing explanation – that a Census responder is truly absent from the administrative file – is also unlikely. If linkage errors were not accounted for mathematically, the Dual System Estimation would conflate these two types of linkage failure and seriously overestimate the number of people missing from the administrative data. Stats NZ used statistical formulas that do account for the linkage error, so it should not lead to bias in the estimates. [Table 4, Stats NZ (2019c)]

Timeliness of data is important because Dual System Estimation fundamentally relies on the two lists being for the same population, but the New Zealand population is continually changing (people move house, are born or die, or enter or leave New Zealand). To the extent that the two lists are not from the same time point, the number of discrepancies between the lists will be larger than the number of actual omissions from either list and the population size will be overestimated.

Stats NZ invested substantial effort and expertise in generating the Dual System Estimation estimates. Their ability to validate the estimates further was limited by the available time, and it is conceivable that further research would show poor performance in some subpopulations, but the estimates are based on established statistical methods and reasonable modelling assumptions.

3.3 How much were these mitigation methods used?

Figure 2.2 in section 2 shows how the data sources for the usually resident population count vary by territorial area or local board. The fraction of the count based on census forms is mostly between 80 percent and 90 percent. The majority of the administrative records are added in at the meshblock level (357,300) rather than as whole households (141,400) or into responding dwellings (20,600).

Figures 1 and 2 in section 2 show that people counted from administrative sources alone are more likely to be young adults and to be from Māori or Pacific ethnic groups. Young adults are progressively more over-represented among those assigned only to a meshblock and among those not assigned even to a meshblock (and so not counted).

In our final report we will present graphs of data sources by variable for each region and for the SA2 in each region with the lowest fraction of Census 2018 data for that variable. These graphs are likely to show considerable geographic variation in the use of mitigation methods. For nearly all variables there is likely to be at least one SA2 with less than 50 percent of data from Census 2018 forms. When 2013 Census or administrative data or imputation are available, the data quality for these small areas will be very sensitive to the quality of mitigation methods; when administrative data or imputation are not available, there will be substantial missing data.

It is important to note that some **values** are more likely to be taken from administrative data or imputed than others. For example, people of Māori descent were more likely to be non-responders than those not of Māori descent, so that administrative or imputed values for this variable will disproportionately indicate Māori descent.

This differential mitigation rate is a necessary consequence of differential response rates and does not (on its own) indicate that the data are of poor quality, but it does mean that population-wide summaries of the amount of data added can be misleading for subpopulations, and particularly for specific subpopulations at sub-national levels (e.g., Chinese or Indians in Rotorua).

3.4 Appropriateness of mitigation methods

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

It is clear that the use of administrative and past-census data has improved the quality of the Census results. The addition of administrative records from the IDI-ERP Sure population reduced the Census undercount dramatically, and the DSE benchmarks confirm that the undercount is small. The net undercount of Māori and Pacific Peoples, in particular, appears to be lower in this Census than in recent censuses.

The use of administrative and historical data to fill in Census non-response, and the use of imputation where administrative data was unavailable has allowed Stats NZ to claim, and the panel to endorse, ‘High’ or ‘Very High’ quality for some important Census outputs, as described in later sections.

However, the use of administrative data can only go so far: as the 2015 Cabinet Paper *Census Transformation - Promising Future* noted "a census based on administrative data is not yet possible." In section 3.5, and in evaluations of specific variables, we indicate where the mitigation will be most successful and where the data quality will still be of concern for some uses.

3.4.1 To what extent have the statistical methods been used before?

Dual System Estimation is a standard and well-understood statistical technique. It has been used frequently in official statistics and in other areas of statistics when population size estimation is needed. The use of DSE to construct population benchmarks involves more reliance than usual on some of the assumptions of the method, because the goal is very precise estimation. We believe that Stats NZ have considered the assumptions carefully.

CANCEIS (Canadian Census Edit and Imputation System) is high-quality software from Statistics Canada, implementing well-understood methods. It was used by the UK Office of National Statistics for imputation in their 2001 and 2011 Censuses and in the 2010 Brazilian Census. Stats NZ is using imputation for a larger fraction of data than is usual in a Census, so the quality of the resulting data will inevitably be lower.

The linkage techniques used to construct the IDI and to link the Census to the IDI Spine are standard and appropriate.

The approach to modelling the correctness of households added from administrative data is based on work by the US Bureau of the Census and is described in a peer-reviewed publication (Keller et al., 2018).

The model for the correctness of a meshblock assignment, used to decide whether to add individuals to a meshblock or leave them out, is novel, but is a straightforward statistical model based on standard regression approaches.

3.4.2 Ethical and legal questions

The use of linked data from the past census and from other government administrative records raises three key questions about privacy and consent:

- Was the data linkage legal?
- Did it follow Stats NZ policies for data integration?
- Was it consistent with the actual or reasonable expectations of the people ultimately providing the data?

Stats NZ have published statements about what they see as the legal bases for the use of administrative data in the 2018 Census.[10] They have also briefed the panel on the Privacy Act 'repurposing' powers, which allow a ministry to share information externally (disclose for a different purpose without consent from the individual) where this is for the purpose of statistics and research.

[10] https://www.stats.govt.nz/privacy-impact-assessments/creating-the-2018-census-dataset-by-combining-administrative-data-and-census-forms-data-our-privacy-considerations

Data integration policy

The Stats NZ Data Integration Manual (Stats NZ, 2015) sets out the following principles:

- Principle 1: The public benefits of integration outweigh both privacy concerns about the use of data and risks to the integrity of the Official Statistics System, the original source data collections, and/or other government activities.
- Principle 2: Integrated data will only be used for statistical or research purposes.
- Principle 3: Data integration will be conducted in an open and transparent manner.
- Principle 4: Data will not be integrated when an explicit commitment has been made to respondents that prevents such an action.

Principle 1 – The benefits of the improved census dataset argue for this principle being satisfied.

Principle 2 – Is clearly satisfied for the Census.

Principle 3 – Stats NZ published information about the 2018 Census data integration after the need for large scale data integration became apparent. This includes an updated Privacy Impact Assessment in April 2019 [11].

Principle 4 – The panel has concerns about Principle 4, which are spelt out more fully below.

The UN Census manual states that "The political decision concerning the use of administrative data in a census can be highly influenced by public approval or refusal. In the run-up to implementing a new or modified census methodology it is helpful to inform the public about the project" (2017, p. 26). Māori and Pacific peoples, who were among those most affected by the extensive use of alternative government data, did not have the opportunity to be informed about the revised census methodology prior to its implementation, or to have input into decision-making. This is particularly important for Māori given Stat NZ's Treaty obligations and the increasing call for governments to take account of Māori and Indigenous data sovereignty (Canatacci, 2018; Jonas, 2018; Kukutai & Taylor, 2016; UN Special Rapporteur on the Right to Privacy, 2019).

The panel also has some concerns about Principle 4. The panel was not able to assess whether integration undertaken as part of census mitigation aligned with commitments made to respondents at the time of data collection for data subsequently included in the IDI and used as alternative data sources. The panel were not provided with information on the commitments for the individual administrative data sources from the IDI that were drawn on to produce the Census 2018 dataset.

Another concern is data provided to the IDI under rule 11(2)(c)(iii) of the Health Information Privacy Code (1994). The code states that information about an individual can be disclosed without specific authorisation of the individual if: the health agency (Ministry of Health [MoH]) believes (reasonably) that it is not desirable or practical to obtain authorisation; the information is to be used for research purposes (for which ethics committee approval has

[11] https://www.stats.govt.nz/privacy-impact-assessments/creating-the-2018-census-dataset-by-combining-administrative-data-and-census-forms-data-our-privacy-considerations

been obtained, if required); and will not be published in a form that could reasonably be expected to identify the individual.

Information of this sort would be legitimately present in the IDI, but it is not clear how its use to fill in Census non-response would be a research purpose. In addition, it is unclear to the panel whether the MoH confirmed that this use of health data meets the rule, or whether ethics committee approval was required and, if so, whether it was sought and obtained.

Social licence

Beyond the question of whether there were specific commitments given to the individuals who ultimately provided the data, there is a broader question of whether the linking of alternative government data to census data enjoys social licence (i.e., tacit approval from the New Zealand public). A recent New Zealand study of social licence and integrated data notes that *"social licence cannot be conferred if the relevant community is unaware of the agency seeking it"* (Gulliver et al., 2018). Prior to undertaking the census, Stats NZ acknowledged in papers published on its website that it intended to use alternative data sources to reduce non-response (rather than as a core part of the census model, Stats NZ, 2017). It also published an independent Privacy Impact Assessment on Census 2018 (Simply Privacy, 2017).

The privacy report, published less than a year before the census, stated that there is good reason to trust Stats NZ but that the agency "should assume a low level of social licence and target its practices at developing openness and transparency to show value, build trust and start to earn one" (Simply Privacy, 2017, p. 13). It also recommended that Stats NZ "provide clear notice to the public about key issues that may impact on public perception, including the retention and use of names and addresses and integration with the IDI and explain that this is legitimate and adds value".

The individual and dwellings census forms did not contain this information[12]. However, the guide notes accompanying the individual form stated that names were retained in order to match individual forms to individuals listed on dwellings forms to help determine family and household structures, and for the "initial matching process" when census data is added to the IDI.

The UN advises that (2017, p. 13), "linkage operations should be undertaken with caution, ensuring not only that all national laws are met but also that the trust of the public in the census and the statistical systems is maintained". Retaining the trust of Māori is especially important, given that Māori tend to have lower levels of generalised and institutional trust but are among those most impacted by the extensive use of administrative data for census mitigation. In addition, Te Mana Raraunga, the Māori Data Sovereignty network, has argued that social licence is a necessary but insufficient condition for the trusted use of Māori data, and that a collective mandate or "cultural licence" is needed (Jenkins, 2018)[13]. Cultural

[12] The individual and dwellings census forms noted that the Public Records Act 2005 required census forms to be retained.

[13] Te Mana Raraunga (2017). Te Mana Raraunga Statement on Social Licence. Retrieved from https://www.temanararaunga.Māori.nz/panui The cultural licence to operate is also noted in the Primary Sector Science Roadmap.

licence is defined as the ability of an organisation to use and share data in a legitimate and acceptable way, based on the trust that iwi and Māori Treaty partners have.

The panel notes that there has not yet been a comprehensive and open public consultation with New Zealanders to gauge the acceptability of the revised census approach. As a result, we are not confident that people in New Zealand understand the extent of data sets that are linked to the census, nor that it would not affect their willingness to provide data if they did understand. Perhaps more importantly, we have not seen convincing evidence that this is the case.

Also relevant to social licence is the process used to obtain prisoner and defence data for Census 2018. In past censuses the enumeration of remand and sentenced prisoners was completed as with any other form of non-private dwelling. For Census 2018 there appears to have been issues with the implementation of agreed census enumeration procedures at the prison level, leading to high levels of under-enumeration of prisoners. Although the data provided by the Department of Corrections was primarily used for matching purposes, rather than as a primary data source, the change in method might still be perceived as pushing the bounds of social licence for a population with already limited rights.[14] For those in Defence establishments, the data used to populate the records of census non-responders were sourced directly by the Ministry of Defence as there was no alternative Defence data in the IDI. In neither case was consent to share the information sought or obtained from the individuals concerned as this was not deemed necessary.

3.5 Statistical limitations

In this section, we present our overall views on the likely limitations in data added to mitigate non-response. More details for specific variables and uses will be provided in our second report due in December 2019.

3.5.1 Timeliness

A census has a common reference date (6 March 2018 for this census) and seeks to collects data for everyone in the population relating to that reference date (even if collected before or after the event). Values in the administrative data sets could have been collected up to five months after the census date or potentially a long time before that date. The time offset is a combination of the time intervals at which other agencies send updates to the IDI and the time intervals at which members of the public supply new data to the agencies. For example, a value of ethnicity from education data might have been supplied to the IDI in December 2017, from an enrolment in February 2017, which might itself have defaulted to the value given the first time that student enrolled at that school. With data from the 2013 Census the delay is, of course, five years.

For variables that rarely change over time (sex, birthplace) or that change in a predictable way (age, time since moving to New Zealand) the issue of timeliness is not particularly important. For variables that do change, such as household composition, education status,

[14] In 2010 the Electoral Act was amended so that no prisoner in New Zealand is able to vote in a national election. Previously, only prisoners serving a sentence of three years imprisonment or more were prohibited from voting.

or smoking, the timeliness is more important. People's self-perception of 'identity' characteristics, such as ethnicity and descent are known to change over time for a proportion of the population (see section on ethnicity).

Moreover, for some variables that rarely change, knowing when they change and for whom is often of interest (e.g., understanding who quits or begins smoking, or understanding who changes their ethnic identification).

3.5.2 Mobility and 'usual residence'

Household mobility and change in household structure are a special case of the general problem of timeliness. The age of respondents where census response was lowest, young adulthood, is also where household membership and usual residence change most. On top of this, most administrative data collection is for individuals, not households. If administrative data on members of a household are collected at different times, and so have different addresses attached, it is hard to tell whether this represents a change in household structure or a move of the whole household to a new dwelling. In the 2013 Census, 21 percent of individuals indicated less than one year's residence at their current address, so mobility is a significant issue. The importance of time lags in updating administrative address information is a major reason that the statistical models for address accuracy were important in adding administrative households (see section 3.2.4).

Household structure uncertainty affects derived measures of household composition, family types/ sizes, extended family type, density of occupation, occupancy type vs family type, sole parents, child dependency status and absentees.

3.5.3 Variation in concept and instrument

Administrative records contain information that is necessary for the administration of laws. They are not influenced by social, demographic and economic thinking which usually inform social surveys and influences the content of a census.

Thus, administrative data on nominally the same construct as a Census variable may not measure the same thing. The administrative variable may be targeting the same concept but with a different question, or it may be targeting a different concept. For example, taxable income as collected by the Inland Revenue is not the same as personal income as collected by the Census. Ethnicity is the same concept in administrative data as in the Census, but, for example, data from education and health sectors is not collected with the same prompts and has a lower proportion of ethnicity recorded and output at the more detailed Levels 3 and 4.

3.6 How will the mitigations tend to affect fitness for use?

The panel will release a second report where we will consider the fitness for use of each Census variables, based on all the information available. For this interim report, we are giving guidelines on how the mitigation methods should generally be expected to affect the usefulness of Census data.

3.6.1 Single variable national and regional summaries

National and regional counts of variables present in the 2013 Census and administrative data or imputed are generally likely to be of good quality. In particular, comparisons with dual system estimation indicate the mitigation strategies have dramatically reduced the total undercount for Māori and Pacific Peoples and for young adults.

Variables that are not present in alternative data sources and not reliably imputable will not have benefited from the mitigation strategies. For example, questions on activity limitations and housing quality are new to the census in 2018 and have no alternative data source to obtain the data where census could not.

Another notable example is iwi affiliation. In April 2019, Stats NZ announced this variable would not be a census output, and the panel endorses the decision. There does not appear to be a robust or reliable way to address missing iwi data in Census 2018 given that:

- Iwi administrative data are sparse; iwi data are only regularly collected by the Ministry of Education, with some data also collected by Corrections and NZ Police. The data that do exist are of poorer quality than the census.
- At the aggregate level, there is very significant inter-censal change in iwi identification. Thus, it would be difficult to justify the use of an individual's 2013 census response to replace their missing 2018 response.
- There were significant changes to the iwi classification in 2017 which saw the inclusion of a number of iwi and iwi-related groups for the first time. For those iwi, no prior census data exists.

Other variables where mitigation was not possible include the new questions on activity limitations and housing quality.

3.6.2 Single variable summaries for small areas

Stats NZ examined the population counts for meshblocks (Dowrick & Johnstone, 2019) and commissioned a sensitivity analysis of small-area summaries to the modelling thresholds used in deciding which individuals were added from administrative data (Dot Loves Data, 2019).

These reviews show that most meshblocks and most SA2s have good quality Usual Resident counts, but for some the quality is either lower, or is unknown.

Adding individuals into administrative households and adding them in to meshblocks (but not into specific dwellings within meshblocks) will likely have improved meshblock population counts in most cases, but some additions will have been incorrect. For example, a very small number of individuals were added from administrative files to meshblocks that should be empty.

Some non-private dwellings (e.g., university halls of residence) were not enumerated in the Census, and Auckland Central has 7 of the 40 most affected meshblocks (Dowrick & Johnstone, 2019). Furthermore, some non-private dwellings were enumerated but incorrectly placed in a different meshblock than in 2013. The total number of people involved is small, but the impact on a very few small-area totals will be large.

The small areas (SA2s) whose populations have most sensitivity to the modelling thresholds are those with the largest fractions of Māori, Pacific Peoples, and young adults. They also tend to have more renters, more people born overseas, and higher smoking rates, and are at the worse end of the neighbourhood deprivation index. These small areas and the people who live in them, may be of particular interest for policy reasons, and it is important to be aware that the counts are less reliable. They are also the areas where accurate Census enumeration would always be expected to be most difficult.

As we will discuss further in our next report, analyses for lower-response SA2s rely heavily on administrative data or imputation and will have high levels of missing data for variables where mitigation was not possible.

These same small areas with less reliable population estimates will also provide less reliable information for variables not available in administrative sources, such as disability and housing quality.

3.6.3 Single variable summaries for subpopulations

Given the distinct characteristics of the households of New Zealand's ethnic communities, statutory obligations to understand whānau, and the impact of ageing on household structure, the loss in the capacity to match individuals within dwellings to establish households and families is significant. The impact is even more pronounced for many ethnic communities who because of the small, generally inadequate sample size of household surveys, will lose some of the limited visibility that they previously had in official statistics.

As noted above, Māori response rates to the census were lower than the population average. Consequently, the variables that are obtained solely in the census were not obtained for the large share of administrative Māori records, except for those where alternative sources could be used. For Māori, and smaller ethnic communities, the range of lost information is therefore much higher. For young adults, Māori, and Pacific ethnic groups, a much greater fraction of the data will be taken from administrative sources or will be imputed than for the whole New Zealand population. For these subpopulations, the quality of the administrative or imputed data will be correspondingly more important. For many of the key variables, administrative data sources do not exist. This has also occurred in 2018 for several key census topics, particularly occupation, languages spoken, unpaid activities, smoking, and disability. For dwellings, there are few alternative sources of information on tenure, housing quality standards, and mortgages.

There are some variables that will also be less stable over time for subpopulations, and so the timeliness of administrative sources will be more critical. For example, young adults are more likely to change education status and educational qualifications than older adults, and Māori are more likely to change ethnic identification. Education data provides virtually complete information on qualifications gained in New Zealand, and tertiary enrolment, up until December 2017. The main gap is qualifications gained overseas.

3.6.4 Evaluating time trends

An important use of the New Zealand Census is estimating and describing trends over time, whether in age structure, ethnicity, use of te reo, home tenure, or smoking. The missing enumerated data and the use of administrative data and imputation to replace it make the

underlying structural process behind the 2018 Census less comparable to previous data. Time trend estimates including 2018 data will be less reliable than in earlier periods – even when it is the 2018 value that is more accurate.

For some variables the overall population distribution is relatively stable, but a key use of the census data is in studying the characteristics of those who have experienced changes. For example, the research suggests that the ability to speak te reo Māori is unlikely to be a 'fixed' trait for a significant number of Māori. Only a small proportion of Māori speak te reo as their first or home language[15]; for most it is acquired through formal learning and the level of fluency can change over the lifecourse. Thus, using an individual's 2013 census response as a replacement for their missing 2018 te reo Māori response is questionable. Reliable census information is particularly important because the census is the only data source that provides a consistent time series of the number and proportion of te reo speakers nationally and sub-nationally. Similarly, smoking behaviour is fairly stable over time, but the Census should have been a key source of information about the populations whose habits have changed.

Stats NZ has used the magnitude of departure from time trends as a component of their assessment of data quality. In our second report the panel will comment on these assessments for particular variables.

3.5.5 Modelling relationships between variables

Relationships between variables will be diluted by the use of administrative data that were not collected at the same time or with the same questions as Census data, and imputed data that were not collected for the same individual.

As with the related cases of subpopulation characterisation and time trends, getting good data for most of the population will often be insufficient - the primary value of the Census in statistical modelling is the ability to get rich data on small groups of people. In particular, modelling related to social policy or wellbeing will need good quality data on precisely the subset of the population where the 2018 Census data quality is least reliable. To some extent these relationships can be studied in the IDI, but not all variables are available there, and they were not measured simultaneously.

The low linkage error between Census data and the IDI spine means that Census data will still be valuable for research use, but it will be less valuable than the research and policy community would have hoped.

3.5.6 Comment on household data

In our second report, the panel will comment in more detail about the quality of the household data. At this point we note that the mitigation approaches are likely to be less successful for household data because:

[15] In Census 2013, 21.7 percent of Māori who responded to the language question reported that they could converse about a lot of everyday things in te reo. Te Kupenga 2013, the nationally representative Māori Social Survey, found that only 2.6 percent reported speaking te reo as the main language at home and 8 percent reported that te reo was the first language that they learned and still understood.

- accurately constructing household data from administrative sources is challenging, and is especially difficult for subpopulations where people move frequently and for those who do not have a single, well-defined 'usual residence'
- 357,300 people added from the administrative data sources were not placed in a household at all. These were more likely to be young adults, and more likely to be Māori or people from Pacific ethnic groups.

It is worth noting that previous censuses used imputation to fill in age, sex, and number of people for non-responding dwellings. The households added to the 2018 data from administrative sources are likely to be more accurate than imputed households in past years.

3.5.7 Calibrating future surveys to reduce non-response bias

Census data are used by government, academic, and business survey researchers to calibrate (re-weight) survey responses to be more representative of the entire target population. The typical data used for this purpose are population totals by age, sex, Level 1 ethnicity, and region or Territorial Area/Local Board.

The data will be reliable for calibration of surveys at least for the population as a whole and for areas down to the TA/Local board level. It is not currently clear how well calibration will perform at smaller geographical scales.

4 Māori descent

This section discusses both Māori descent data for electoral purposes and the general Māori descent output data.

4.1 Māori descent counts for electoral purposes

4.1.1 The use of Māori descent data for electoral purposes

Under the Electoral Act 1993[16], census counts are used as part of the statutory formula to determine electoral populations and the number of General and Māori electorates. The calculation of the Māori electoral population (MEP) is determined by multiplying the electoral Māori descent usually resident population count (referred to here as Māori descent electoral) by the percent of enrolled Māori voters choosing the Māori roll.[17] The General electoral population (GEP) is the census usually resident population count less the MEP. The number of South Island general electorates is fixed at 16, and dividing the South Island GEP gives the South Island quota (the average size of a South Island general electorate). As the populations for each electorate should be approximately equal, the South Island quota is used to determine the number of Māori and General electorates (for more detail, see Statistics NZ, 2013).

The Electoral Act 1993 defines a Māori person as a: "... *person of the Māori race of New Zealand; and includes any descendant of such a person*". The Māori descent question in the census precedes the question on iwi affiliation. In 2018 it asked *Are you descended from a Māori (that is, did you have a Māori birth parent, grandparent or great-grandparent, etc)?* The response options were: No; Yes; and, Don't know.

4.1.2 Census non-response

As noted in section 2, there are various kinds of census non-response. Table 4.1 shows the distribution of non-response, including Don't know for the Māori descent electoral and no Māori descent electoral counts.

[16] Accessed here: http://www.legislation.govt.nz/act/public/1993/0087/latest/DLM307519.html

[17] In the 2018 Māori Electoral Option, 52 percent of eligible voters of Māori descent enrolled on the Māori Roll and 48 percent on the General Roll. See: www.elections.nz

Table 4.1: Non-response in 2018 Census, Māori descent electoral

	Māori descent electoral		No Māori descent electoral	
	(N)	(%)	(N)	(%)
Complete non-response (no census form)	188,766	21.05	336,087	8.84
Household form only ('partial'), no individual form received	53,880	6.01	149,127	3.92
Individual form received, item non-response (descent not answered on form)	7,758	0.87	48,030	1.26
Individual form received - Don't know Māori descent	20,517	2.29	76,377	2.01
Individual form received – Māori descent answered yes/no	625,647	69.78	3,193,560	83.97
Total	896,565	100.00	3,803,190	100.00

The figures clearly show that for nearly all categories (except item non-response), the level of non-response was much higher in the Māori descent electoral count than in the no Māori descent electoral count. The difference was especially marked for complete non-response. Although not shown here, the level of complete non-response for Māori descent electoral in the 2018 census was substantially higher than in 2013 or 2006.

In 2018 it was compulsory on the online form to answer the question on Māori descent, thus individual form non-response was only possible on individual paper forms (Stats NZ, 2018a). Prior to the 2018 Census, Stats NZ expected that making the descent question mandatory online could result in a relative decline in the imputed share of the Māori descent counts prepared for electoral purposes as the majority of individual form non-responders would select No (Stats NZ, 2018a, p. 17). While item non-response to the ancestry question was low, it was more than offset by the increase in complete non-response.

The tickbox order was changed for 2018, with the options appearing in the following order: Yes, Don't know, No. Previously the order was Yes, No, Don't know

4.1.3 Imputation and administrative data approach

For electoral purposes, all of the usually resident population enumerated since the 1996 Census must be classified as being of Māori descent, or not. A missing or residual response is not allowed.

Between 1996 and 2013, statistical imputation methods were used to assign a valid Māori descent response to non-responding individuals. For the 2018 census, administrative data from the 2013 Census and Department of Internal Affairs birth records were used, for the first time, to substitute non-response. Statistical imputation, including the new CANCEIS method detailed in section 3, were only used when such data were not available.

The approach used to compensate for non-response for 2018 Māori descent electoral was as follows: where neither a Yes or No was recorded for Māori descent, data were used from the 2013 census (1st priority) and then birth records (2nd priority). If the Māori descent electoral response was still *not* Yes or No, household probabilistic imputation was used,

which involved using the Yes or No response of the person of closest age in the usual residence. If response for Māori descent electoral was still *not* Yes or No, 2018 Census iwi response was used. Thus, if there was at least one valid iwi response in Census 2018, then Māori descent was coded to Yes. The residual was addressed using CANCEIS donor imputation to find a donor with a Yes or No response.

As Table 4.2 shows, of the 896,565 total Māori descent count prepared for electoral purposes, only 69.8 percent came from received individual census forms. The remainder was sourced from the 2013 Census (15.0%) and birth records (6.3%), with the residual imputed using probabilistic and CANCEIS methods described in the methodology section above. The low share of received forms reflects the poor overall Māori collection response rate.

For the no Māori descent electoral count, 84.0 percent came from received individual census forms**.** The remaining records drew on individuals' data from the 2013 Census (7.7%), birth records (1.5%), probabilistic imputation (1.3%) and CANCEIS imputation (5.5%).

Table 4.2: Data sources and imputation used for 2018 Census Māori descent and no Māori descent electoral counts

	Māori descent electoral		**No Māori descent electoral**	
	(N)	(%)	(N)	(%)
2013 census	134,292	15.0	293,904	7.7
DIA birth records	56,628	6.3	55,950	1.5
Probabilistic imputation	21,618	2.4	50,610	1.3
CANCEIS imputation	58,383	6.5	209,160	5.5
Received individual form 2018	625,647	69.8	3,193,560	84.0
Total	**896,565**	**100.00**	**3,803,190**	**100.00**

Table 4.3 shows how the use of alternative data sources significantly impacts the final counts of the Māori and no Māori descent electoral counts, while the change in statistical imputation methods between 2013 and 2018 has very little effect. For example, for the Māori descent count, taking only the information from received individual forms and imputing the remainder using the 2013 imputation methods produces a count of 819,633. However, adding available information from the 2013 census and birth registrations to the received forms, and then imputing the rest using 2013 imputation methods produces a substantially higher count of 895,565. Where alternative sources are used, and only the imputation method is changed, the difference in the count is only ~1,200.

An opposite effect can be observed for the no Māori descent count in that the addition of 2013 census and births records decreases the no Māori descent count by around ~76,000, compared to the use of 2013 imputation methods alone. Likewise, whether 2013 or 2018 imputation methods are used makes little difference to the count.

Table 4.3 Comparison of derivation methods for Māori descent and no Māori descent electoral applied to Census 2018 data

Description of method		Māori descent electoral	% of total population
Use only the information in 2018 census responses for Māori descent	Impute all remaining using 2013 method	819,633	17.4
Where available, add information from 2018 census and DIA births to derive Māori descent	Impute all remaining using 2013 method	895,341	19.1
	Impute all remaining using 2018 method	896,565	19.1
		No Māori descent electoral	% of total population
Use only the information in 2018 census responses for Māori descent	Impute all remaining using 2013 method	3,880,125	82.6
Where available, add information from 2018 census and DIA births to derive Māori descent	Impute all remaining using 2013 method	3,804,414	80.9
	Impute all remaining using 2018 method	3,803,190	80.9

4.1.4 Change over time

The Māori descent electoral count in 2018 (896,565) represents an increase of 18.7 percent since 2013. This far exceeds the 4.7 percent increase in the electoral Māori descent count between 2006 and 2013. Proportionately, the Māori descent electoral population share also increased substantially, from 17.8 percent of the total population in 2013 to 19.1 percent in 2018.

The number of people counted as No Māori descent increased by 9.1 percent to 3,808,190 in 2018.

Table 4.4 compares the Māori descent and no Māori descent electoral counts in 2018 with 2006 and 2013. The comparability over time will have reduced because of the changes in methodology, and the need to account for a substantially increased number of complete non-responses. Detailed sensitivity analysis was undertaken to identify the extent to which the inclusion of administrative records affected the number of Māori and General electorates and electorate size (next section).

Table 4.4: Māori descent and no Māori descent electoral counts and intercensal change, 2006, 2013, 2018

	2006	2013	2018	2006-13 (% change)	2013-18 (% change)
Māori descent electoral	721,431	755,598	896,565	4.7	18.7
No Māori descent electoral	3,306,516	3,486,450	3,803,190	5.4	9.1

4.1.5 Consistency of Māori descent electoral responses across different sources

Given the extensive use of alternative data to compensate for Māori descent electoral non-response, we consider the consistency of individuals' designations of Māori descent in the Census 2018 context versus the 2013 Census and DIA birth registrations. For individuals who returned an individual form with a valid Māori descent response in 2018, their records were matched with their 2013 census or birth records. Tables 4.5 and 4.6 show the level of consistency for each available response (Yes, No, and Don't know) between the 2018 forms and the alternative data sources.

For matched 2013 and 2018 Census responses, the level of consistency was higher for No Māori descent than Yes. Thus, 97.8 percent of those who responded No to the Māori descent question on the 2013 Census form also responded No in 2018; for the Yes responses on the 2013 Census form, the share also reporting Yes in 2018 was 92.9 percent. Although Don't know is not a valid response for electoral purposes, we show the responses here as provided in both censuses. Table 4.5 shows that only one in four of the matched individuals recorded as Don't know in 2013 also recorded Don't know in 2018.

Table 4.6 shows a lower level of consistency for Yes responses to the Māori descent responses in the births records and 2018 Census (88.9%), but a similar pattern of consistency for No responses (96.9%).

Table 4.5. Consistency of Māori descent responses, 2018 and 2013 census

	2018 Census			
2013 Census	Yes	No	don't know	Total
Yes	418,300	21,600	10,600	450,500
	92.9	4.8	2.4	100.0
No	17,900	2,300,700	33,600	2,352,300
	0.8	97.8	1.4	100.0
Don't know	12,400	39,500	17,300	69,200
	17.9	57.1	25.0	100.0
Census 2013 total	448,600	2,361,800	61,500	2,871,900

Table 4.6. Distribution of Māori descent responses, 2018 census and DIA birth registrations

	2018 Census			
DIA births	yes	no	don't know	Total
Yes	77,300	6,600	3,100	87,000
	88.9	7.6	3.6	100.0
No	3,500	197,100	2,800	203,400
	1.7	96.9	1.4	100.0
Not sure	2,500	4,900	1,600	9,000
	27.8	54.4	17.8	100.0
Births total	83,400	208,600	7,400	299,400

4.1.6 Sensitivity analyses

The methods for determining the Māori descent and non- Māori descent shares need to be precise enough at electorate level to meet the tolerances specified in the electoral legislation. They have been applied since 1996, and there have been no situations when the adjustments for non-response by the Government Statistician would have altered the number of Māori seats.

Given that the number and average population size of General and Māori electorates depends on both the census usually resident population count, and the Māori descent census electoral count, it is crucial to understand the impacts of decisions on which administrative addresses to include or exclude in the 2018 Census dataset.

Stats NZ commissioned the data science company, Dot Loves Data, to undertake sensitivity analyses of the impact of adding administrative records (from the IDI-ERP) to census responses to determine the number and size of General and Māori electorates. Dot Loves Data's approach was to use Stats NZ's rating of the quality of the information about an individual's address meshblock to determine how different quality thresholds ('alpha') impacted the number and size of electorates.

There were 422,000 administrative records that could be added. Setting the quality threshold at the minimum (alpha=0) results in the addition of all 422,000 records, regardless of the quality of their address. Setting the threshold at the maximum (alpha=1) results in the addition of no additional records. Setting the threshold at the middle point (alpha=0.5; the threshold chosen by Stats NZ) results in the addition of 357,000 records. We note the distribution of quality among addresses is skewed, that is, there are far more records rated as having high quality address information than low quality address information.

As the threshold for adding addresses lowers, the number of people added to the population increases. This results in both: (i) a larger census usually resident population; and (ii) a probable greater share of the census usually resident population being of Māori descent. Those of Māori ethnicity – which relates strongly but not perfectly to Māori descent – were disproportionately missed from the census enumeration, and so are more likely to be in administrative records being considered for inclusion.

The Dot Loves Data sensitivity analysis reached three main conclusions. First, the number of Māori electorates would stay at seven regardless of decisions about the inclusion of administrative records. If NO administrative records were added, the estimated number of Māori electorates would be 6.54, given (i) the Māori descent electoral population, (ii) the percentage of Māori opting for the Māori electoral roll (53%[15]), and (ii) the South Island quota, as described above. Note that 6.56 rounds to 7 (as do all numbers between 6.5 and 7.499). Conversely, if ALL 422,000 administrative records were added, the estimated number of Māori electorates would be 7.32 (which also rounds to 7). At the threshold chosen by Stats NZ (alpha = 0.5), the estimated number of Māori electorates would be 7.23 (again rounding to 7).

The panel considers this to be a very robust finding for the following reason. Adding no administrative records in 2018 results in an increase from the 2013 Māori descent electoral census usually resident population by ~14,500 (from 755,598 to 770,100). This represents a 1.9 percent increase, which is far lower than the increase from 2006 to 2013 (4.7%). Moreover, in order for the number of Māori electorates to be **less than seven** (i.e., less than 6.5 with rounding) the increase in the Māori descent census usually resident population since 2013 would have to be less than or equal to 1 percent, which the panel consider to be an unrealistically low increase, given demographic trends. Thus, it is very unlikely that the number of Māori electorates could be calculated to be less than seven.

Similarly, for the number of Māori electorates to be **greater than seven** (i.e., >=7.5 with rounding), would require an increase in the Māori descent census usually resident population of at least 24.5 percent since 2013. The panel considers an increase this size to be unrealistically large. **As such the panel is confident that the number of Māori electorates should be seven and is confident that this conclusion can be drawn with or without the addition of administrative records to the census.**

Second, the panel agrees with the conclusion that for most thresholds for inclusion of administrative records, the number of General electorates would be 65 (i.e., 16 electorates in the South Island as designated by the Electoral Act 1993, and 49 electorates in the North Island as determined by the North Island GEP). At alpha = 0.5 (the threshold chosen by Stats NZ), the estimated number of seats in the North Island would be 48.58 (which rounds to 49). For the slightly stricter threshold of alpha = 0.6, the estimated number of seats in the North Island would be 48.5 (which rounds to 49). For all alpha < 0.5 (i.e., lower thresholds for the inclusion of administrative records), the estimated number of seats in the North Island would be >48.6 (i.e., rounding to 49 in all cases). However, for all thresholds where alpha > 0.6 (which increasingly restrict the inclusion of administrative records), the estimated number of seats in the North Island would be <48.5 (i.e., rounding to 48 in all cases). The Electoral Act 1993 enables the Government Statistician to exercise a degree of discretion; this would include the selection of alpha < 0.5 or <0.6.

Thirdly, the methods used to determine the Māori descent and no Māori descent electoral counts are sufficiently precise to meet the +/- 5 percent allowance of adjustment of quota, as specified under the Electoral Act 1993.

The most obvious impact would be the inclusion (or not) of a 49th seat in the North Island, which will clearly affect electoral boundaries. It is the role of the Representation Commission to decide where this will go. These sensitivity analyses can support this work and indicate where working more narrowly within the 5 percent tolerance limits set by the

legislation can manage potential uncertainty about the effect of differences in the impact of non-response across New Zealand on electoral calculations. We note that the current Auckland Central electorate, which seems to be particularly sensitive to the addition (or failure to add) administrative records, may need to be managed in such a way.

Additional points to note about the sensitivity analysis

The panel notes some minor limitations of the Dot Loves Data analysis.

First, sensitivity tests around thresholds for the inclusion of administrative records were limited to those records (n=422,000) considered of insufficient quality to be included directly in households, but of sufficient quality to potentially be included in meshblocks. There was no consideration of the roughly 141,000 administrative records considered of sufficient quality to be placed into households.

The panel considers this a minor limitation as those records with good enough address information to be placed in a household would – necessarily – have good enough address information to be placed in the meshblock containing the household. Note that electorates are made up of combinations of meshblocks; resolution to the household level is not required. Furthermore, a decision to **not** include these 141,000 records would have likely resulted in a population with unrealistically low Māori descent census usually resident population growth, as explained above.

Second, Dot Loves Data had to impute Māori descent for around half of those with alpha <0.5 (around n=32,000 of the n=65,000 with alpha<0.5 needed imputation), because these records had not yet been through census processing. The panel also considers this a minor limitation given that Stats NZ is not considering using a threshold of alpha <0.5, as this would result in the inclusion of records with low quality addresses. As such, Dot Loves Data's imputation of Māori descent for those with alpha<0.5 is not likely to have an impact on electorate calculations.

4.2 Māori descent (output)

Māori descent is a Priority 1 variable. It provides the subject population for the iwi variable (which will not be released as official statistics from Census 2018) and the raw data from which the Māori descent electoral count is derived.

In previous censuses, non-response to the Māori descent question was close to 10 percent. The mitigation responses employed for Census 2018 reduced non-response to zero but also created a major break in the Māori descent time-series, with the count being significantly larger (30.1%) in 2018 than in 2013.

The major usage of the Māori descent variable is to provide data for electoral purposes, which has already been described in detail above. This section briefly outlines quality issues with the Māori descent (output) variable that were not discussed in the foregoing section.

The panel rates the Māori descent output variable as being of high quality. However, because Māori descent data are usually only published in iwi tables, and Stats NZ will not produce official iwi output for Census 2018, the Māori descent output variable is unlikely to be widely used in Census 2018 data releases.

4.2.1 Distribution of Māori descent responses

Distinct from the Māori descent electoral variable, Māori descent (output) includes Don't know as a valid response. The distribution of responses to the Māori descent question in the Census 2018 dataset are shown in Table 4.7, along with comparator data for 2006 and 2013 censuses.

Table 4.7. Māori descent counts and intercensal change, 2006, 2013, 2018

	Numbers			%		
	2006	**2013**	**2018**	**2006**	**2013**	**2018**
Māori Descent	643,977	668,721	869,850	16.0	15.8	18.5
No Māori Descent	2,917,311	3,065,487	3,715,050	72.4	72.3	79.0
Don't know	78,774	87,234	114,855	2.0	2.1	2.4
Total Stated	3,640,062	3,821,445	4,699,755			
Response Unid.	3,111	2,694	0	0.1	0.1	0.0
Not Stated	384,774	417,906	0	9.6	9.9	0.0
Total Population	4,027,947	4,242,048	4,699,755	100.0	100.0	100.0

At 18.5 percent, the Māori descent population represents a significantly higher proportion in 2018 than in recent censuses, due to both an absolute increase in size, but also complete coverage for the Māori descent variable reducing non-response. In 2006 and 2013, the Māori descent share of the total population was 16.0 percent and 15.8 percent respectively. However, as a proportion of the total stated, it was 17.7 percent and 17.5 percent respectively, much closer to the 2018 descent share.

In absolute terms the Māori descent count in 2018 represented a 30.1 percent increase on 2013, compared with just 3.8 percent between 2006 and 2013.

The number of individuals recorded as being of Māori descent increased in all regions in 2018, with Northland having the largest increase (44.3%) between 2013 and 2018, and West Coast having the smallest (15.9%).

The number of people counted as No Māori descent increased by 21.2 percent from 3,065,487 in 2013 to 3,715,050 in 2018.

Direct time series comparison is not advisable because of the change in methodology with the addition of alternative government data.

Although not shown here, there was a large increase in counts of people who had no Māori descent, but recorded Māori ethnicity. Counts of people with no Māori descent, but either only Māori ethnicity, or Māori in combination with at least one other ethnicity, increased by 248.7 percent and 205.6 percent respectively between 2013 and 2018, albeit from a small base[18]. These differences may partly be a consequence of the two variables being imputed independently of each other when both are missing.

[18] For no Māori descent, Māori ethnicity only, the number increased from 1,410 to 4,917; for no Māori descent, Māori ethnicity combined, the number increased from 2,802 to 8,562.

4.2.2 Data sources used

As Table 4.8 shows, of the **Māori descent (output) count**, **71.9 percent** came from received individual census forms. The remainder was sourced from the 2013 Census (14.0%) and birth records (6.1%), with the residual imputed using CANCEIS and probabilistic methods.

For the **no Māori descent count**, **86.0 percent** came from received individual census forms**.** The remaining records drew on individuals' data from the 2013 census (6.8%), birth records (1.3%) and statistical imputation (5.9%).

For the **Don't Know Māori descent count, 83.1 percent** came from received individual forms. The remainder was drawn from the 2013 Census (10.4%), administrative data (4.3%), or imputed (2.1%).

Table 4.8. Data sources and imputation used for 2018 Census Māori descent output

	Māori descent		No Māori descent		Don't know Māori descent	
	(N)	(%)	(N)	(%)	(N)	(%)
2013 census	121,908	14.0	254,352	6.8	11,949	10.4
DIA birth records	52,626	6.1	47,067	1.3	4,995	4.3
Probabilistic imputation	20,169	2.3	50,610	1.4	2,421	2.1
CANCEIS imputation	49,500	5.7	169,461	4.6	-	-
Received individual form 2018	625,647	71.9	3,193,560	86.0	95,493	83.1
Total	869,850	100.0	3,715,050	100.0	114,855	100

4.2.3 Imputation approach

The method used to address missing data in 2018 differed from that used in 2006 and 2013. In those censuses, missing data were treated as non-response.

In 2018, where neither a Yes, No or Don't know was recorded for Māori descent, data were used from Census 2013 (1st priority) and then birth records (2nd priority). If the electoral Māori descent response was still *not* Yes, No or Don't know, household probabilistic imputation was used, which involved using the Yes, No or Don't know response of the person of closest age in the usual residence. If response for electoral Māori descent was still not Yes, No or Don't know, 2018 Census iwi response was used. Thus, if there was at least one valid iwi response for an individual in 2018 Census, then Māori descent was coded to Yes. The residual was addressed using CANCEIS donor imputation to find a donor with a Yes, No or Don't know response.

4.2.4 Consistency of Māori descent responses across different sources

Residual missing data was reduced to zero percent once Stats NZ added in historical census data, administrative data, and imputed missing variables. On that basis, the coverage could be rated high. However, the quality of the data being used to substitute for missing data, and the accuracy of the information used, also matters. A relatively high share (28.1%) of

the Māori descent count was drawn from outside of Census 2018. Thus, the use of historic data to replace missing values, or Don't Know responses, depends on the assumption that Māori descent identification is relatively stable over time.

The Stats NZ Metric 1 quality rating for data source and coverage tries to assess this by computing a weight that measures the consistency of individuals' Māori descent responses in the 2018 Census with their responses in the 2013 Census or their birth record. The source weight is essentially a consistency measure and is derived by comparing exact Yes and No and Don't Know responses from 2018 Census to exact Yes and No and Don't Know responses in the 2013 Census and birth records. The 2018 Census weight was 1.00 (as the reference point), the 2013 Census weight was 0.95, and birth records was 0.92. The weight for each source is then multiplied by its proportion (of the overall count) to give a score contribution that could sum to 1.00. The total score given by Stats NZ was 0.97 (high). We note that comparisons can only be made for individuals with both a 2018 and 2013 Census form or birth record. This will be less than 72 percent for the Māori descent (count output) population (i.e., the share for whom a 2018 form was received).

Table 4.3. Stats NZ quality rating calculation table

Source	Total(%)	Weight	Score contribution
2018 Census form	83.30	1.000	0.83
2018 Census (missing from individual form)	0.00	1.000	0.00
Historic (2013 Census)	8.26	0.950	0.08
Admin data sourced	2.23	0.920	0.02
Probabilistic imputation (from other variables or members of the UR household)	1.56	0.800	0.01
CANCEIS nearest neighbour (2018 Census form)	3.87	0.600	0.02
CANCEIS nearest neighbour (2018 Census form, missing from individual form)	0.00	0.600	0.00
CANCEIS nearest neighbour (Historic donor)	0.60	0.570	0.00
CANCEIS nearest neighbour (Admin donor)	0.00	0.552	0.00
CANCEIS nearest neighbour (probabilistic donor)	0.19	0.480	0.00
CANCEIS nearest neighbour (donor has missing value)	0.00	0.000	0.00
Missing	0.00	0.000	0.00
Overall quality rating			**0.97**

Because the above quality rating is computed for the overall Māori descent output variable, and the source distributions for the Māori descent response categories vary significantly, we computed the quality ratings separately for the Yes, No and Don't Know categories. Doing so produced the following quality ratings – 0.96 for Māori descent, 0.98 for no Māori descent, and 0.88 for Don't Know.

4.2.5 Final comments

There has been growing interest in the use of Māori descent as an alternative or complement to Māori ethnicity data. It is often used alongside iwi data, although iwi data will not be produced as official counts from Census 2018 due to quality issues.

The coverage for the Māori descent output variable has been improved through the mitigation methods employed by Stats NZ, with non-response reduced to zero through the

addition of administrative data and imputation. However, the change in methods has resulted in a break in the Māori descent time-series.

Caution is also needed when making comparisons with iwi data. For example, in 2018, 58.6 percent of respondents of Māori descent indicated that they knew the name of their iwi. This is perhaps unsurprising given that missing iwi data could not be plugged with alternative government data. If we only look at those individuals who provided a valid response to the Māori descent question on the census form, the share reporting at least one iwi was 79.8 percent. This is more in line with previous census iwi shares[19].

[19] In the 2001, 2006 and 2013 censuses, the share of the Māori descent population reporting at least one iwi ranged from 80 percent to 83 percent (for those providing a valid response to the iwi question).

5 Ethnicity

5.1 Background

Ethnicity is one of the few variables that is specifically required by legislation to be collected in the New Zealand census and is designated a 'Priority 1' variable by Stats NZ. The census ethnicity variable uniquely enables the entire population to be grouped by the ethnic affiliations of its members. Ethnicity is a key public policy variable in New Zealand and ethnic inequities are a major societal issue. Census ethnicity data are required for the design, delivery and monitoring of many Government policies and services, and are also crucial for communities' ability to plan and to monitor outcomes and inequities by ethnicity. In this section, in addition to assessing the quality of the ethnicity variable, we have also considered the impacts of the low response rates on the availability of characteristic information for specific ethnic populations.

The *Statistical standard for ethnicity* ('the Standard') outlines the definition of ethnicity and the recommended standard approach to collection and recording, while the hierarchical *Ethnicity New Zealand standard classification* (ETHNIC05) details the four ethnicity levels (Statistics New Zealand, 2017[20]). Level 1 has six "major ethnic groups"[21]: European; Māori; Pacific Peoples; Asian; Middle Eastern/Latin American/African; and, Other. With the exception of Māori, the other level 1 categories are all broad groupings of specific ethnic groups included at levels 2, 3, and 4 of the classification. Level 4 has 185 categories, of which five are residuals. Māori is the only ethnic group that appears at all four levels.

As high-level aggregations, the level 1 ethnic categories obscure a great deal of diversity between the constituent groups. The Asian category subsumes Chinese, Indian, South East Asian, and other Asian groups (Vietnamese, Cambodian, Sri Lankan, Japanese, Korean), while the Pacific Peoples category includes Samoan, Tongan, Cook Islands Māori, Niuean, Tokelauan, Fijian, and other Pacific peoples. The European grouping identifies 30+ distinct nationalities, many of which comprise New Zealand-born along with long-term residents and more recent arrivals. Immigration and settlement policies rely on high quality census data to distinguish these different parts of ethnic communities with respect to their previous country of birth and characteristics. The Middle Eastern/Latin American/African (MELAA) category is the most rudimentary of the level 1 categories, aggregating ethnic groups across no less than three continents.

High quality census data are particularly important for groups that have a special status, rights or interests. For Māori, Te Tiriti o Waitangi creates distinct obligations and te reo Māori is the country's second official language. The Cook Islands and Niue are part of the Realm of New Zealand (New Zealand is officially responsible for defence and foreign affairs

[20] There are multiple versions of the classification. For Census 2018, Stats NZ is using CEN.ETHNIC05_2018CENS

[21] http://aria.stats.govt.nz/aria/#ClassificationView:uri=http://stats.govt.nz/cms/ClassificationVersion/l36xYpbxsRh7IW1p

in those countries), and Tokelau is a dependent territory. Samoa was administered by New Zealand from 1919 through a League of Nations mandate and then was a United Nations Trust Territory until the country received its independence on 1 January 1962.

The Standard defines ethnicity as the "ethnic group or groups that people identify with or feel they belong to". Ethnicity is intended as a measure of self-perceived cultural affiliation and people can belong to more than one ethnic group. Stats NZ records up to six ethnicities per person at level 4 of the classification. The 2018 Census ethnicity question was the same as that used in recent censuses.

The Standard recognises that ethnicity is not a fixed characteristic and that people can and do change how they self-identify their ethnicity over time and by context (also see Didham, 2016; Caron-Malenfant, 2014). Some of this 'ethnic mobility' reflects changes in ethnic affiliation, including changes in the reporting of the number and combination of ethnic groups, while others may reflect a change in collection from next-of-kin (e.g., a parent or carer) to self-report (Cormack & Harris, 2009). The assessment of the quality of ethnicity needs to recognise such inherent limitations.

5.2 Major uses of census ethnicity data

Under the 1975 Statistics Act, ethnic origin is one of the 'particulars' that is required to be collected from every individual in the census (alongside name, address, sex, and age). There are a number of major uses of census ethnicity data, summarised briefly below.

Human rights

Ethnicity data are necessary for the Government to report on various obligations under UN human rights conventions, as well as to monitor human rights locally. The *Human Rights Act 1993* specifies 'ethnic or national origin' and 'race' as unlawful grounds for discrimination.

Te Tiriti o Waitangi

Ethnicity data are important for monitoring the Crown's Treaty obligations to Māori (including iwi). Ethnicity data have been critical for measuring the impacts of Crown policies and programmes on Māori, identifying inequities and monitoring over time. Ethnicity data are often used to generate evidence in support of claims to the Waitangi Tribunal about Treaty breaches.

Policy development and monitoring/reporting by ethnicity

Government agencies use ethnicity data when developing policies and to monitor the impacts of these policies. Census ethnicity data are the basis for monitoring social, economic, and cultural outcomes by ethnicity and ethnic inequities over time. Government agencies routinely report social and economic outcomes by ethnicity. For example, District Health Boards are required to report specific outcomes by ethnicity as part of routine performance monitoring. Researchers, academics, non-governmental organisations, and communities also use ethnicity data to monitor and report the impact of policies, programmes and interventions by ethnicity. Census data are also crucial for monitoring changing ethnic diversity at national, regional, and local levels.

Targeting resourcing, funding, and services

Ethnicity data are used to target resources, funding, and services (e.g. the Population-Based Funding Formula, see section 2.1 of this report), to model the impacts and costs of policy changes, and to forecast expenditure on services for particular groups.

Community advocacy

When community and ethnic groups, or NGOs who advocate for them, bid for resources and grants, they will routinely make use of census ethnicity data to illustrate the need for funding and intervention.

Population estimates and projections

Census ethnicity data provide the population counts that are used as the base to produce intercensal population estimates for Māori. These estimates are used as denominators for the calculation of ethnic specific rates as well as base data for European, Māori, Asian, and Pacific population projections.

Granular ethnicity data

The census is one of the very few data collections that has detailed level 4 ethnicity data available for the whole population. Were ethnicity required only for level 1 of the classification, a sample survey rather than full population census would suffice. While birth and death registrations have moved to collecting data at level 4, this only apply to births and deaths registered after 1995 when collection processes were aligned with the Standard.

5.3 Census response and non-response

Overall, 84.1 percent of ethnicity data in the 2018 census dataset came from individual census forms (paper and online combined). A further 8.3 percent were sourced from 2013 census data, and 6.3 percent of records were from administrative data. The remaining 1.2 percent were derived from various forms of imputation.

In 2018, it was mandatory to complete the ethnicity question on the online form, so non-response on individual forms only derives from paper forms.

The numerical dominance of the level 1 European group means that the aggregate ethnicity figures obscure significant inter-ethnic variation. Table 5.1 shows that, among the level 1 categories, the percentage of census ethnicity data sourced from a 2018 individual census form was lowest for Māori and Pacific peoples (70.87% and 67.55% respectively), and highest for European ethnic groups (88.48%). At level 2, the share of ethnicity data sourced from the 2018 census was highest for Other European (89.50%), Chinese (89.13%), and New Zealand European (89.12%). Excluding the ‘nfd’ level 2 categories, the share of 2018 census sourced data was lowest for Tongan (63.37%) and Fijian (66.09%).

Table 4: Distribution of data sources for level 1 ethnicity data (grouped total response)

Data source / Level 1 Ethnicity	**European**		**Māori**		**Pacific Peoples**		**Asian**		**MELAA**		**Other ethnicity**	
	Count	*%*	*count*	*%*	*Count*	*%*	*count*	*%*	*count*	*%*	*count*	*%*
2018 Census form	2,918,034	88.48	549,807	70.87	257,805	67.55	591,414	83.58	56,037	79.67	47,889	82.49
Historic (2013 Census)	223,845	6.79	116,592	15.03	63,120	16.54	46,134	6.52	5,856	8.33	5,184	8.93
Admin data sourced	123,885	3.76	101,334	13.06	55,368	14.51	55,551	7.85	6,735	9.58	4,491	7.74
Probabilistic imputation (from other variables or members of the UR household)	10,641	0.32	4,311	0.56	2,661	0.70	5,286	0.75	555	0.79	222	0.38
CANCEIS nearest neighbour (2018 Census form)	18,675	0.57	2,721	0.35	1,752	0.46	8,010	1.13	999	1.42	210	0.36
CANCEIS nearest neighbour (Historic donor)	1,788	0.05	741	0.10	480	0.13	753	0.11	90	0.13	18	0.03
CANCEIS nearest neighbour (Admin donor)	819	0.02	303	0.04	405	0.11	345	0.05	36	0.05	39	0.07
CANCEIS nearest neighbour (probabilistic donor)	177	0.01	30	0.00	51	0.01	105	0.01	18	0.03	3	0.01
Total	**3,297,864**	100.0	**775,836**	100.01	**381,642**	100.01	**707,598**	100.00	**70,332**	100.00	**58,053**	100.01

We did not have access to detailed data showing the distribution of individual forms received at level 4. However, aggregate data show that 77 percent of European level 4 groups, 74 percent of Asian level 4 groups, and 73 percent of MELAA level 4 groups had at least 85 percent individual forms (the national total). For Pacific level 4 ethnic groups, the share of received forms was much lower at just 26 percent. Only five of the 19 level 4 Pacific ethnicities had at least 85 percent or more census responses.

Consistent with Stats NZ's 'total response' approach to outputting ethnicity data (Stats NZ, 2017), individuals are counted in every broad ethnic category with which they identify so that the total of the six main ethnic groups exceeds the subject population.

After census mitigation involving the use of historic census data, administrative data and statistical imputation, ethnicity non-response was reduced to zero. By comparison, the level of non-response in Census 2013 was 5.3 percent, which was mainly from substitute records created for 'partial' responses, with some item non-response.

While ethnicity data coverage (i.e., completeness) has been achieved for the entire Census 2018 dataset, the way that this was achieved differs by ethnic group. For Māori and Pacific ethnic groups, in particular, completeness was only possible by drawing extensively on alternative data sources. The data sources used to complete ethnicity data were (in order of prioritisation): Census 2013; DIA Births; Ministry of Education (tertiary enrolments); Ministry of Health (subset of fields from National Health Index); and, Ministry of Defence and Department of Corrections. Census 2013 data are collected at level 4 of the classification, as are DIA Births from 1995. MoE data are collected at level 3. MoH data are reported at level 2 (although the 2017 revised ethnicity data protocols for the health and disability sector require collection and recording of ethnicity data at Level 4 (MoH, 2017)).

The lack of consistency in ethnic granularity across administrative datasets has resulted in higher than usual numbers of individuals in 'not further defined' and 'not elsewhere classified' categories at levels 2, 3, and 4. For example, the 'Other nec' (not elsewhere classified) level 4 category increased from 150 in Census 2013 to 4,017 in Census 2018, an increase of over 2,500 percent, albeit from a very low base. The 'Southeast Asian nfd' category also increased by 300+ percent from 1,275 in 2013 to 6,219 in 2018. There were also significant increases in the level 2 'European nfd', 'Pacific Peoples nfd' and 'Asian nfd' categories, rendering at least this level of the classification less reliable than for previous censuses.

5.4 Imputation and administrative data approach

Table 5.2 shows the breakdown of administrative data sources used for each level 1 major ethnic groups. Of the 101,337 admin data records provided for Māori, 50,568 (49.9%) were sourced from DIA births data, 37,221 (36.7%) from MoE data, and 13.2 percent from Ministry of Health. That means that of all Māori ethnicity records, 6.5 percent were sourced from birth data reported by someone else (a parent/carer), 4.8 percent from MoE data (tertiary enrolments), and 1.7 percent from Ministry of Health. For Pacific ethnic groups, a higher share of administrative data was sourced from MoH data (26.98%, or 3.91% of all Pacific ethnicity records).

Table 5.2 also shows that a small number of ethnicity records (n=99) were drawn from data provided directly to Stats NZ by the Department of Corrections for those in prison at the time of the census[22] (for more details, see section 3.2.4 of this report); along with 795 administrative records from the Ministry of Defence for those in military establishments. Analogous data for level 2 ethnic groups are included in appendix 2.

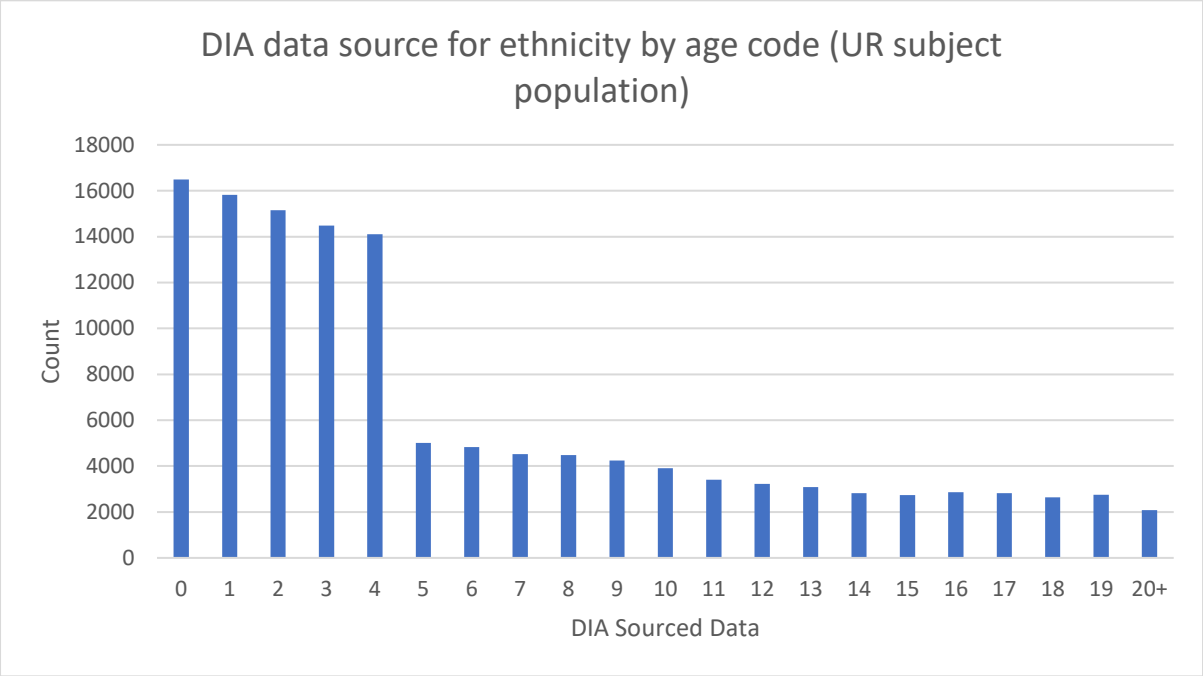
Table 5: Distribution of administrative data sources for Level 1 ethnicity data (grouped total response)

Ethnic group Level 1	Administrative data source					
	COR	DEF	DIA	MoE	MoH	Total
European	21	303	41,979	42,162	39,435	123,990
	0.02%	0.24%	33.88%	34.03%	31.83%	
Māori	48	129	50,568	37,221	13,371	101,337
	0.05%	0.13%	49.90%	36.73%	13.19%	
Pacific Peoples	24	36	26,727	13,644	14,937	55,368
	0.04%	0.07%	48.27%	24.64%	26.98%	
Asian	s	21	10,455	21,951	23,124	55,551
	s	0.04%	18.82%	39.52%	41.63%	
MELAA	s	s	1,332	1,974	3,426	6,732
	s	s	19.79%	29.32%	50.89%	
Other	6	306	387	3,345	447	4,491
	0.13%	6.81%	8.62%	74.48%	9.95%	

Figure 5.1 below shows that the majority (57.8%) of birth records used for ethnicity in the 2018 Census dataset were for 0-4 year olds who would not have had records in the 2013 census. Indeed, 75.4 percent of the DIA births data used was for children aged under ten years. This is reassuring as ethnicity at birth will have been assigned by parents/carers and it is unlikely that substantive shifts will have occurred with a shift to self-identification. There were, however, just over 2,700 19-year olds who had their ethnicity sourced from birth records and an unknown proportion will have changed their ethnic self-identification since birth.

[22] The number of prisoner records provided to Stats NZ by the Department of Corrections far exceeded 99 (see section 3.2.4), however most of the prisoner ethnicity data were drawn from existing administrative data in the IDI rather than the ethnicity data provided directly by Corrections.

Figure 5.1. DIA births data source for ethnicity by age code



5.5 Potential quality issues

Before undertaking a quality assessment of ethnicity data using the Stats NZ quality ratings, we identify some of the potential quality issues associated with the use of historic census and administrative data sources.

There are four main quality limitations that could bias responses, especially with regard to multiple ethnic affiliation. These are: reporting by others; use of outdated concepts or categories during data collection; limiting respondents' ability to report multiple ethnic affiliations; and use of output classification levels that are inconsistent with the census (i.e., ethnicity not classified to level 4).

2013 Census data

Although the ethnicity question remained the same for the 2013 and 2018 Censuses, the ethnicity data drawn from the 2013 Census is historical. Given that individual reporting of ethnicity can change over time, there is the potential that some of these data are out of date.

Administrative data

The proportion of ethnicity records sourced from administrative sources differed significantly between ethnic groupings. While only 6.29 percent of ethnicity data overall were sourced from administrative sources, the share was 13.06 percent for Māori, 14.51 percent for Pacific, 7.85 percent for Asian and 9.58 percent for MELAA. This compares with just 3.76 percent of records for European. As such, any quality issues arising from the use of administrative data will disproportionately affect non-European ethnic groups.

DIA births data

Overall, 2.0 percent of ethnicity data was drawn from DIA Births. However, the proportion was higher for Māori (6.5%) and Pacific peoples (7.0%), and lower for European (1.3%), Asian (1.5%), and MELAA (1.9%) ethnic groupings. These different proportions relate to the differential likelihood of group non-response and likely relate to age and country of birth as well.

Birth registration forms ask for ethnicity for the baby (next-of-kin report), and the parents (self-report). Data quality will be impacted by changes in the question and classification over time, as well as differences in collection approach and response patterns. The ethnicity question on birth registrations has only aligned with the census since the passing of the *Births, Deaths, and Marriages Registration Act 1995,* when the question was aligned with the ethnicity question from the 1996 census. For those aged over 14 years, a previous version of the statistical standard will have been used – but these represent 12 percent of the births records used[23].

Ministry of Education tertiary education data

The second prioritised dataset for filling gaps in ethnicity data was MoE tertiary enrolment data. Stats NZ note that the question used to collect this MoE data is inconsistent with the Standard, confirming previous research showing that many of the collections from the education sector do not align with the Standard (Cormack, 2010). Although the Standard states that people should not be forced to identify with a broader general category, and be able to write-in an ethnicity, MoE tertiary ethnicity collections do not always meet these conditions. The use of MoE data may be contributing to the increase of people with a broad, rather than a specific ethnicity, due to both the use of level 3 of the classification and collection forms that contain broad categories without the ability to write-in an ethnicity.

In addition, it is unclear whether individuals who are re-enrolling in a tertiary qualification are required to update or confirm their ethnicity data if it has already been collected previously. In these cases, the data could potentially be many years old, even if an enrolment is recent (Stat NZ used the most recent enrolment information).

Ministry of Health data

There are known quality issues with Ministry of Health ethnicity data that have been documented over a number of years (Cormack & McLeod 2010). These include undercounting of Māori in health sector datasets, disagreement between different datasets, lower than expected reporting of multiple ethnicities, and inconsistent application of a standard approach to ethnicity collection, recording and output (Cormack & McLeod 2010).

[23] While the census ethnicity question changed in 2001, the birth registration form retained the 1996 census ethnicity question until mid-2005, when it moved to align with the 2001 census ethnicity question. This question is essentially the same as that used in the 2006, 2013 and 2018 censuses. Data sourced from birth registrations between 1995 and mid-2005 will, therefore, have been collected using a question that is different to the 2018 or 2013 censuses. In addition, the official classification for ethnicity changed in 2005, with the introduction of the new ethnicity standard (ETHNIC05). Therefore, birth registration data from before DIA aligned with the new statistical standard will have been classified using a different classification to both Census 2018 and Census 2013. This will impact on some specific groups where there were changes between the standards over time.

In a comparison of 2013 census Level 1 total response output with administrative data sources, Stats NZ found that the ratio of Māori in the census compared to Māori in Ministry of Health data was just 0.79. Ratios were higher for European (0.90), Pacific peoples (0.91) and Asian (0.89). Conversely, MELAA (1.07) and Other (3.67) had higher ratios in the Ministry of Health data than in the census (Reid, Bycroft & Gleisner 2015). In short, there are quality issues with MoH ethnicity data, with particular impacts for Māori in terms of undercounting, and for the MELAA and 'Other' categories in terms of overcounting.

Reporting of multiple ethnicities in administrative sources

In Census 2018, 13.0 percent of the subject population reported two or more ethnicities. Of those responding to the ethnicity question in Census 2013, 11.2 percent reported two or more ethnicities. We do not have available the specific breakdowns of ethnic combinations. However, we do know that Māori and Pacific peoples are much more likely than others to report multiple ethnicities in the census. As such, these groups would be disproportionately affected by the use of data sources that do not accurately reflect multiple affiliations. In Census 2013, 53.5 percent of Māori identified with at least one other ethnic group[24]. For the other major ethnic groups, the share reporting two or more major ethnic groups was 32.3 percent for Pacific peoples, 12.3 percent for European, 8.6 percent for Asian, and 16.0 percent for MELAA.

The likelihood of reporting and/or recording multiple ethnicities varies across administrative sources, but in ways that are difficult to account for. Health data often has much lower recording of multiple ethnicities than would be expected based on census patterns. The under-reporting of multiple ethnicities in health might be related to: the context of data collection (i.e. administrative situations compared with survey settings); systems issues including the ability (or not) of systems to record more than one ethnicity; and, prioritisation[25] of ethnicity data at source (this has been reported to occur in both health and education settings).

A Stats NZ study found that there was less than 50 percent agreement of reporting of multiple ethnicity for Māori in MoH data compared with Census 2013 responses, that is, less than 50 percent of Māori who identified as Māori and at least one other ethnic group in Census 2013 were recorded as Māori and another ethnic group in the administrative health data (Bycroft et al., 2015). This pattern was consistent for other Level 1 ethnic groupings that reported multiple ethnicities in 2013, with under 50 percent recorded the same at Level 1 in either the MoE tertiary collection or MoH (Reid et al., 2015).

5.6 Quality rating assessments

Stats NZ assessed the quality of the final Census 2018 dataset using its quality assurance framework. A quality rating scale was developed for three metrics relating to data sources

[24] The single/combined ethnic composition was: 46.5% Māori only; 43.5% Māori and European; 3.9% Māori, Pacific and European; 3.8% Māori and Pacific, and 2.3% residual combinations. Table available on NZ.Stat. Ethnic group (detailed single and combination) by age group and sex, for the census usually resident population count, 2013 Census (RC, TA).

[25] Prioritised ethnic groups involve each person being allocated to a single ethnic group based on a pre-defined hierarchy which is typically: Māori, Pacific, Asian, MELAA, Other and European.

and coverage (metric 1), consistency and coherence (metric 2) and data quality (metric 3), and an overall quality rating given. These different aspects of ethnicity data quality are considered below.

The Stats NZ quality descriptions for the three quality metrics can be found in Appendix 1.

5.7 Data sources and coverage

Stats NZ provided a "high quality" rating for the data sources used to complete the overall ethnicity variable in the final Census 2018 dataset. This process involved Stats NZ matching individuals' ethnicity responses in their 2018 Census forms with their ethnicity responses in their 2013 Census or administrative data, as well as comparisons with statistical imputations. For each source, the resulting quality weighting (0.00 to 1.00) was multiplied by its proportional contribution to the total ethnicity output and summed to derive an overall score. The resultant quality rating will reflect points of difference between Census 2018 data and alternative sources including: possible linking errors; coding and scanning issues in the administrative source collection; differences in classifications and classification levels; and, inconsistencies in individuals' ethnic identification.

While this approach is informative for gauging the extent to which alternative data accurately reflects individuals' ethnic self-designations, there are several important limitations:

- The rating is only based on matched records and the matched proportion varies significantly across ethnic groups. As the final column in Table 5.3 shows, consistency could only be computed for 70.9 percent of Māori, and between 63.4 percent and 70.5 percent of level 2 Pacific ethnic groups;
- Patterns of ethnic mobility for the unmatched population may vary from the matched population in ways that are difficult to account for – we know, for example, that those missed by the census were more likely to be young, male, Māori or Pacific, and concentrated in areas of higher deprivation and Māori population share. It is unclear how these biases will affect ethnic identification;
- The matches are not exact – the weightings are based on a 'total count' rather than on the combination of ethnicity responses given (e.g., 'single/combination' output). Thus, an individual who identified as Māori and European in 2018 but only Māori in 2013 would be counted as a consistent match for Māori in 2018; and,
- The process does not account for an individual's consistency (or lack thereof) in ethnicity reporting across all alternative data sources (e.g., across DIA births, MoE, MoH). An individual's ethnicity might be reported consistently in the 2018 and 2013 Censuses, but differ in other administrative data sources. Taking account of all of the variability would provide a fuller picture of individual-level consistency in ethnic reporting. This matters because the goal of mitigation is to get as close as possible to the 'true' value, in this case, how an individual would self-designate in Census 2018 if given the opportunity to do so. This 'true' value is best assessed by examining individuals' ethnic designations across multiple time points and contexts.

Stats NZ has assessed data source quality for ethnicity overall[26] but we have done so for level 2, which is the lowest level that we are able to assess, and only at the national level. There will likely be significant deviations at the sub-national level and at smaller geographies. In addition, we do not have data by which to assess the quality ratings for multiple ethnicities. Although Stats NZ have two standard output methods ('total response' and 'single/combination'), we only have quality weightings for total response at levels 1 and 2.

Table 5.3 below shows the overall quality rating for the data sources used to constitute the Level 2 ethnic groups. Unsurprisingly, given the high response rate, the quality rating is very high (98-100) for New Zealand European, as well as for Chinese and Indian. The data source quality rating is also high (95-<98) for Māori and for most Level 2 Asian, MELAA and Pacific groups (Niuean, Tokelauan and Cook Island Māori are close to the moderate threshold). We reiterate that quality ratings could only be produced for between 63 percent to 71 percent of individuals in these groups

By contrast, the Fijian and African groups have moderate quality ratings (90-<95). For Fijian the lower rating was driven both by a low percentage of data sourced from Census 2018, and poorer consistency between ethnicity responses on the received forms and the 2013 census and administrative data.

Quality weightings were not provided for lower levels of the classification. However, we can be confident that the overall quality will be lower at levels 3 and 4, and that more groups would have moderate or below moderate ratings.

Table 6: Data source quality rating for level 2 ethnic groups

Level 2 ethnic group	Quality rating	% sourced from 2018 Census forms
European nfd	N/A	20.1
New Zealand European	Very high	89.1
Other European	High	89.5
Māori	High	70.9
Pacific Peoples nfd	N/A	30.7
Samoan	High	68.7
Cook Islands Māori	High	69.0
Tongan	High	63.4
Niuean	High	70.5
Tokelauan	High	69.1
Fijian	Moderate	66.1
Other Pacific Peoples	Moderate	74.2
Asian nfd	N/A	38.2
Southeast Asian	High	84.4
Chinese	Very high	89.1

[26] We note that the historic and administrative metric weightings are based on data at level 2 of the ethnicity classification.

Indian	Very high	78.7
Other Asian	High	86.0
Middle Eastern	High	80.2
Latin American	High	81.6
African	Moderate	76.0
Other	Poor	82.5

Metric 1: Panel rating:

Level 1: National-level only: European: very high; Māori: high; Pacific peoples: high; Asian: very high; MELAA: high; Other: moderate

Level 2: National-level only: As above in Table 5.3

5.8 Consistency and coherence

The metric for ethnicity data consistency and coherence is based on comparability with Dual System Estimation benchmarks, which can only be produced for level 1 Māori, Pacific, and Asian populations[27], expectations for population proportions in national projections, and time series consistency (level 1 to Territorial Authorities and Auckland Council Local Board Areas (TALB); level 4 national). Each of these are discussed below.

5.8.1 Benchmarks

The level 1 counts and comparisons against the DSE benchmarks are shown in Table 5.4.

Table 7: Comparisons of level 1 ethnic counts in Census 2018 with DSE benchmarks

DSE	Benchmark	2018 count	Undercount	% undercount
Māori	807,900	775,836	32,064	4.0%
Pacific peoples	397,200	381,642	15,558	3.9%
Asian	727,400	707,598	19,802	2.7%

Table 5.4 shows that the 2018 Census counts for the level 1 groups with DSE benchmarks available (Māori, Pacific peoples, and Asian) were undercounts relative to the DSE national benchmarks. At the national level the level 1 Asian grouping was closest to the benchmark, with Māori having the largest relative undercount.

It is important to note that, for all recent censuses, the major ethnic group counts were smaller than the Estimated Resident Population (ERP), which includes usual residents temporarily overseas. To illustrate, the Māori 2013 Census count (598,602) was 13.5

[27] The DSE calculations are applied within strata to produce estimates by key demographic breakdowns: single year of age, sex, TALB, and ethnic group. The strata are necessary to produce a range of demographic population estimates, and also to meet underlying assumptions of the DSE method. However there is a limit to the level of detail that can be produced, since fine demographic breakdowns can produce very small cell sizes, which can result in bias to the DSE estimates. The DSE has been produced only for three level 1 ethnic groups: Māori, Pacific, and Asian, because too many small cells would be introduced if any further ethnic groups were attempted. A limiting factor for ethnic groups is the need to apply the DSE to unique combinations of multiple ethnic groups. For example, to produce these three ethnic groups eight mutually exclusive combinations are used.

percent below the 2013 Māori ERP (692,300). Differences for European, Asian, and Pacific peoples ranged between about 8.5 percent and 14.1 percent. The confirmed undercounts subsequently derived from the 2013 Post-Enumeration survey were: 6.1 percent for Māori; 4.8 percent Pacific; 3.0 percent Asian; and 1.9 percent European (Statistics New Zealand, 2014).

The changed methodology for Census 2018 means that the estimated undercounts in Table 5.4, which are considerably lower than in 2013, are likely to be relatively close to the 2018 PES undercount results. The PES results are due to be published early in 2020.

At the sub-national level, the level 1 undercounts, relative to DSE benchmarks, varied substantially. Here we only focus on TALBs with populations of at least 1,000. For Māori in the North Island, census undercounts were highest in South Wairarapa (5.2%), Rotorua (5.0%), and Hamilton City (4.8%). Māori undercounts over 5 percent were also evident in some South Island TAs but the populations were relatively small (e.g., Buller, Kaikoura). Within Auckland, the largest Māori undercounts, relative to DSE benchmarks, were in Waitematā (6.7%), Waiheke (6.3%), Māngere-Ōtāhuhu (6.2%), and Whau (6.2%).

For Pacific peoples the census undercounts were greatest in Hastings (10.44%), Western Bay of Plenty (9.3%), and Taupo (7.6%). Within Auckland LBAs, the level 1 Pacific undercounts, relative to benchmarks, were less marked than for Māori. The largest undercount was in Waitematā (7.7%), followed by Franklin (4.5%).

For the level 1 Asian grouping, the largest proportionate undercounts were in Western Bay of Plenty (7.5%), Hastings (6.4%), and Queenstown-Lakes (5.9%). Within Auckland, the Asian undercount was largest in Waitematā (8.8%), followed by Otara-Papatoetoe (4.1%).

Unfortunately, there are no benchmarks by which to judge the comparability of the MELAA and European sub-national counts, or ethnic groups beyond the high-level major ethnic categories.

5.8.2 Expectations against projections

To further gauge consistency, comparisons were made between the Census 2018 counts and level 1 projections for 2018 from the National Ethnic Population Projections: 2013 (base) - 2038 update[28]. The projections base (2013) data are derived from population estimates which include usual residents temporarily overseas.

In terms of population shares, the 2018 counts closely matched the projected population percentages for Pacific Peoples and Asian groups (Pacific: 8.1% census compared to 8% projected; Asian: 15.1% census compared to 15% projected). The census Māori proportion was slightly above expectation at 16.5 percent compared with the projected 16 percent. The ‘European/Other’ group was projected to comprise 72 percent of the total population in 2018 but the census share was slightly below at 71.3 percent. The 2018 proportion for MELAA of 1.5 percent was somewhat lower than the expectation of 2 percent.

[28] The stochastic projections indicate the future population usually living in New Zealand for eight broad and overlapping ethnic groups: 'European or Other (including New Zealander)', Māori, Asian, Pacific, Chinese, Indian, Samoan, and MELAA.

5.8.3 Time series

Ethnic census counts were also compared against counts from previous censuses to assess comparability. The national-level comparisons are shown in Table 5.5.

Table 8: Level 1 time series comparisons, 2006, 2013 & 2018 census

Level 1 ethnicity	2016 census	2013 census	2018 census		
Count					
European	2,609,589	2,969,391	3,297,864		
Māori	565,329	598,602	775,836		
Pacific Peoples	265,974	295,941	381,642		
Asian	354,552	471,708	707,598		
MELAA	34,746	46,953	70,332		
Other Ethnicity	430,881	67,752	58,053		
Total Stated	3,860,163	4,011,399	4,699,755		
NEI	167,784	230,649	0		
Total Population	4,027,947	4,242,048	4,699,755		
Population percentage (%)				**Increase 2006-13 (%)**	**Increase 2013-18 (%)**
European	67.6	74.0	70.2	13.8	11.0
Māori	14.6	14.9	16.5	5.9	29.6
Pacific Peoples	6.9	7.4	8.1	11.3	29.0
Asian	9.2	11.8	15.0	33.0	50.01
MELAA	0.9	1.2	1.5	35.1	49.8
Other Ethnicity	11.2	1.7	1.2	-84.3	-14.3
Total Stated				3.9	17.2
NEI	4.2	5.4	0.0	37.5	-100.0
Total Population				5.3	10.8

For level 1, the counts are higher for all groups, except for the Other category. The Māori ethnic group count is clearly much higher than would be expected on the basis of recent census counts and intercensal growth. The increase of nearly 30 percent between 2013 and 2018 far exceeds the increase of just under 6 percent between 2006 and 2013. Clearly, this growth is due more to changes in methodology rather than 'real world' demography. While the intercensal increases for Pacific peoples, Asian, and MELAA are also much higher than for 2006–13, the magnitude of the difference is not as large as it is for Māori.

Given the lack of imputation for ethnic non-response in the 2006 and 2013 censuses, and the extensive use of alternative government data sources in Census 2018, these results are not surprising. However, it does make time-series analysis very problematic, particularly for Māori.

Time series analysis becomes more complex at lower levels of the classification. At level 2, for example, the increase in the Cook Island Māori and Niuean counts between 2013 and 2018 (30.2% and 29.3% respectively), were far greater than the 2006–13 increases (6.6% and 6.3% respectively). Large increases were also recorded for the Chinese and Indian groups between 2013 and 2018 (44.5% and 54.1% respectively) compared to 2006–2013 (16.2% and 48.4%).

There are also issues at the national level with time series data for level 4 of the classification, which Stats NZ indicates is likely to be from changes to methodology and the use of administrative data that records ethnicity at levels 2 (MoH) and 3 (MoE) of the classification. Level 4 timeseries are further complicated by changes to the ethnicity classification, which introduced new categories and merged others.

Stats NZ have also noted some inconsistences at lower levels of geography, which we will assess in the next report.

It is the view of the panel 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.

This break has wider implications for the understanding of health and other important social phenomena. For example, the larger count may artificially deflate rates when used as denominators where the numerator data are sourced elsewhere (e.g., health). In such cases it would be advised to use the ERP ethnic figures, although this is not always possible (e.g, morality by occupational grouping).

Metric 2: Panel rating:

Level 1: National-level only: European: high; Māori: moderate; Pacific peoples: moderate; Asian: moderate; MELAA: moderate.

5.9 Data quality

The data quality 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 rated the ethnicity variable as ‘high’ on this metric; the Panel has rated it as ‘moderate to high’. The Stats NZ data quality assurance framework defines data quality as ‘moderate’ where “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.” The Panel considers that some data quality issues associated with a ’moderate’ rating are apparent for ethnicity.

Overall, the level of nfd ethnicity responses is minimal, comprising just 0.9 percent of total responses at level 2 and 1.0 percent at level 3. However, the levels of nfd vary significantly

across categories, with nfd responses ranging from 0.7 percent of all level 3 Pacific ethnicities to 2.5 percent of level 3 Asian ethnicities.

At level 4 the share of nfd responses is very high for the disaggregated MELAA categories: 30 percent nfd for Middle Eastern; 38 percent for South American, and 43 percent for African. We note that the percentage increase in these categories between 2013–18 was lower than for 2006–13.

Metric 3: Panel rating: Moderate to high

Overall rating:

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.

5.10 Other comments relating to the uses of census ethnicity data

The quality of ethnicity information from Census 2018 ultimately depends on the correct designation of ethnicity to census non-responders. However, its usefulness, particularly to ethnic communities, depends greatly on the ability to cross-tabulate ethnicity with other census variables to identify the characteristics and conditions of groups at different levels of ethnic and spatial disaggregation.

While it has been possible to place almost all individuals in Census 2018 into the ethnicity classification with a high degree of reliability, it is only the European and Asian major ethnic groups for which the majority of characteristics obtained from the census questionnaire exist for more than 80 percent of people. At level 2 it includes: New Zealand European; Other European; Chinese; Southeast Asian; Other Asian; Middle Eastern; Latin American; and Other.

For Māori and Pacific peoples, and other groups with 80 percent or less of their ethnicity data sourced from Census 2018, much more caution will be needed when analysing characteristics based on variables where a high level of imputation has been used, or where the quality rating of the alternative data source used is less than high. For some variables, such as language and cigarette smoking, there are no alternative data sources except for historic 2013 Census data.

These issues are important to acknowledge given Stats NZ's Treaty and equity obligations. If data are difficult to use or are subject to an unreasonable number of caveats, their usefulness and value to ethnic communities and for key uses of census ethnicity data may be compromised. The panel will explore the issue of ethnic characteristic data more fully in its next report.

6. Data quality of key variables

6.1 Summary

Table 6.1 summarises the panel's views on quality of the Priority 1 data that is to be released by Stats NZ on 23 September 2019. More detail is given on key variables below, except for ethnicity, which was reported on in section 5. Quotes from Stats NZ included in this section are taken from internal quality assurance documents which were shared with the panel.

Table 6.1: External Data Quality Panel ratings

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/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. See section 5.
Māori descent – electoral	High	High – See section 4.
Māori descent – output	High	High - See section 4.
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.

Absentees is a Priority 1 variable which has been rated by Stats NZ as of very poor quality and is not being released on 23 September. The quality assurance panel endorses this assessment.

6.2 Usual Resident Count

Stats NZ rate the quality of 'Counts of Individuals – Usual Residents' as very high. The quality assurance panel endorses this assessment.

Stats NZ endorse the usual resident counts at regional council/Territorial Authority/Auckland local board, and Statistical Area 2 (SA2). They have also carried out analyses at Meshblock level that confirms the data quality at this level – which the panel also endorses. There are a small number of meshblocks where NPDs have been allocated to different meshblocks compared to 2013 (this can unfortunately happen if a large NPD spans more than one meshblock) – which will distort changes at this very low level of geography.

Most uses of census data rely on the usual resident count. This count tells us about those people usually resident in New Zealand who consume public services, contribute to the economy, live in families etc. It is a critical variable, and its quality impinges on all other census outputs. Table 6.2 summarises the source of information for the usually resident count.

Table 6.2: Sources of information for usual resident count

Unit record source	Number	%
Individual form	3,971,892	84.5
Individual from household listing	202,914	4.3
Field enumerated rough sleeper	99	<0.1
Response	*4,174,902*	*88.8*
Enumeration in admin household – occupied	99,159	2.1
Enumeration in admin household – unoccupied	42,252	0.9
Enumeration in admin household	141,411	*3.0*
Admin enumeration at responding private dwelling	20,643	0.4
Admin enumeration at prison or penal institution	4,707	0.1
Admin enumeration at defence establishment	798	<0.1
Other dwelling-based admin enumeration[1]	26,148	*0.6*
Meshblock enumeration	357,294	7.6
Total	*4,699,755*	*100.0*

In summary, 84.5 percent of census usual residents came from individual census forms, with a further 4.3 percent from household listing information on dwelling forms (but where there was no individual form). A further 3 percent of records were admin enumerations (of households) into non-responding census dwellings (either classified as occupied or unoccupied on census night). A total of 20,643 records (0.44%) were for individual admin records added to responding census private dwellings (i.e., adding people where there was evidence of gaps in the census returns).

Admin records (0.12% of the total) were also used to fill gaps in responses from prisons and defence establishments. On top of this, a further 357,294 records (7.6%) were admin records that could not reliably be associated with an address and were therefore assigned

only to a meshblock. Overall 11.17 percent of 2018 Census usual residence records were sourced from admin data.

The effects of the response issues with the 2018 Census were not evenly spread across New Zealand, so there is regional variation in the reliance on non-2018 Census data. Whilst nationally 15.5 percent of 2018 Census data came from sources other than the 2018 Census, the equivalent figure for Gisborne was 22 percent; for Northland it was 20 percent; and for each of Auckland, Hawke's Bay, and Bay of Plenty it was 18 percent.

Comparison with expectations

The 2018 total count of 4,699,755 is 68,800 people (1.4%) lower than the estimated resident population count of 4,768,600, which is felt to be within an acceptable range.

Stats NZ have carried out analyses of usual resident counts at meshblock level – this involved looking at the 51,000 meshblocks and focusing on, for example, those with the biggest changes since 2013. Of these short-listed meshblocks, approximately 500 were investigated. This analysis concluded that:

- the majority of meshblock level changes appear valid
- of the problems identified, they appear to be distributed across New Zealand; problems do not appear to be concentrated in any one area
- there are a small number of NPDs that have been put into a neighbouring meshblock compared to 2013, resulting in large rates of change between 2013 and 2018. Whilst creating odd changes at the meshblock level, these will not be visible at higher levels of geography.

The analysis did not find significant issues with the meshblock level dataset. The type of issues identified were consistent with known issues in previous censuses.

Data for the usually resident population follows similar trends to those recorded in 2006 and 2013. Overall, observed changes in the 2018 data from previous censuses are in line with expectations. There were particular response problems for Prisons and Defence Establishments, and admin records were used for 4,707 prisoners and 798 staff in Defence Establishments.

6.3 Census Night Count

Stats NZ rate the quality of 'Counts of individuals – census night' as moderate. This is mostly due to comparability with previous census estimates, particularly for overseas visitors. The quality assurance panel endorses this assessment.

The census night population count is a count of all people enumerated by census, who were present in New Zealand on census night. The census night population counts people where they are in New Zealand on census night and includes overseas visitors as well as New Zealand usual residents.

Most uses of census data rely on the usual resident count. The census night count is less significant as an output in its own right but is important for some local areas, such as those that receive large numbers of tourists and visitors.

Table 6.3: Sources of information for census night count

Unit record source	Number	%
Individual form	4,065,300	84.8
Individual from household listing	203,106	4.2
Field enumerated rough sleeper	99	<0.1
Response	4,268,505	89.1
Enumeration in admin household – occupied	99,159	2.1
Enumeration in admin household – unoccupied	42,252	0.9
Enumeration in admin household	141,411	3.0
Admin enumeration at responding private dwelling	20,643	0.4
Admin enumeration at prison or penal institution	4,707	0.1
Admin enumeration at defence establishment	798	0.0 <0.1
Other dwelling-based admin enumeration1	26,148	0.5
Meshblock enumeration	357,294	7.5
Total	4,793,358	100.0

The census night population of New Zealand in 2018 was 4,793,358, up 15.7 percent from 2006 and 10.1 percent from 2013. The Census night population was made up of 4,699,755 New Zealand usual residents and 93,606 overseas visitors. While most New Zealand residents are at home on census night, 162,441 New Zealand residents were away from home on census night.

In contrast to the growth seen for New Zealand usual residents which increased by 10 percent from the 2013 Census, the population of overseas visitors decreased 15.8 percent in 2018.

As shown above in Table 6.3, 89 percent of records for New Zealand usual residents are sourced from census forms, and 11 percent from admin records and imputations. All overseas visitors are sourced from census forms.

Census night data for New Zealand residents follows similar trends to those recorded in 2006 and 2013. Overall, observed changes in the 2018 data from previous censuses is in line with expectations. Stats NZ state “For New Zealand residents, in general, comparisons at regional council, territorial authority, Auckland local board, and Statistical Area 2 (SA2) level are *reasonable*.”

When compared with expectations and previous years, the increase of the census night population was evenly distributed across age and sex but not geographically. The geographical inconsistencies when compared with expectations and previous years become more pronounced at lower geographical levels.

2018 Census night counts sourced from admin records may not have been counted at their census night address – it is likely that admin sources capture people at their usual residence – for those away from home on census night this will not be a good proxy.

Stats NZ state "We do not know where admin enumerated individuals would have been enumerated on census night if they had completed forms." A census night location was imputed elsewhere in New Zealand for admin enumerations in dwellings classed as 'residents away' on census night. For those admin enumerations placed in meshblocks, the census night location is imputed from a 'nearest neighbour', so the distribution of 'at home' and 'away from home' on census night follows a similar pattern to received census responses.

The census typically undercounts overseas visitors in New Zealand on census night, but census night coverage for overseas visitors is lower in 2018 that in 2006 and 2013. In 2018 there was a substantial undercount of overseas visitors, which affected the census night count for some geographic areas more than others. Based on external migration estimates of the number of overseas visitors in New Zealand on census night, census coverage of overseas visitors was around 60 percent in the 2006 and 2013 Census and dropped to 35 percent in 2018 Census.

Stats NZ state "Census night coverage for overseas visitors differs from the 2006 and 2013 data. There was a substantial undercount of overseas visitors which affected some geographic areas more than others. Administrative enumeration is not able to count overseas visitors, and enumeration of non-private dwellings (NPDs) was poor."

Since the main use of the census night count is at local geographic level, particularly those areas that rely on tourism, the 2018 census night count is rated as moderate quality due to the undercount of overseas visitors.

6.4 Dwelling occupancy status

Stats NZ do not give a quality rating for dwelling occupancy status.

Dwelling occupancy status classifies all dwellings according to whether they are occupied, unoccupied, or under construction etc. during the time period of the census data collection. The occupancy status is applied to dwellings on the Census Dwelling Frame.

Dwelling occupancy status is defined during the field operation and includes judgements by census staff (e.g., field staff or contact centre staff) as well as being based on information from respondents. The quality of the occupancy status will therefore be variable, partly depending on local judgements – but these local judgements are likely to be no better or worse than for previous censuses.

Table 6.4: Dwelling occupancy status

Dwelling occupancy status	Count
Occupied dwellings	1,673,880
Residents away	98,664
Empty dwelling	97,842
Dwelling under construction	16,128
Total dwellings	1,886,517

6.4.1 Background to the Census Dwelling Frame

The Census Dwelling Frame was derived from the Statistical Location Register (SLR) which was itself produced to support the 2018 Census - an address list to locate and enable respondents. The SLR was formed by geocoding the 2013 Census dwellings and is updated monthly by Land Information NZ (LINZ) and NZ Post data.

A 2018 Census Operational File was derived from the SLR, which was verified by on-the-ground canvassing by field officers. Private dwelling addresses were divided between those to be mailed out (with an internet access code – 80 percent of dwellings) and those to be visited in the field during the list/leave phase' of census operations (20% of dwellings). The mailout meshblocks were canvassed between June and August 2017 to check and update the address information; the field 'list-leave' and 'targeted' areas were canvassed as part of the census field operation.

The Census Operational File was updated during the census operations. The changes made indicate that initial dwelling classifications were of high quality, especially for private dwellings (99.3% of all dwellings). Further corrections and improvements were made between May and September 2018, when the dwelling frame was ready for processing and evaluation. The function of the dwelling frame was now to support statistical processing and produce a clean unit record file (CURF) to enable the production of outputs from the 2018 Census data.

6.5 Usual residence address

Stats NZ rate the quality of 'Usual residence address' as high quality. The panel endorse this assessment at the national and regional council/TALB level. We note however that there must be greater uncertainty at the SA2 and meshblock level as Stats NZ have not been able to fully assess that quality.

Stats NZ state that "Usual residence addresses are of a high quality via most of the near 4 million individual forms (IFs) received having been completed online (87%) where they are of a high data quality, compared with on paper forms (13%) where they are of a lesser data quality."

For around 216,000 partial responses there is good quality usual residence address information from the household summary form or dwelling form.

There were around 523,000 admin records added to the census file to count people who were missed. Almost all (99.6%) of the usual residence addresses for these records were

sourced from admin data, with most of the remaining 0.4 percent from CANCEIS nearest neighbour.

Stats NZ state that "The usual residence address data quality assessment would have been assessed as very high if not for the 357,000 people that could only be assigned to a meshblock."

Note that the records assigned to a meshblock impact on the quality of the usual residence address but do not impact in the same way on the quality of usual residence count (hence their different quality ratings).

6.6 Census night address

Stats NZ rate the quality of 'Census night address' as moderate quality. The panel endorse this assessment at the national and regional council level. There is greater uncertainty at lower levels of geography.

Stats NZ state "There was a substantial undercount of overseas visitors which affected some geographic areas more than others. … When compared with expectations and previous years, the increase of the census night population was evenly distributed across age and sex but not geographically. The geographical inconsistencies when compared with expectations and previous years become more pronounced at lower geographical levels (e.g., territorial authorities and local boards)."

Census night address data for New Zealand residents follows similar trends with respect to age to those recorded in 2006 and 2013. However, census night address data for overseas visitors does not.

6.7 Age

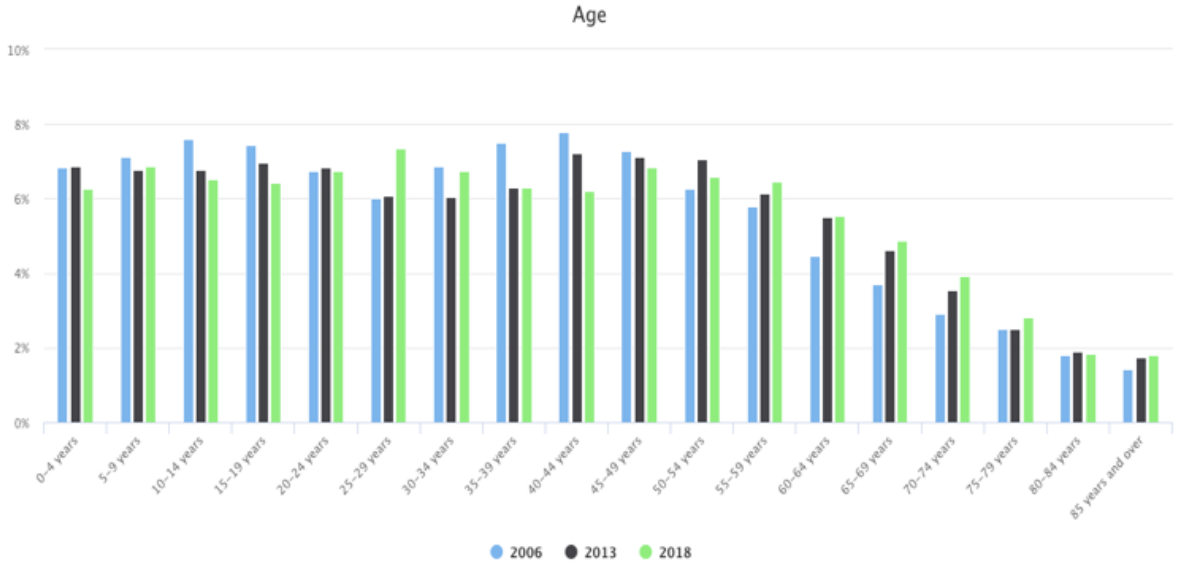
Stats NZ rate the quality of Age as very high quality. The panel endorse this assessment at the national and regional council levels of geography. The panel has not seen analyses below regional council level.

Stats NZ state "Overall, any change in the data in 2018 from previous censuses can largely be attributed to real world change, with minimal issues associated with data quality, non-response, or changes to the census collection methodology."

Age was collected on the individual form (IF) from question 2 'When were you born' which has a date of birth (day, month and year of birth) answer. Age was also collected on the dwelling form) and is taken from question 17 which lists each person at the dwelling on census night and asks their age. Age on the dwelling form is an age and is not a date of birth.

Figure 6.1 shows the percentage distribution of the New Zealand population in five-year age bands for 2018, 2013 and 2006. The increasing proportion of older people is in line with an ageing population. What is most striking is the change in pattern for those aged 25-29; 30-34; and possibly 35-39, with an increasing share of the population compared to falls over the previous two censuses.

Figure 6.1: Counts by five-year age bounds, 2006 to 2018



Stats NZ state "A lot more people aged in their twenties and thirties, particularly males". (This could be due to the high level of migration over the last few years). Previous censuses have had the missing male problem where males in their 20s and 30s don't get enumerated in the census giving a lower number of males for these ages. It is possible that the data from admin sources has picked up some of these missing males meaning the data quality could be better than in previous years."

This means there could be a break in the time series from previous censuses when looking at males (and possibly females) in their 20s and 30s.

Table 6.5 summarises the source of the information for the 2018 Census age counts.

Table 6.5: Source of information for age counts

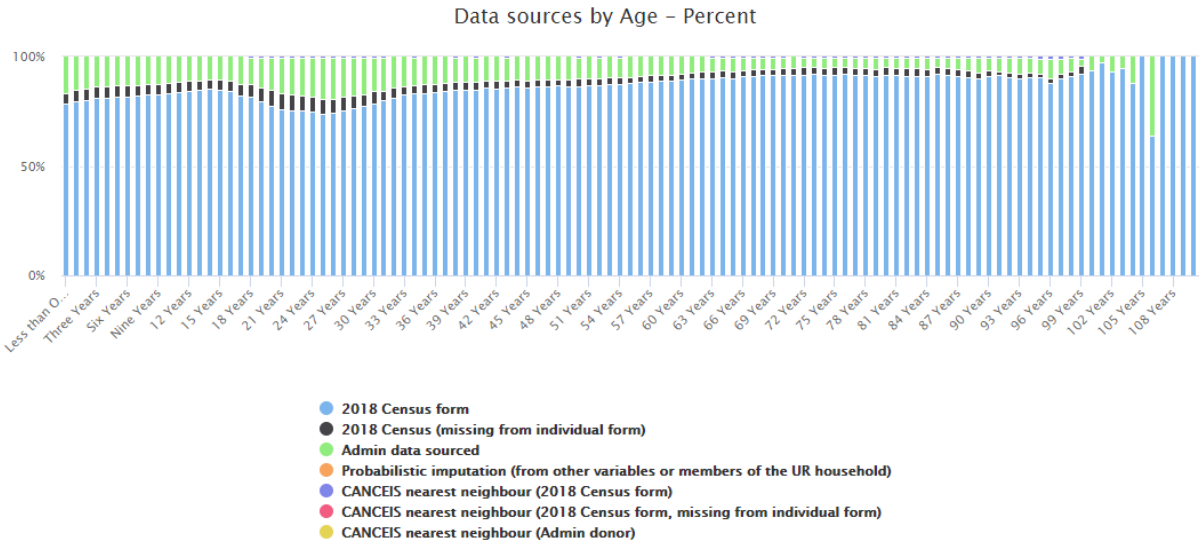
Data source	Number	Percent
2018 Census form	3,966,819	84.4
2018 Census (missing from individual form)	195,522	4.2
Admin data sourced	524,853	11.2
Probabilistic imputation (from other variables or members of the UR household)	15	<0.0
CANCEIS (imputation) nearest neighbour (2018 Census form)	9,075	0.2
CANCEIS (imputation) nearest neighbour (2018 Census form, missing from individual form)	3,129	0.1
CANCEIS (imputation) nearest neighbour (admin donor)	312	0.0
Total	4,699,755	

Admin data for age has been rated by Stats NZ as of the same quality as 2018 Census data, and of higher quality than 2013 Census data. Stats NZ state that the majority of age data from admin sources will come from birth registrations, passport, and tax information, and give a date range up to July 2018.

Figures 6.2 and 6.3 show data sources used, by age, indicating that higher proportions of admin data records were used for:

- young children (parents often leave them out of the census)
- young adults (who are harder to enumerate), and
- the very elderly (99+) where there can be recording errors in date of birth (wrong century recorded).

Figure 6.2: Data sources by age (%)



Stats NZ have analysed the data sources by regional council area. The use of admin data ranged from 16.7 percent in Gisborne, 15.8 percent in Northland, 12.5 percent in Auckland, and 8 percent in the Canterbury region.

Figures 6.3 and 6.4 show the single year of age for males and females from the 2006, 2013, and 2018 censuses.

Figure 3.3: Age by sex - male

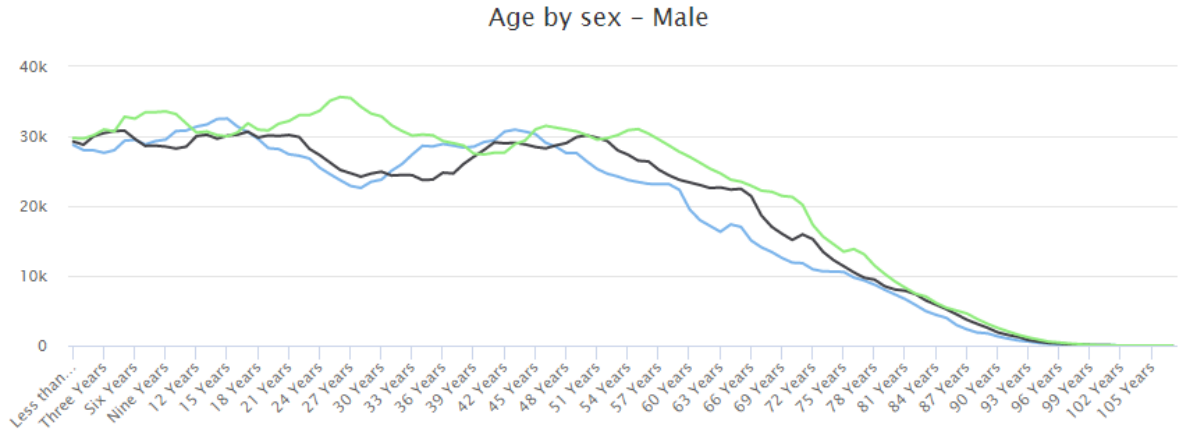
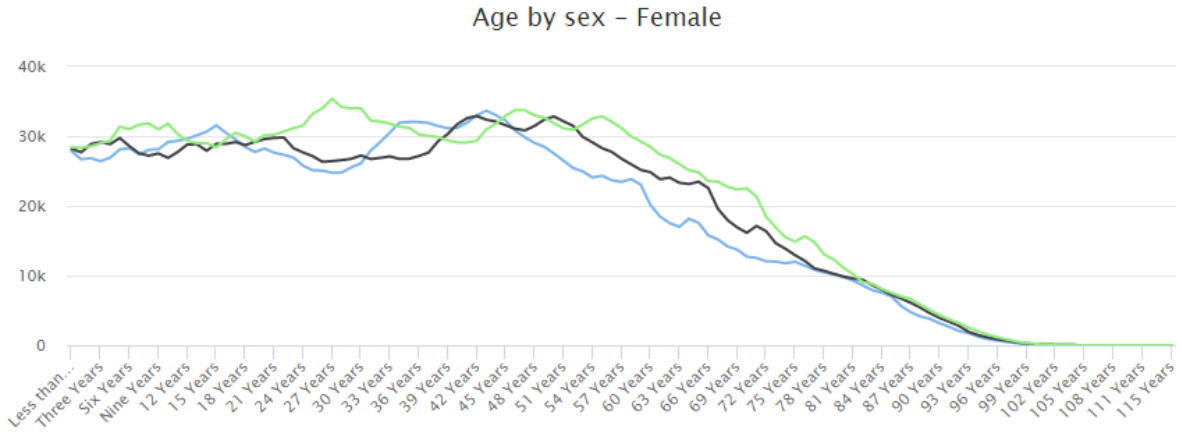


Figure 6.4: Age by sex - female



The graphs show the expected right shifting of patterns by the number of years between the censuses, except for young adults in the 2018 Census. Young adults are typically hard to count in the census and the 2018 Census use of admin data is probably a quality improvement compared to previous censuses, but may have introduced a break in the time-series for at least this age group.

6.8 Sex

The final rating by Stats NZ is 'very high quality' down to the SA2 level of geography. The panel endorses this assessment.

Stats NZ state "While there are new imputation methods and the use of administrative data for this variable, the overall quality of the data is sound and comparable with 2006 and 2013 data."

Sex is a key and fundamental output from the census. Almost all other variables are cross tabulated by sex (and often age and sex). The question on sex in the 2018 Census did not change from that in 2013. The classification has not changed (male/female). The sex question was mandatory in the online form (you could not submit the form without completing this question). Only 4,830 paper forms did not contain sex information.

Table 6.6 below summarises information sources for the 2018 data on sex.

Table 6.6: Source of information for sex

	Number	%
2018 Census form	3,962,430	84.3
2018 Census (missing from individual form)	206,226	4.4
Admin data sourced	524,853	11.2
Probabilistic imputation (from other variables or members of the UR household)	1,317	0.0
CANCEIS (imputation) nearest neighbour (2018 Census form)	4,146	0.1
CANCEIS (imputation) nearest neighbour (2018 Census form, missing from individual form)	516	0.0
CANCEIS (imputation) nearest neighbour (Admin donor)	228	0.0
CANCEIS (imputation) nearest neighbour (Within household)	36	0.0
Total	4,699,755	100

In summary, 84.3 percent of census sex data came from individual census forms. 1,916,217 for men; 2,046,213 for women), with a further 4.4 percent from dwelling forms. 11.2 percent came from admin data records (524,853 records - 291,948 for men; 232,905 for women). Only 0.1 percent of records were imputed.

Of the fully responding individuals who answered the sex question, 87.2 percent answered online and 12.8 percent answered on paper forms. The percent of male and female responses to the sex question online was similar at 87.4 percent and 87.1 percent respectively (12.6% of male responses to the sex question were on paper forms; the figure was 12.9% for female responses)

Sex is a basic characteristic collected in admin data systems. Admin data on sex is therefore likely to be highly reliable. Whilst it is possible for individuals' self-identification of sex or gender to change over time, this is not common. The sex data from admin sources will therefore be valid as at the census date, even if it was recorded in an admin data system some time previously.

Stats NZ have previously documented that people who identify as intersex had the option to complete a paper-based form and tick both male and female. However, they would be imputed to either male or female in the output. However, Stats NZ note that there were a lot of false positives with people marking female with a tick (on the paper form) that went into the male field and was picked up by scanning in both tick boxes. These had to be resolved.

7. Tests of quality

Following the first release of 2018 Census data in September 2019, Stats NZ has a sequence of further data releases planned.

From the work of the External Data Quality Assurance Panel additional analyses on all of the main variables produced as part of the census dataset will be available.

Few of the statistics produced from the 2018 Census have been prepared in the same manner as the information from earlier censuses and understanding the effect on well used measures has become most important. Along with the statistical results from the 2018 census there will be a wide range of quality information, with which users will need to become familiar. This includes response rates, scale and sources of imputed information, and expert assessments of quality made by both Stats NZ and the External Data Quality Assurance Panel.

Basic counts of individuals by age, sex, ethnicity, and place have fewer problems from response at enumeration time than experienced in earlier censuses because of the quality of address and person frame that is now available to compare against the enumerated population and find substitute responses, on a scale that has not been practical before 2018. Information in imputed populations is readily available.

While there are always changes in content and differences in the response rates from people depending on age, place, ethnicity, those differences have not played a key part in the supporting advice given by Stats NZ in the use of statistics from earlier censuses. There has been a presumption that the error structure inherent in a traditional census would not change dramatically from one census to the next. This assumption will not hold for the 2018 Census, and it is likely that the changes between 2013 and 2018 reported from these censuses will reflect to some degree changes in method. The same concerns are likely to arise later when comparing 2023 Census statistics with those from 2018. Analyses that need to be interpreted with caution will result when:

- *The census variable that has been analysed is only available from information provided on the questionnaire.*
- *The population in the ethnic, age or place group being analysed was not defined for a significant minority of respondents by information found on completed census forms.* This also affects comparative studies of occupations.
- *The information that the census statistics are based on requires individuals to be placed in a specific dwelling, rather than located at the level of meshblock*
- *Analyses require information from two or more topics with a low response rate such as ethnicity and languages spoken to be combined.*
- *Census variables have classifications with multiple levels of aggregation.* It is important to examine the non-response rates and imputation levels of the components of interest at the most detailed level of the classification, as huge

variances that affect the use may be obscured if the cancel out at any of the aggregated levels of the classification (e.g., place, ethnicity).

- *The values of the same variable for different people have not been obtained for the same time period*. In the income question, the reference period may be lagged by a year as the information is usually measured for a tax year, rather than cumulatively up to some selected point. The ethnicity measure has been obtained from several sources which refer to different points in time before the census reference date.

- *Potential non-response bias.* Those who do not complete questionnaires are not representative of the total population. For example, where travel to work measures are only available for those whose address and place of work or school has been sufficiently well identified during the same time period.

In the 2018 Census, much of the information available will need to be associated with assessments and measures of quality, as the quality that has usually been provided in not possible. We would advise that for any analyses of the 2018 Census of Population and Dwellings that is intended to influence public and community welfare in any way that user should apply the quality information available from Stats NZ to validate their fitness for use, and assess the degree of confidence in the analyses and its impact on decision-making.

References

Bakker, C. (2014). *Valuing the census*. A report prepared for Stats NZ which quantifies the benefits to New Zealand from the use of census and population information. Retrieved from https://www.stats.govt.nz/assets/Research/Valuing-the-Census/valuing-the-census.pdf

Cannataci J. 2018. Report of the Special Rapporteur on the right to privacy. 17 October. Retrieved from A/73/45712. https://www.ohchr.org

Caron-Malenfant, E., Coulombe, S., Guimond, E., Grondin, C. and Lebel, A. (2014). Ethnic mobility of Aboriginal peoples in Canada between the 2001 and 2006 censuses. *Population, 69*(1), 29-53

Cormack, D., and Harris, R. (2009). *Issues in monitoring Māori health and ethnic disparities: An update*. Wellington, New Zealand: Te Rōpū Rangahau Hauora a Eru Pōmare.

Didham, R. (2016). Ethnic mobility in the New Zealand Census, 1981–2013: A preliminary look. *New Zealand Population Review, 42,* 27–42.

Dot Loves Data (2019). Sensitivity analysis of 2018 Census for electoral boundaries. Unpublished report provided to Statistics NZ.

Electoral Act 1993. Retrieved from http://www.legislation.govt.nz/act/public/1993/0087/latest/DLM307519.html

Gleisner, F, Downey, A, & McNally, J (2015). *Enduring census information requirements for and about Māori.* Retrieved from http://archive.stats.govt.nz/methods/research-papers/enduring-census-requirements-maori/

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

Jenkins, K. (2018). Can I see your social licence please? *Policy Quarterly, 14*(4), 27-35.

Jonas, M. (2018). Ethics and the integrated data infrastructure (IDI). *Ethics Notes,* August. Retrieved from https://mailchi.mp/hrc/ethics-notes

Keller, A., Mule, V.T., Steeg Morris D., & Konicki S. (2018) A Distance Metric for Modeling the Quality of Administrative Records for Use in the 2020 U.S. Census. Journal of Official Statistics, Vol. 34, No. 3, 2018, pp. 599–624, http://dx.doi.org/10.2478/JOS-2018-0029

Kukutai, T., Thompson, V. & McMillan, R. (2015). Whither the census? Continuity and change in census methodologies worldwide, 1985–2014. *Journal of Population Research.* DOI 10.1007/s12546-014-9139-z

Kukutai T, and Taylor J, (Editors). (2016). *Indigenous data sovereignty: toward an agenda*. Canberra: ANU Press.

Ministry of Health (2017). HISO 10001:2017 Ethnicity Data Protocols. ISBN 978-1-98-850291-5 (online)

Simply Privacy Ltd. (2017). *2018 Census. Independent privacy impact assessment*. Retrieved from https://www.stats.govt.nz/privacy-impact-assessments/2018-census-independent-privacy-impact-assessment

Statistics New Zealand (2014). *Post-enumeration survey: 2013*. Retrieved from http://archive.stats.govt.nz/browse_for_stats/population/census_counts/PostEnumerationSurvey_HOTP13.aspx.

Statistics New Zealand (2015). Measuring disability in New Zealand: Current status and issues (PDF).

Stats NZ (2017a). *2005 New Zealand standard classification of ethnicity.* Retrieved from http://archive.stats.govt.nz/methods/research-papers/topss/comp-ethnic-admin-data-census/classification-of-ethnicity.aspx

Stats NZ (2017b). *Post-enumeration survey: 2013*. Retrieved from http://archive.stats.govt.nz/browse_for_stats/population/census_counts/PostEnumerationSurvey_HOTP13.aspx

Stats NZ (2018a). *2018 Census: Changes and how they might affect the data*. Retrieved from https://www.stats.govt.nz/methods/2018-census-changes- and-how-they-might-affect-data

Stats NZ (2018b). *Vote Statistics Four Year Plan 2017/18 – 2020/21*

Stats NZ (2019a). *Addition of administrative records to the New Zealand 2018 Census dataset: An overview of statistical methods*. Retrieved from www.stats.govt.nz

Stats NZ (2019b). *Data quality assurance for 2018 census*. Wellington: Stats NZ.

Stats NZ (2019c). Dual system estimation combining census responses and an admin population. Available from www.stats.govt.nz.

Stats NZ (2019d). *Linked administrative sources for the census and population statistics in New Zealand*. Retrieved from www.stats.govt.nz

Treaty of Waitangi Fisheries Commission. (2003). *He kawai amokura –this report represents the 'full particulars' of a model for allocation of the Fisheries Settlement assets.* Wellington: Treaty of Waitangi Fisheries Commission.

United Nations (2017). *Principles and Recommendations for Population and Housing Censuses, Revision 3.* Department of Economic and Social Affairs. Retrieved from https://unstats.un.org/unsd/demographic-social/Standards-and-Methods/files/Principles_and_Recommendations/Population-and-Housing-Censuses/Series_M67rev3-E.pdf

UN Special Rapporteur on the Right to Privacy. 2019. *Draft recommendation on the protection and use of health-related data*. Retrieved from https://www.ohchr.org/EN/Issues/Privacy/SR/Pages/HealthRelatedData.aspx

UN Economic and Social Council. (2015). *Statistical Commission. Report on the 46th Session.* New York: United Nations. Retrieved from https://unstats.un.org/unsd/statcom/46th-session/documents/.

Appendix 1 – Stats NZ data quality assurance definitions for 2018 Census

Stats NZ's 'Data quality assurance for 2018 Census' (Stats NZ, 2019b) outlines the quality rating scale and quality assurance framework 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 2 – Admin data sources by ethnicity level 2

Ethnic Group	COR	DEF	DIA	MOE	MOH	Total
European nfd - count	**S**	**6**	**186**	**10,428**	**14,187**	**24,807**
European nfd - percent	*S*	*0.0%*	*0.7%*	*42.0%*	*57.2%*	*100.0%*
New Zealand European - count	**21**	**207**	**39,657**	**29,634**	**25,554**	**95,073**
New Zealand European - percent	*0.0%*	*0.2%*	*41.7%*	*31.2%*	*26.9%*	*100.0%*
Other European - count	**S**	**90**	**4,605**	**2,931**	**S**	**7,626**
Other European - percent	*S*	*1.2%*	*60.4%*	*38.4%*	*S*	*100.0%*
Maori - count	**48**	**129**	**50,568**	**37,221**	**13,371**	**101,337**
Maori - percent	*0.0%*	*0.1%*	*49.9%*	*36.7%*	*13.2%*	*100.0%*
Pacific Peoples nfd - count	**21**	**S**	**S**	**564**	**1,080**	**1,665**
Pacific Peoples nfd - percent	*1.3%*	*S*	*S*	*33.9%*	*64.9%*	*100.0%*
Samoan - count	**S**	**12**	**12,378**	**5,679**	**7,101**	**25,170**
Samoan - percent	*S*	*0.0%*	*49.2%*	*22.6%*	*28.2%*	*100.0%*
Cook Islands Maori - count	**S**	**9**	**6,813**	**2,868**	**1,395**	**11,085**
Cook Islands Maori - percent	*S*	*0.1%*	*61.5%*	*25.9%*	*12.6%*	*100.0%*
Tongan - count	**S**	**S**	**7,500**	**2,964**	**3,156**	**13,620**
Tongan - percent	*S*	*S*	*55.1%*	*21.8%*	*23.2%*	*100.0%*
Niuean - count	**S**	**S**	**2,472**	**930**	**501**	**3,903**
Niuean - percent	*S*	*S*	*63.3%*	*23.8%*	*12.8%*	*100.0%*
Tokelauan - count	**S**	**S**	**573**	**336**	**192**	**1,101**
Tokelauan - percent	*S*	*S*	*52.0%*	*30.5%*	*17.4%*	*100.0%*
Fijian - count	**S**	**6**	**942**	**1,119**	**1,809**	**3,876**
Fijian - percent	*S*	*0.2%*	*24.3%*	*28.9%*	*46.7%*	*100.0%*
Other Pacific Peoples - count	**S**	**S**	**1,146**	**S**	**S**	**1,146**
Other Pacific Peoples - percent	*S*	*S*	*100.0%*	*S*	*S*	*100.0%*
Asian nfd - count	**S**	**S**	**72**	**1,590**	**4,746**	**6,408**
Asian nfd - percent	*S*	*S*	*1.1%*	*24.8%*	*74.1%*	*100.0%*
Southeast Asian - count	**S**	**S**	**1,833**	**2,355**	**3,885**	**8,073**
Southeast Asian - percent	*S*	*S*	*22.7%*	*29.2%*	*48.1%*	*100.0%*
Chinese - count	**S**	**S**	**2,670**	**5,274**	**3,639**	**11,583**
Chinese - percent	*S*	*S*	*23.1%*	*45.5%*	*31.4%*	*100.0%*
Indian - count	**S**	**S**	**4,692**	**11,400**	**10,893**	**26,985**
Indian - percent	*S*	*S*	*17.4%*	*42.2%*	*40.4%*	*100.0%*
Other Asian - count	**S**	**9**	**1,410**	**1,536**	**S**	**2,955**
Other Asian - percent	*S*	*0.3%*	*47.7%*	*52.0%*	*S*	*100.0%*
Middle Eastern - count	**S**	**S**	**597**	**738**	**1,254**	**2,589**
Middle Eastern - percent	*S*	*S*	*23.1%*	*28.5%*	*48.4%*	*100.0%*
Latin American - count	**S**	**S**	**306**	**477**	**1,458**	**2,241**
Latin American - percent	*S*	*S*	*13.7%*	*21.3%*	*65.1%*	*100.0%*
African - count	**S**	**S**	**432**	**771**	**717**	**1,920**
African - percent	*S*	*S*	*22.5%*	*40.2%*	*37.3%*	*100.0%*
Other Ethnicity - count	**6**	**306**	**387**	**3,345**	**447**	**4,491**
Other Ethnicity - percent	*0.1%*	*6.8%*	*8.6%*	*74.5%*	*10.0%*	*100.0%*

Appendix 3 - 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.
Alpha (α)	A score reflecting the probability that an administrative meshblock reflects the true meshblock of usual residence for an individual (based on a statistical model). This score is used as a threshold to determine the administrative records to be included in the census usual resident population.
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)
CURF	Clean unit record file. A finalised approved data file made available for reporting and analysis where responses have been validated and the available information will meet the confidentiality protection requirements
DIA	Department of Internal Affairs.
Donor imputation	Method of imputation which uses data from similar individuals or households to 'impute' (fill-in) data for missing responses/respondents
DSE benchmark	The estimated usually resident population count as at census night based on dual system estimation (DSE). Interim coverage and response rates are calculated using this benchmark.
Dual system estimation (DSE)	A method used to estimate the total population using population estimates from two or more sources (e.g., the DSE benchmark combines census results and data from the Integrated Data Infrastructure (IDI)).
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'; 36 categories at 'level 3'; and 180 categories at 'level 4'.
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.
General Electoral Population (GEP)	The census usually resident population count minus the Māori Electoral Population.
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.

IDI Spine	The primary person-level dataset in the IDI to which all other person-level datasets are linked. The current (prototype) spine used in the IDI is formed by linking together tax (IRD) records since 1999, New Zealand birth records from 1920, and long-term visa records from 1997.
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.
Linkage	The process of combining two or more data sets so that a data set with more information can be created which can then usually be used as though the information came from the same source.
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.

Māori Electoral Population (MEP)	The Māori descent electoral count multiplied by the proportion of enrolled Māori voters choosing the Māori roll.
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).
MOE	Ministry of Education
MOH	Ministry of Health
MOJ	Ministry of Justice
MSD	Ministry of Social Development
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).
NZTA	New Zealand Transport Agency
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)
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 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
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.
South Island Quota	The South Island General Electoral Population divided by 16.
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% undercount).