Stats NZ
Tatauranga Aotearoa

2023 Census: Methodology for filling gaps for gender and sex concepts

Consultation technical paper: December 2022

Introduction

We are seeking feedback as we develop the methodology for using alternative data sources for producing the most complete census possible. We plan to use alternative data sources for filling gaps in census responses to the gender and sex concepts questions in the 2023 Census. We want to make sure we deliver quality data and having the most complete, representative data is key for doing this.

We have started some investigations and we are interested in feedback on options that we have identified so far. We plan to carry out further investigations as we develop the methodology.

Context

Combined census model

For the 2023 Census, to give Aotearoa New Zealand the most complete census possible, Stats NZ will be adding existing data from other sources to data collected in the census. We call this a combined census model.

A combined census model involves collecting data as part of the 2023 Census and using administrative (admin) and historical census data to fill gaps in data (if required and where the data exists). This is called a ‘combined methodology’ because it involves combining different data sets.

Admin data is data that is collected by government agencies or other organisations while conducting their normal business, such as delivering a service or recording an event.

There are gaps in admin data for key population groups and variables, so participation in census field collection remains the primary means of producing quality census data.

Alternative data sources for gender and sex concepts

As part of the combined census model, we will be using admin data to fill gaps in census response data. This methodology is created on a variable-by-variable basis, and there are options for how we can use admin data for missing responses for each variable. We would like feedback on our proposed approach to the use of alternative data sources and statistical imputation.

The four sex and gender concepts to be covered as part of this consultation are:

- gender
- sex at birth
- sexual identity
- variations of sex characteristics.

The 2023 Census content, form design and questions have been decided by the Deputy Government Statistician, Census and Collection Operations, on behalf of the Government Statistician. For more information, please see 2023 Census: Final content report and Design of forms for the 2023 Census.

We can use three alternative data sources for filling in gaps in census responses, where available.

1. Historical census responses

Information from the 2013 or 2018 Census. Both censuses asked a single question: ‘are you male/female?’ Previous censuses did not differentiate between gender and sex at birth and did not ask about sexual identity or variations of sex characteristics.

2. Admin data

Information taken from an admin data source. As part of the combined model, census respondents will be linked to the Integrated Data Infrastructure (IDI). The IDI is a large research database that includes data from a range of government agencies, Stats NZ surveys (including the 2013 and 2018 Census), and non-government organisations. The data is linked and anonymised to form the IDI. See the Integrated Data Infrastructure page for more information, including how the IDI keeps people's information safe.

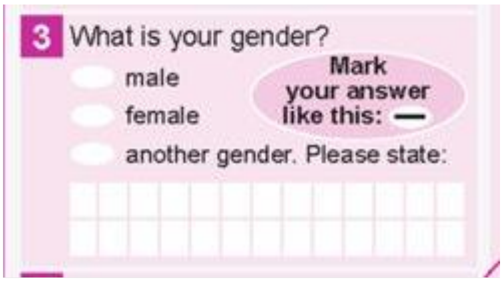
Once census responses are linked to the IDI, we will have a choice of which admin data sources to use to fill in gaps in 2023 Census variables such as gender. The choice of data source for the variables of interest will be based on the quality of the variables in the admin data sources. We will need to consider all of the six dimensions of quality – accuracy, completeness, uniqueness, timeliness, consistency and validity – as outlined in the 2018 Census data quality management strategy.

3. Statistical imputation

The term ‘imputation’ refers to the replacement of missing or unusable information with values from a statistical process, in contrast to methods of sourcing real values from admin or historical census data. Imputation is not intended to capture precise information about the specific individual but instead aims to find similar individuals and use that information to ensure a realistic distribution for those who have not responded. Therefore, consistency of individual values is not as directly relevant as it is for the other sources.

Gender

2023 Census question



3 What is your gender?

○ male

○ female

○ another gender. Please state:

Mark your answer like this: —

Existing decisions

The following decisions have been already made by the 2023 Census programme and will impact the available methodology options.

- One of gender or sex at birth must be answered on the online form.
- Multiple response answers will not be output.
- We will produce a complete time series for gender. The gender variable must be complete for all people. Where the response is not complete, we must fill in or impute that variable in some way.
- Gender is a priority one topic. Priority one topics, explained in the 2023 Census: Final content report, are those that make up the core reason for conducting a census and have the highest output quality need.
- Where gender cannot be obtained from the 2023 Census individual form, gender will be obtained from the paper dwelling form or online household summary form. If gender is not available from either of those sources, we will use a combination of the 2023 Census sex at birth response, historic census responses or admin data to source gender information.
- For the 2023 Census we will not produce any derived variables, such as cisgender and transgender statuses, and LGBTQI+ indicator, on alternatively sourced data. We will only report on complete responses given on the 2023 Census individual form.

Key questions

For gender, when we need to fill in the gaps in census responses, we have the following data sources available to us: admin data, previous responses to the census question ‘are you male/female?’, 2023 Census response to the question ‘Sex at birth’, and statistical imputation. We are seeking feedback on the following questions and recommendations:

1. What is your preferred order of use of these alternative data sources? Where should we first look for a response to the question of gender? Admin data, 2023 Census sex at birth response, or 2018 Census?
2. If we are to use historic census data, how far back can we go? How much do you think these responses are subject to change over time?
3. Are there any admin data sources that clearly distinguish between the sex at birth and gender concepts?

Options and proposed methodology

Decision G1: Should we use admin data to fill in gender?

	Pros	**Cons**
OPTION a Use combined admin sex/gender data (mixed concepts)	• Less reliance on statistical imputation and more use of information about the specific person • More accurate than statistical imputation • Potentially a timely source of gender, where gender has been updated recently	• Data sources use mixed concepts which potentially could lead to false enumeration • Individuals may not have their gender correctly recorded in admin data which also could lead to false enumeration • Under-reporting of the ‘another gender’ category
OPTION b No	• Does not introduce the cons of using combined admin sex/gender data	• Greater reliance on statistical imputation contributes to lower quality data • If we are unable to use admin data sources to fill in gaps in gender, then because we need to provide a complete time series for this variable, we must use statistical imputation

Recommendation G1: Option a

We are required to achieve full completeness for the gender variable. However, previous census responses may not reflect gender because in previous censuses, the sex question 'are you? male/female', was not clearly identified as asking about either the concept sex or gender. We expect that, in most scenarios, mixed concept admin data is likely to be more accurate than both historic census data and statistical imputation.

Decision G2: Should we use 2023 census 'sex at birth' responses to fill in gender?

	Pros	**Cons**
OPTION a Yes, use 2023 census 'sex at birth' responses	• Less reliance on statistical imputation and more use of information about the specific person • More accurate than statistical imputation • Source is respondent themselves • Respondents who did not fill out a historic census form can be enumerated with this method	• Gender will be misrepresented if the missing response would have been different to their sex at birth
OPTION b No, do not use 2023 census 'sex at birth' responses	• The respondent's gender could be different to their sex at birth response.	• Greater reliance on statistical imputation and historical census responses potentially leading to a lower quality output • No historical census data to fall back on for those who were missing or not in the country for previous censuses (2018 in particular)

Recommendation G2: Option a

For a significant portion of the population the sex at birth response will be the same as current gender. We assume that most missing gender responses can be attributable to a sex at birth response that has already been given. We also assume that if a respondent's gender was different to their sex at birth, they would include a separate gender response.

Decision G3: Should we use historical census data to fill in gender?

	Pros	**Cons**
OPTION a Yes, use prior census sex responses	- Less reliance on statistical imputation and more use of information about the specific person - More accurate than statistical imputation - Source is respondent themselves	- Prior censuses only have two categories (male/female) - Prior censuses were unclear as to whether sex or gender was being collected
OPTION b No use of prior census sex responses	- Does not introduce the cons of using historical census data	- Greater reliance on statistical imputation contributes to lower quality output - If we are unable to use admin data sources to fill in gaps in gender, then because we need to provide a complete time series for this variable, we must use statistical imputation

Recommendation G3: Option a

For a significant portion of the population, the historic census sex response will be the same as current gender. As the historic census sex question was not clear if it was collecting sex or gender, for some respondents where sex at birth and gender are not the same, the historic census response may reflect gender. For both reasons, historic census sex data will likely be more accurate than statistical imputation.

Decision G4: How should we impute the gender variable using statistical imputation?

	Pros	**Cons**
OPTION a Impute two (male/female) or three categories of gender	- Statistical imputation may improve the representativeness of the data (over missing information) contributing to higher quality data. This will be particularly important for	- We have no population level data to tell us what the distribution of the three gender options is, we will not know if the responses we are seeing are representative - May emphasise biases if we get more/less ‘another

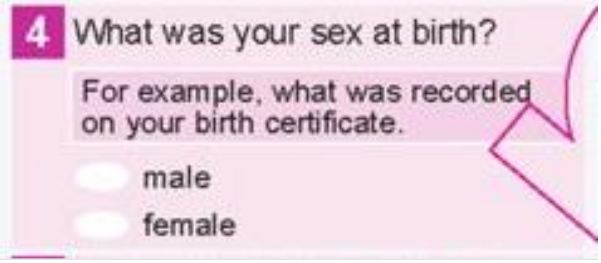
(male/female/ another gender) Note that further work is required to investigate if three category imputation is possible	the ‘another gender’ category	gender’ responses than is representative • We may not be able to identify good matching variables for donor imputation • The impact of imputation on cross tabulations of variables is unknown and will need to be investigated if imputation is done
OPTION b Don’t impute	• Records in the census dataset with no missing information are more accurate as they represent all 2023 Census responses	• It is a priority one variable and we must provide a complete dataset for gender. Where we do not have or do not use admin data or historic census data then statistical imputation is the best alternative to filling in missing information

Recommendation G4: Option a

We are required to produce a complete dataset for gender. Achieving a complete dataset for gender is not possible without imputation.

Sex at birth

2023 Census question



4 What was your sex at birth?

For example, what was recorded on your birth certificate.

○ male

○ female

Existing decisions

- One of gender or sex at birth must be answered on the online form
- Multiple responses will not be output
- Historical sex will be used for the output timeseries, as for gender
- We will use alternative data sources, that is, admin or historic census data, to fill in gaps in the response to this question
- If sex at birth response is missing on the 2023 Census individual form, then we will use historic census or admin data to source sex at birth information
- As with gender, sex is a priority one topic.

Key questions

We have similar options for sex at birth as we do for gender. However, the sex question in previous censuses did not specify sex at birth, so there may be some variation in how respondents interpreted the question. Therefore, previous census responses are unlikely to be used to fill in gaps for the sex at birth variable.

The downstream impacts of our sex at birth decisions on the ‘Number of children born’ variable will also have to be considered as it uses a subject population of Sex at birth: Female and Age: 15 and over. We are seeking feedback on the following questions and recommendations:

1. What is your preferred order of use of the methods? Where should we look for a response to the question of sex at birth? Admin data or 2023 Census gender response?
2. If we use admin data and/or 2023 Census responses and have remaining gaps, should we fill these in using statistical imputation or leave gaps in the data?
3. Are there any admin data sources that clearly distinguish between the sex at birth and gender concepts?

Options and proposed methodology

Decision SB1: Should we use admin data to fill in sex at birth?

	Pros	**Cons**
OPTION a Use combined admin sex/gender data (mixed concepts)	• Less reliance on statistical imputation and more use of information about the specific person	• Data sources use mixed concepts which potentially could lead to false enumeration • Individuals may not have their sex at birth correctly recorded in admin data

	• More accurate than statistical imputation • Potentially a timely source of sex at birth, where sex at birth has been updated in admin data recently • Better enumeration for variables such as ‘Number of children born’ in both overseas and NZ born populations	which also could lead to false enumeration
OPTION b Use selective admin data to fill in all three categories (that minimises the inclusion of mixed concepts data)	• Less missing information • More accurate than statistical imputation and using mixed concepts admin data • Potentially a timely source of sex at birth, where sex at birth has been updated in admin data recently • Better enumeration for variables such as ‘Number of children born’ in both overseas and NZ born populations	• If we don’t use mixed data, then there will be more missingness and therefore this will contribute to lower data quality • There may be no admin data sources that are not mixed concepts • Individuals may not have their sex at birth correctly recorded in admin data which also could lead to false enumeration
OPTION c No	• Does not introduce the cons of using admin data	• More missing information contributes to less representative output • Higher reliance on statistical imputation leading to less accurate unit level data, especially where the historic sex response does match their sex at birth

Recommendation SB1: Option b

We will use admin data sources that we are confident are made up of mostly sex at birth responses and will try to avoid mixed concept data sources.

Decision SB2: Should we use historical census data to fill in gaps in sex at birth?

	Pros	**Cons**
OPTION a Yes, use prior Census responses	• Less missing information • More accurate than statistical imputation • Source is respondent themselves	• Prior censuses only have two categories (male/female) • Prior censuses were unclear as to whether sex or gender was being collected • Data is not recent, potentially leading to false enumeration
OPTION b No	• Does not introduce the cons of using historical census data	• More missing information contributes to less complete output • Higher reliance on statistical imputation leading to less accurate unit level data, especially where the historic sex response does match sex at birth

Recommendation SB2: Option b.

As the historic census sex question was not clear if it was collecting sex or gender, for some respondents where sex at birth and gender are not the same, the historic census response may reflect gender. For sex at birth we recommend not to use historic census information for this reason.

Decision SB3: Should we impute the sex at birth variable using statistical imputation?

	Pros	**Cons**
OPTION a Impute	• Less missing information contributes to more complete dataset and potentially more representative output	• We have no population level data to tell us what the distribution of sex at birth responses is, and we will not know if the responses we are seeing are representative

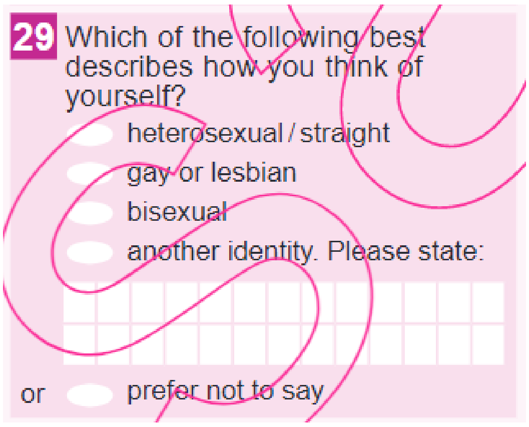
		• The impact of imputation on cross tabulations of variables is unknown and will need to be investigated if imputation is done
OPTION b Don't impute	• Records in the census dataset with no missing information are more accurate as they represent all 2023 Census responses	• More missing information contributes to less complete output and potentially lower quality output

Recommendation SB3. Option a

Sex at birth is a priority one output of wide use. Higher incidences of missing information will limit the utility of this variable.

Sexual identity

2023 Census question



29 Which of the following best describes how you think of yourself?

○ heterosexual / straight

○ gay or lesbian

○ bisexual

○ another identity. Please state:

or ○ prefer not to say

Existing decisions

- Not mandatory to respond
- Multiple responses will not be output

Key questions

We do not have historic census data for sexual identity as the question is new, and the topic has limited coverage from admin data sources. The only option would be to use imputation to reduce the missingness. We are seeking feedback on the following questions and recommendations:

1. Should we impute sexual identity?
2. Are there any admin data sources that hold this information?

Options and proposed methodology

Decision SI1: Should we impute the sexual identity variable using statistical imputation?

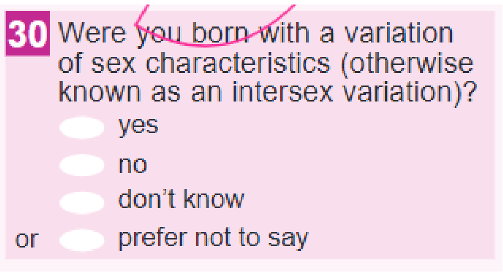
	Pros	**Cons**
OPTION a Impute	• Less missing information contributes to a more complete dataset and potentially more representative output	• We have no population level data to tell us what the distribution of the sexual identity responses is, we will not know if the responses we are seeing are representative • We are not confident in being able to identify good matching variables for donor imputation • May emphasise biases in data making the output less representative • The impact of imputation on cross tabulations of variables is unknown and will need to be investigated if imputation is done
OPTION b Don’t impute	• Records in the census dataset with no missing information are more accurate as they represent all 2023 Census responses	• More missing information contributes to less complete output and potentially lower quality output

Recommendation SI1: Option b

We do not have complete and accurate historic census data or admin data sources to fill in the gaps or understand the distribution of this variable. Without other sources of this information, we are limited in our ability to identify good matching variables for donor imputation which means the quality of the imputation is likely to be lower.

Variations of sex characteristics

2023 Census question



30 Were you born with a variation of sex characteristics (otherwise known as an intersex variation)?

○ yes

○ no

○ don't know

or ○ prefer not to say

Existing decisions

- Not mandatory to respond
- Multiple response will not be output

Key questions

We do not have historic census data for variations of sex characteristics as the question is new, and the topic has no admin data sources. The only option would be to use imputation to reduce the missingness. We are seeking feedback on the following questions and recommendations:

1. Should we impute this variable?
2. Are there any admin data sources that hold this information?

Options and proposed methodology

Decision VS1: Should we impute the variations of sex characteristics variable using statistical imputation?

	Pros	**Cons**
OPTION a Impute	• Less missing information contributes to more complete dataset and potentially more representative output	• We are not confident in being able to identify good matching variables for donor imputation • May emphasise biases in data making the output less representative

		- The impact of imputation on cross tabulations of variables is unknown and will need to be investigated if imputation is done
OPTION b Don’t impute	- Records in the census dataset with no missing information are more accurate as they represent all 2023 Census responses	- More missing information contributes to less complete output and potentially lower quality output

Recommendation VS1: Option b

We do not have complete and accurate historic census data or admin data sources to fill in the gaps or understand the distribution of this variable. Without other sources of this information, we are limited in our ability to identify good matching variables for donor imputation which means the quality of the imputation is likely to be lower.