Waipapa Taumata Rau University of Auckland

Growing Up in New Zealand

Data Output Guide

March 2026

Prepared by

Growing Up in New Zealand

Prepared for

Growing Up in New Zealand Data Users

Making Aotearoa a better place to call home

Version	Date updated	Date approved	Approved by	Reason for change
V1.0	03/02/2025	12/02/2025	Data Access Committee	N/A
V1.1	04/02/2026			Updated the following section: • 8.3 – Publications
V1.2	26/03/2026			Updated format to new *GUiNZ* style guidelines

Table of Contents

1. Purpose of this guide

This guide describes the methods and rules that researchers must use to maintain participants' confidentiality in outputs produced from *Growing Up in New Zealand* (*GUiNZ*) data. This guide also outlines the process for obtaining and publishing output(s) from the secure data platform.

If you are a researcher using *GUiNZ* data, following the rules in this guide ensures there is a high chance your output will be released, and your application to publish will be approved. These rules do not cover every eventuality. If you have any questions or concerns, please contact the Data Access Coordinator.

If you produce an output that violates a confidentiality rule or is not covered by any of the rules in this guide, you will be required to explain why the output contains no disclosure risks. Please submit your explanation during the output submission email to the Data Access Coordinator. See section 8.1 for instructions on how to submit your output for release.

All final output and publications must include the appropriate disclaimers outlined in section 8.2. If you have any questions about your output or application to publish, please contact the Data Access Coordinator.

This guideline has incorporated ideas and information from the StatsNZ Microdata Output Guide (fifth edition).

2. Principles of confidentiality for data access

GUiNZ provides approved researchers with access to de-identified research datasets. Please refer to the Data Access Protocol for the principles of access for the *GUiNZ* data.

The research datasets generated by the study constitute a valuable national resource that can have significant utility for research scholars and social policy analysts who wish to investigate questions concerning child development and family functioning. When using the *GUiNZ* data, it is of paramount importance that the privacy of study participants and their families is protected.

We treat the *GUiNZ* research datasets with extreme care and only allow access to the data under specific conditions that meet the requirements of the Data Access Protocol.

For information on the longitudinal research datasets and technical documentation, please refer to the *GUiNZ* website and *GUiNZ* Data User Guide. The Data User Guide also contains further information on the *GUiNZ* de-identification.

2.1. Why confidentiality is important

We rely on the *GUiNZ* participants' trust and goodwill to continue participating in the study so that we can produce the data for research to improve the outcomes of all people growing up in New Zealand. Therefore, maintaining privacy, confidentiality, and

data security is a core value of *GUiNZ*. We are committed to ensuring the privacy, security, and confidentiality of all our information. This includes the collection, use, storage, and distribution of the information we collect.

The terms privacy, security, and confidentiality are often used interchangeably, but each term has a different meaning:

- **Privacy** refers to the ability of a person to control the availability of information about themselves.
- **Security** refers to how the agency stores and controls access to the data it holds.
- **Confidentiality** refers to the protection of information from and about individuals and organisations and ensuring that the information is not made available or disclosed to unauthorised individuals or entities.

2.2. Goals for confidentiality

The output rules and checking processes covered in this guide are part of the risk management framework. A disclosure may occur when a person recognises or learns something they did not already know about an individual or organisation through microdata or output produced from microdata. For a disclosure to occur, this information must enter the public domain.

Microdata output rules are designed to meet the following four goals:

- **Utility** – we want the research output to be as rich, detailed, and unmodified as possible.
- **Safety** – we manage the risk of disclosure of particulars about data subjects down to the level required by our ethical obligations and the preservation of trust.
- **Simplicity** – we want the rules to be as simple to apply and check as possible.
- **Consistency** – we aim to maximise consistency across output produced by different channels and across similar output from different source collections.

The first two goals (utility and safety) lead to rules that aim to release as much detail as possible and to protect the entities that need to be protected.

We aim to maximise the potential utility of research outputs by ensuring that the unit record data provided through the *GUiNZ* secure data platform is as rich and detailed as possible and by not adding confidentiality perturbation to this data.

2.3. Legislative requirements

GUiNZ is required to protect the information we collect. These requirements are outlined in the Privacy Act 2020 and in the General Data Protection Regulation 2016.

3. The Five Safes

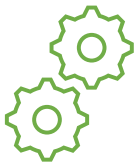
Data access is only provided to the research datasets if all of the 'five safes' conditions are met.

Safe People

Researchers are vetted through the Data Access process to ensure they have the capability to use *GUiNZ* data. Once approved, researchers are required to sign a contract where they agree to follow our rules and protocols to ensure a commitment to data safety. Researchers who break our protocols can be banned, blacklisted, or prosecuted.

Safe Projects

To gain access to the *GUiNZ* research datasets, researchers must have a research project they can demonstrate is in the public interest and aligns with *GUiNZ* principles. Please find more information on the principles in the Data Access Protocol.



Safe Settings

GUiNZ research datasets can only be accessed through a secure data platform (AWS Workspace). A variety of security layers further protect the information:

- Multifactor authentication is required to access AWS Workspace.
- Workspaces are not connected to the network.
- Users cannot transfer information in or out of their AWS Workspace without approval from the *GUiNZ* data access team.

Safe Data

The *GUiNZ* research datasets have had identifying information removed, and researchers only receive access to the specific datasets they need for their approved projects. This means that data available to researchers is de-identified. Participants receive an encrypted ID, which can be linked to each data collection wave without using identifying information (e.g., names and addresses).

Safe Outputs

All information released from the secure data platform is checked to ensure it does not contain any identifying information. Results that could potentially identify participants will not be released. The following describes the methods and guidelines researchers use to maintain confidentiality of their outputs produced from the *GUiNZ* research datasets.

4. Considerations for analysis and interpretation of data by ethnicity

The research proposal must demonstrate how it protects and promotes Māori rights and aspirations in alignment with the *GUiNZ* Kaitiaki principles (Paine et al., 2022). The research proposal should demonstrate the relevance of the research to Māori and how it can contribute to informing the elimination of ethnic inequities (Reid et al., 2017) and/or apply other criteria for strengthening the reporting of health research involving Indigenous people (Huria et al., 2019).

Fundamental to the overarching intent of *GUiNZ* is the recognition that ethnicity is a complex, multi-faceted and fluid construct, and its conceptualisation, measurement, use in analysis and interpretation require careful considerations (Atatoa Carr et al., 2022; Yao et al., 2021; Yao et al., 2022). Ethnicity (or ethnic identification) is widely used to measure and analyse differences between population subgroups for research and policy purposes. The use and interpretation of the *GUiNZ* ethnicity data should therefore demonstrate considerations to uphold rights for Māori in research (Paine et al., 2020). The research proposal should clearly describe and justify the use of the *GUiNZ* ethnicity variables and data including grouping and classification (Statistics New Zealand, 2005), and consider representation of multiple ethnic identification patterns and groups that are typically under-represented in research (e.g. Cormack & Robson, 2011; McLeod & Harris, 2023).

The research team should further demonstrate that it has involved adequate expertise to inform the proposed approach, methods, analysis and interpretation of data by ethnicity.

5. Descriptive statistics

Cell count refers to the number of observations (i.e., children or mothers) that possess certain characteristics.

Data users must use a series of statistical methods to protect the confidentiality of *GUiNZ* data. First users are required to protect “sensitive” cells using suppression or aggregation, after which they are required to use perturbation to add random noise into data outputs.

A cell is considered "sensitive" when knowledge of the value would permit an unduly accurate estimate of the contribution of an individual. In the case of the *GUiNZ* data, a cell is deemed sensitive when it has a count of less than **10** (i.e., less than 10 individuals have contributed to the cell).

5.1. Cell suppression

Suppression (primary suppression) is the removal of a cell's value when it has been deemed sensitive.

When suppressing output, use consistent notation such as <10 or 'S'.

7.1.1 When sensitive cells occur, other cells or marginal totals in the table must also be suppressed to stop the first cell from being recalculated. This later stage is called secondary suppression. This is important because if secondary suppression is not applied when appropriate, a user can recalculate the sensitive cell's value.

7.1.2 There are no universal guidelines for applying secondary suppression, except there must be enough secondary suppression to ensure primary suppressed values cannot be derived. Where secondary suppression is required, cell counts may be rounded **up** to the nearest multiple of 10 (i.e., n < [multiple of 10]).

7.1.3 An Excel macro to perform suppression is available. Secondary suppression will need to be applied manually. If using R both primary and secondary suppression will need to be applied manually using the R suppression function.

Example:

Before *suppression is applied (before perturbation)*

	Group 1	Group 2	Group 3
Group 1	11	47	58
Group 2	27	32	33
Group 3	4*	31	20
Total	**42**	**110**	**111**

After *suppression has been applied (before perturbation)*

	Group 1	Group 2	Group 3
Group 1	<20^	47	58
Group 2	27	32	33
Group 3	<10*	31	20
Total	**42**	**110**	**111**

**This cell is primary suppressed. ^This cell has been secondary suppressed.*

5.2. Aggregation

Aggregation is a method for protecting sensitive cells by collapsing categories.

Example:

Before *aggregation is applied (before perturbation)*

	Group 1	Group 2	Group 3
Group 1	11	47	58
Group 2	27	32	33
Group 3	4*	31	20
Total	**42**	**110**	**111**

**Sensitive cell*

After *aggregation has been applied (before perturbation)*

	Group 1	Group 2	Group 3
Group 1	11	47	58
Group 2 *(prev. 2&3)*	31	63	53
Total	**42**	**110**	**111**

The sensitive cell has been aggregated into a group.

5.3. Perturbation

Perturbation is a best-practice method and works by adding a random value to the data, to mask the data. This is called adding "random noise".

5.3.1 All cell counts must be randomly rounded to base 3 which involves randomly changing each count in a table to a multiple of 3.

5.3.2 Apply random rounding to base 3 by rounding values to the nearest multiple of 3 with a probability of 2/3, the second nearest multiple of 3 with a probability of 1/3, and leaving values that are already multiples of 3 unchanged.

Example: An original (unrounded) count of 17 would be rounded to 18 with a probability of 2/3 and rounded to a probability of 15 with 1/3.

5.3.3 When producing the same count in the same cell, apply the rounding to the count in the same direction, even if it is the same cell but in a different set of output. This check must be done manually.

5.3.4 Marginal totals of these counts can be independently and randomly rounded to base 3. Alternatively, you can calculate marginal totals by summing the rounded counts.

5.3.5 Macros that perform random rounding to base 3 in RStudio and Excel are available from *GUiNZ*.

To use the R function: The user is expected to have their tabular output in vector or data frame format and follow the instructions to apply the "rrn" function script to the output. The R macro will allow you to apply base 3 rounding only.

To use the Excel macro: The user is expected to copy the tabular output to the Sheet1 tab and follow the instructions in the Excel macro. The Excel macro will do both suppression and base 3 rounding.

Example:

Before *random rounding to base 3 is applied (with suppression applied)*

	Group 1	Group 2	Group 3
Group 1	<20	47	58
Group 2	27	33	33
Group 3	<10	33	21
Total	**42**	**110**	**111**

After *random rounding to base 3 is applied (with suppression applied)*

	Group 1	Group 2	Group 3
Group 1	<20	48	57
Group 2	27	32	33
Group 3	<10	31	18
Total	**42**	**108**	**111**

5.4. Value magnitudes (cell totals and means)

Value magnitudes refer to measures (cell totals and means) from a numerical variable (e.g., household income). If a tabular output contains information about individuals or households, the individuals or households need protection to ensure their contribution to a value magnitude cannot be estimated with accuracy. To protect individuals and households, you must calculate value magnitudes from at least 10 observations.

5.4.1 Value magnitudes (cell totals and means) should be suppressed if the unrounded cell count (i.e., the number of individuals contributing to that value magnitude) is less than 10. Calculate means from **rounded** counts.

5.5. Maximum and minimum values

Maximum and minimum values are normally suppressed. Maximum and minimum values are respondent values and may be outliers that pose a high risk of disclosure.

5.5.1 Aggregation of maximum and minimum values into variable groups could be considered.

5.5.2 This rule also applies to any descriptions of the data in the methods section.

5.5.3 Where a maximum or minimum value is not identifying, it may be considered for release. If you produce a maximum or minimum value that you believe is not identifying, please provide an explanation in your submission form.

Example: Income and expenditure are examples of sensitive variables. Maximum and minimum values for these variables are normally suppressed.

5.6. Medians, quantiles, and percentiles

5.6.1 Suppress medians if the unrounded cell count is less than 10. For other quantiles and percentiles, use the table below to find out how many observations are needed for each quantile or percentile.

Quantile or percentile	Number of observations needed overall
0.01	500
0.05	100
0.10	50
0.25	20
0.50	10
0.75	20
0.90	50
0.95	100
0.99	500

5.7. Percentages, proportions, and ratios

5.7.1 You must derive all percentages, proportions, and ratios (including odds ratios) using the rounded counts.

5.7.2 Round percentages to 1 decimal place.

5.7.3 Suppress percentages, proportions, or ratios where either or both of the cell counts used to calculate the percentage, proportion, or ratio have been suppressed.

6. Analytical output

6.1. Regression models

Regression output does not usually have confidentiality issues, except in the circumstances listed below. However, you must ensure that pieces of output are not equivalent to statistics subject to other confidentiality rules, particularly low cell counts or statistics based on low cell counts.

6.1.1 Regression output may contain counts or lead to the calculation of counts. Suppress regression output if the underlying cell count is less than 10.

6.1.2 Classification and regression tree models may produce the equivalent of detailed count tables. If this occurs, suppress cell counts less than 10.

6.1.3 Regression outputs that are equivalent to other forms of output need to have the relevant rules applied. For example, coefficients produced by ordinary least squares regressions with binary (0/1) right-hand-side variables are equivalent to cell means and, therefore, need to comply with the means rule.

7. Graphical output

The same suppression rules as in the above sections apply to graphical outputs, including secondary suppression and perturbation.

There are four main types of graphs:

Type A: Graphs produced from aggregated data or tables that have had confidentiality rules applied (e.g., frequency histograms, bar charts of magnitudes).

Type B: Graphs produced directly from the unit record data but aggregated in the process by the software (e.g., frequency histograms, kernel density plots).

Type C: Graphs produced directly from the unit record data and displaying unit record values (e.g., scatterplots, residual plots).

Type D: Graphs produced from the results of modelling or derivation that use the unit record data (e.g., regression curves).

You can format graphs in the following ways:

- Static – the graph is simply a picture with no data attached.[1]
- Interactive – can be modified by the software that contains the data.

7.1.1 Release type A graphs in either static or interactive format.

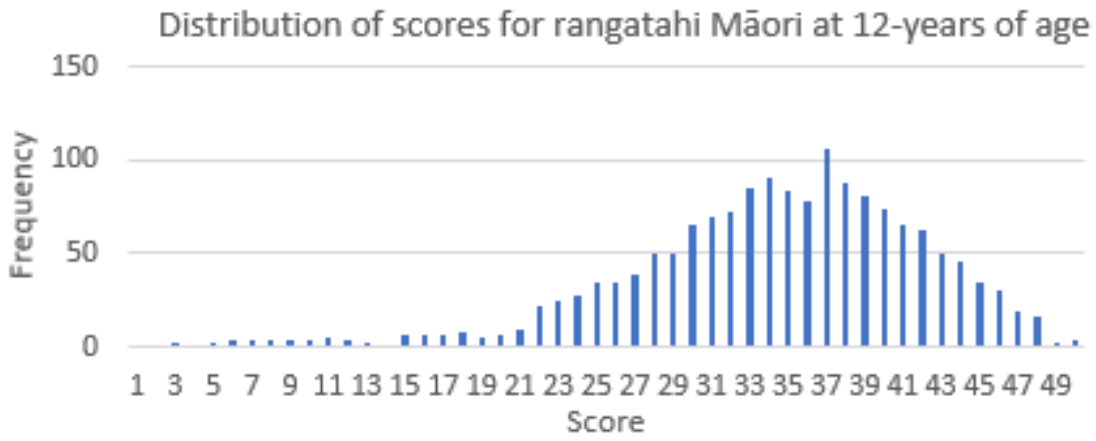
7.1.2 Release type B graphs in static format, only if the graph provides a high level of uncertainty.2

7.1.3 Release type C graphs in static format after further processing at the discretion of the output checker. For this type of graph to be released, you need to ensure that individuals cannot be recognised and that values can only be estimated with a high level of uncertainty.2 Further processing can include but is not restricted to cutting off the tails of a distribution, removing outliers, jittering the actual values, and removing or modifying axis values.

7.1.4 Release type D graphs either in static or interactive format, but only if the values shown in the graph cannot be used to find the original data values (i.e., where the modelling or derivation cannot be reversed to find the original value for each individual).

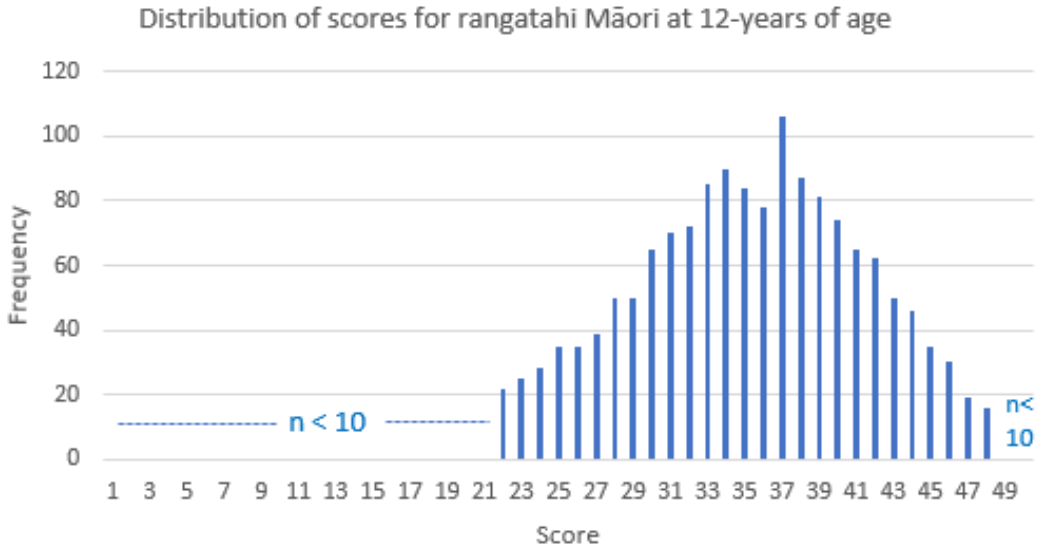
Example:

Before suppression is applied



[1] When graphs are released in this format, you need to ensure that the points on the graph cannot be recalculated in some way (e.g., by counting the pixels).
[2] The level of uncertainty is high if the level of uncertainty about the data values is equal to or larger than that in the table for which the confidentiality rules have been applied. For graph type B you do not need to provide the underlying data, but you do need to include a justification in your submission form explaining why the graph has a high enough level of uncertainty to be released. For graph type C, you need to include the underlying data and a justification in your submission form.

After suppression has been applied



8. Data output release and publication

8.1. Data output release

Once your analyses have been completed, and before you can request that the outputs be released from the Secure Data Access Platform, please ensure that all outputs submitted for release meet the confidentiality requirements outlined. **To ensure that the confidentiality requirements have been applied correctly (when necessary), users should provide:**

1) Both raw and processed outputs for ease of comparison (*see example below*)
2) For any graphical outputs, the associated counts used should be clearly displayed in a table so that they can be reviewed

Example:

	Counts (raw)				Counts (processed)		
	Group 1	**Group 2**	**Group 3**		**Group 1**	**Group 2**	**Group 3**
Group 1	11	47	58	**Group 1**	<20	48	57
Group 2	27	32	33	**Group 2**	27	32	33
Group 3	4*	31	20	**Group 3**	<10	31	18
Total	**42**	**110**	**111**	**Total**	**42**	**108**	**111**

Output requests are to be submitted to the Data Access Coordinator at dataaccess@growingup.co.nz and should specify the output file name and file path. The *GUiNZ* team will review the output file and either approve the release request or provide feedback if the output does not meet the specified requirements. If the request is approved, the Data Access Coordinator will release the file(s) to you via email. The output checking process may take up to 5 working days.

8.2. Disclaimer

All final output produced using *GUiNZ* data must include the following and disclaimer (attach to your output(s) requested from your Workspace):

> *"Access to the data used in this study was provided by Growing Up in New Zealand (GUiNZ) under conditions designed to uphold the security and confidentiality provisions of the GUiNZ Data Access Protocol. The results presented in this study are the work of the author(s) and do not necessarily represent the view(s) of GUiNZ."*

8.2.1 If your publication contains data that has been suppressed (primary or secondary), you may wish to include the following disclosure statement to inform the publisher.

> *"Access to the data used in this study was provided by Growing Up in New Zealand (GUiNZ) under conditions designed to uphold the security and confidentiality provisions of the GUiNZ Data Access Protocol. These provisions outline the requirement for primary and secondary suppression to maintain confidentiality and to protect participant data."*

8.2.2 If your publication requires description of the Growing Up in New Zealand study, you may wish to include the following description to inform the publisher.

> *"Growing Up in New Zealand is Aotearoa's largest longitudinal study of child health and wellbeing, following the lives of more than 6000 rangatahi and their families. The rich information collected by the study is helping create flourishing futures for all young people in Aotearoa. growingup.co.nz"*

8.3. Publications

Investigators agree to:

8.3.1 Provide a copy of all publications (journal articles and research reports) through the Data Access Coordinator at dataaccess@growingup.co.nz.

8.3.2 Use the standard descriptor of the Growing Up in New Zealand study outlined in Section 8.2.2 of this Data Output Guide.

8.3.3 Include the disclaimer outlined in Section 8.2 of this Data Output Guide.

8.3.4 Manage small cell sizes in accordance with the procedures outlined in Section 5 of this Data Output Guide.

8.3.5 Interpret subgroup analyses by ethnicity according to the guidelines outlined in Section 4 of this Data Output Guide.

8.4. Publicity

Once your publication has been approved, *GUiNZ*'s communication manager may be in touch with you to discuss potential media opportunities. Where appropriate, we encourage media coverage of *GUiNZ* papers to raise the study's profile and to show study families that the study is producing interesting and valuable findings. However, you must obtain approval from the *GUiNZ* communications manager before distributing a press release or giving press interviews or comments.

Your publication will also be linked on the *GUiNZ* website once it has been published.

9. References

Atatoa Carr, P., Langridge, F., Neumann, D., Paine, S.-J., Liang, R., Taufa, S., Fa'alili Fidow, J., Fenaughty, J., & Kingi, T. K. (2022). 'Seeing'our tamariki in longitudinal studies: exploring the complexity of ethnic identification trajectories within Growing Up in New Zealand. Journal of the Royal Society of New Zealand, 1-17.

Cormack, D., & Robson, C. (2011). Classification and output of multiple ethnicities: considerations for monitoring Māori health. Te Rōpū Rangahau Hauora a Eru Pōmare.

General Data Protection Regulation 2016 (E.U.). https://gdpr-info.eu/

Growing Up in New Zealand. (2017). Data Access Protocol V11. The University of Auckland. https://cdn.prod.website-files.com/63a70013e473f3b2807218ee/63c1923084d7a21ce8b37c1c_DA_protocol_V11.pdf

Growing Up in New Zealand. (2024). Data User Guide: May 2024. The University of Auckland. https://cdn.prod.website-files.com/63a70013e473f3b2807218ee/66fde3aa2fed68bc48fb3e5b_DCW12_Data_User_Guide_September_2024v2.pdf

Huria, T., Palmer, S. C., Pitama, S., Beckert, L., Lacey, C., Ewen, S., & Smith, L. T. (2019). Consolidated criteria for strengthening reporting of health research involving indigenous peoples: the CONSIDER statement. BMC Medical Research Methodology, 19, 1-9.

McLeod, M., & Harris, R. (2023). Considerations for Māori Data Analyses. A report for Te Aka Whai Ora: Māori Health Authority. https://www.tewhatuora.govt.nz/assets/Publications/Maori-health/Ethnicity-analysis-report-Sept-2023.pdf

Morton, S.M., Atatoa Carr, P.E., Grant, C.C., Robinson, E.M., Bandara, D.K., Bird, A., Ivory, V.C., Kingi, T.K., Liang, R., Marks, E.J., Perese, L.M., Peterson, E.R., Pryor, J.E., Reese, E., Schmidt, J.M., Waldie, K.E., & Wall, C. (2012). Cohort Profile: Growing Up in New Zealand. International Journal of Epidemiology, 42(1), 65-75. https://doi.org/10.1093/ije/dyr206

National Ethics Advisory Committee. (2021). National Ethical Standards. https://neac.health.govt.nz/ national-ethical-standards

Paine, S.-J., Cormack, D., Reid, P., Harris, R., & Robson, B. (2020). Kaupapa Māori-informed approaches to support data rights and self-determination 1. In Indigenous data sovereignty and policy (pp. 187-203). Routledge.

Paine, S. J., Neumann, D., Langridge, F., Peters, A., & Kingi, T. K. (2022). Kaitiakitanga – principles for protecting and promoting tamariki and rangatahi wellbeing in Growing Up in New

Zealand. Journal of the Royal Society of New Zealand, 52(3), 254–264. https://doi.org/10.1080/03036758.2022.2066142

Privacy Act 2020 (N.Z.). https://www.legislation.govt.nz/act/public/2020/0031/latest/LMS23223.html

Reid, P., Paine, S.-J., Curtis, E., Jones, R., Anderson, A., Willing, E., & Harwood, M. (2017). Achieving health equity in Aotearoa: strengthening responsiveness to Māori in health research. New Zealand Medical Journal, 130(1465), 96-103.

Shulruf, B., Morton, S., Goodyear-Smith, F., O'Loughlin, C., & Dixon, R. (2007). Designing Multidisciplinary Longitudinal Studies of Human Development: Analyzing Past Research to Inform Methodology. Evaluation & the Health Professions, 30(3), 207-228. https://doi.org/10.1177/0163278707304030

Statistics New Zealand. (2005). Statistical Standard for Ethnicity. Statistics New Zealand.

Statistics New Zealand. (2020). Microdata output guide (fifth edition). https://www.stats.govt.nz/assets/Methods/ Microdata-Output-Guide-2020-v5-Sept22update.pdf.

Statistics New Zealand. (2020). Ngā Tikanga Paihere: a framework guiding ethical and culturally appropriate data use. https://data.govt.nz/assets/data-ethics/Nga-Tikanga/Nga-Tikanga-Paihere-Guidelines-December-2020.pdf

Yao, E. S., Meissel, K., Bullen, P., Carr, P. A., Clark, T. C., & Morton, S. M. (2021). Classifying multiple ethnic identifications. Demographic Research, 44, 481-512.

Yao, E. S., Meissel, K., Bullen, P., Clark, T. C., Carr, P. A., Tiatia-Seath, J., Peiris-John, R., & Morton, S. M. (2022). Demographic discrepancies between administrative-prioritisation and self-prioritisation of multiple ethnic identifications. Social Science Research, 103, 102648.