April 2021

Ngā Tapuae technical guidance

Analysing Māori student transitions in the Ngāi Tahu takiwā

SOCIAL WELLBEING AGENCY | TOI HAU TĀNGATA

New Zealand Government

Authors

Marianna Pekar, Joel Bancolita

Acknowledgements

Simon Anastasiadis[1], Dr Charles Sullivan[1], Louise Pirini[1], Dr Porsha London[2], Dr Eruera Tarena-Prendergast[2], Jo Aldrige[2], Maui Brennan[2], Madison Simons[2], Delane Luke[2], Integrated Data Unit[3]

1=Social Wellbeing Agency, 2=Tokona Te Raki, 3=Statistics New Zealand

Creative Commons Licence

Liability

Citation

Social Wellbeing Agency 2021. Ngā Tapuae technical guidance. Wellington, New Zealand.

ISBN 978-0-473-57413-0 (online)

Published in April 2021 by
Social Wellbeing Agency
Wellington, New Zealand

Integrated Data Infrastructure Disclaimer

The results in this paper are not official statistics, they have been created for research purposes from the Integrated Data Infrastructure managed by Statistics New Zealand. The opinions, findings, recommendations and conclusions expressed in this paper are those of the author(s) not Statistics NZ, or other government departments.

Access to the anonymised data used in this study was provided by Statistics NZ in accordance with security and confidentiality provisions of the Statistics Act 1975. Only people authorised by the Statistics Act 1975 are allowed to see data about a particular person, household, business or organisation and the results in this paper have been supressed to protect these groups from identification.

Careful consideration has been given to the privacy, security and confidentiality issues associated with using administrative and survey data in the Integrated Data Infrastructure. Further detail can be found in the Privacy impact assessment for the Integrated Data Infrastructure available from www.stats.govt.nz.

The results are based in part on tax data supplied by Inland Revenue to Statistics NZ under the Tax Administration Act 1994. This tax data must be used only for statistical purposes, and no individual information may be published or disclosed in any other form, or provided to Inland Revenue for administrative or regulatory purposes.

Any person who has had access to the unit record data has certified that they have been shown, have read, and have understood section 81 of the Tax Administration Act 1994, which relates to secrecy. Any discussion of data limitations or weaknesses is in the context of using the Integrated Data Infrastructure for statistical purposes and is not related to the data's ability to support Inland Revenue's core operational requirements.

These results are not official statistics. They have been created for research purposes from the Integrated Data Infrastructure (IDI) which is carefully managed by Stats NZ. For more information about the IDI please visit https://www.stats.govt.nz/integrated-data/.

The results are based in part on tax data supplied by Inland Revenue to Stats NZ under the Tax Administration Act 1994 for statistical purposes. Any discussion of data limitations or weaknesses is in the context of using the IDI for statistical purposes, and is not related to the data's ability to support Inland Revenue's core operational requirements.

Contents

Purpose

This technical guide describes the quantitative analysis techniques we applied in a mixed method research project. The project was co-designed and conducted by the Social Wellbeing Agency (the Agency) and Tokona Te Raki, a Ngāi Tahu-led collaborative established to increase Māori participation, success and progression in education and employment outcomes. This guide provides guidance on the methodology and direction on how to use the accompanying codes published on GitHub.

Previous applications of the quantitative methodology have been published by the Agency (SWA, *Representative timelines – modelling people's life experiences.* 2019). This guide should therefore be read in conjunction with the Agency's previous publication, which provides in-depth information about the origins, details, strength, and weaknesses of the method.

A comprehensive report including the results of the qualitative findings will be published by Tokona Te Raki with the Agency's input and oversight on quantitative results.

Research co-designed with iwi

The Agency is working together in partnership with Ngāi Tahu iwi.[1] More specifically, we engaged with Tokona Te Raki to participate in joint mixed methods research which focuses on data discovery to learn actionable insights to help young Māori to succeed. Tokona Te Raki named this research Ngā Tapuae (Stepping Stones) to emphasise the search for positive and hopeful journeys for rangatahi (young people).

Finding what helps for young Māori

The research question was to find the most important barriers, levers and boosters that help young Māori to succeed. From this the research objectives were formulated as:

- Visualising the complexity of life for all Māori in the Ngāi Tahu takiwā (tribal area)[2] during the critical life stage of leaving secondary education and transitioning into the next stages of life: further studies, employment or other pathways.
- Incorporating qualitative techniques to ensure that:
  - The results of exploratory data analysis are validated by rangatahi and whānau voice.
  - We go where purely quantitative data cannot go: provide further insights and understanding.
- Enabling compelling storytelling that empowers by providing evidence and clarity needed to focus on the most important systemic levers to create the conditions for young Māori to unleash their potential.
- Providing insights to stakeholders that can be actioned, e.g. when designing interventions.

[1] An iwi, also referred to as a Māori tribe, is one of the largest Māori kinship groupings. It is generally made up of several hapū that usually descend from a common ancestor. Hapū are clusters of whānau; the whānau usually consists of children, parents, grandparents, and other closely related kin. Source: Stats NZ.

[2] More information is provided in the chapter Population of interest.

- Developing data capability of rangatahi by involving Ngāi Tahu interns in the project work, leading by example and teaching while doing.

Combining quantitative research and lived experience

The research builds on the existing evidence base investigating the need for better transition pathways for Māori aged 15-29 years. Our team has used mixed methods to answer the research questions. The qualitative component used a Kaupapa Māori Method (Smith, 1999) alongside inductive analysis. The quantitative analysis is based on New Zealand's Integrated Data Infrastructure (IDI) to apply social sequence analysis with optimal matching to identify and summarise wellbeing outcomes to discover journey patterns leading to these outcomes.

The Agency proposed to group sequences of social events, such as education, training, and employment using an optimal matching algorithm and then summarising similar experiences. Optimal matching allows the calculation of similarity between sequences, based on which similar timelines are clustered together. To create representative timelines of the individuals' life experiences in each cluster, a greedy algorithm was used to determine the typical timing of each event (measure) for each clusters. A mathematical description of this process is described in SWA (2019).[3] The output is a longitudinal view of a population's early career outcomes and their interactions with social services. The output is presented as counts of events or people and does not imply causality or statistical significance.

Tokona Te Raki facilitated the qualitative research components to provide a better understanding of events and experiences not observed in the data. Focus group interviews were conducted before and after the quantitative results were revealed to the research team and discussed during a participatory data analysis phase or "data party".

We aim to equip rangatahi with future-focussed insights

Our aim through this research was to equip rangatahi, whānau and Rūnaka (tribal councils) with future-focused insights. This includes:

- Gaining insights into the challenges and opportunities of rangatahi, so they determine and navigate their education and career pathways towards wellbeing.
- Discover actionable insights that lead to success while giving the opportunity for viewers of the data to observe how success might look like for them.
- Along with a secondary outcome to build stronger relationships between the Agency and Ngāi Tahu to better understand how we can work together with iwi to advance shared aspirations for change.

[3] SWA 2019, pp 14-18 and Appendix

Applying the Agency's representative timelines methodology

This section explains the analytic techniques applied in the quantitative research. Guidance on assembling the codes to conduct the analysis in the IDI is available on GitHub.[4]

Reusing our representative timeline codes

The Agency has previously completed research using cross-sector representative timelines of families' experiences around the birth of a child in South Auckland. We co-designed this methodology with an NGO, The Southern Initiative (TSI), who conducted the qualitative research and published results. This work shows that the quantitative methodology creating representative timelines is effective at drawing information together to generate new questions and insights. The techniques to construct representative timelines can be applied to a range of situations. The Agency has developed a set of reusable codes to construct the timelines as well as codes to develop an application to view the timelines and make interpretability easy for non-technical audiences.

This method was adapted to form the basis of our quantitative analysis. A number of events in each person's life were recorded with a description, start and end dates: e.g. attending high school, paying taxes while working in a certain industry, earning within a certain salary band, having a child, living in a certain location, having a health issue. The methodology removes the complexity of the different length of these timelines as well as summarises people's journeys when those journeys are similar. Although this methodology has been developed in the context of the IDI, the approach is applicable in general.

The literature supports the application of the method for studying educational transitions

SWA (2019) presents a brief literature review on the application of the method in social sciences. For Ngā Tapuae our research team completed a complementary literature review focussing on the application of this technique, studying early careers and educational pathways.

A number of the studies we reviewed conclude that sequence analysis applying optimal matching or alignment techniques produce interpretable results in a wide variety of areas, including career pathways or occupational trajectories (Abbott et al, 1990; Abbott et al, 2000 and Halpin et al, 2004). By itself it should only be used as an "exploratory data analysis" technique and it does not imply causality (Halpin et al, 2004).

While this research did not focus on exploring causal relationships between characteristics of groups and their outcomes, there are ways of expanding the research scope and investigating causality. In the literature the identification of trajectories with sequence and cluster analysis are often complemented with the application of multinomial logit models to reveal characteristics associated with different pathways,

[4] https://github.com/nz-social-wellbeing-agency/nga_tapuae

such as age, gender, work type, marital status, number of children, education (Fuller et al, 2014; Struffolino, 2019 and Biemann, 2012). For example, results of such regression estimates show that women—especially if lower-educated—are more likely than men to experience pathways characterised by instability and more precarious early careers (Struffolino, 2019).

Besides changing the study population and measures of interest, there were some additional minor modifications to the original timeline codes, informed by the literature reviewed. Studying the dynamics of employment trajectories requires a longitudinal perspective, therefore the time window for analysis is longer that in the original work completed by the Agency.[5] The time increments that make up a sequence event consequently also need to be longer.[6]

Modifications to the original timelines codes

Customising the codes for Ngā Tapuae takes the following steps:

1. Defining the target population to be analysed and determining the journey events. These events have start and end dates, and a description.
2. Individual timelines are 8 and 15 years long, depending on the sub-population analysed.[7] Time increments are created to account for each event having a different length. In this case we have quarters (13 weeks for simplicity). In each quarter (for each person) we record whether an event happened, based on the event's start and end dates.
3. Individual journeys carry a great deal of information, however, to comply with microdata output checking rules they cannot be observed outside of the IDI environment for confidentiality reasons (Stats NZ, 2020). Therefore the journeys of similar individuals – and some groups of interest – will have to be summarised. The methods we use are as follows:
    - 3.1 Algorithmic clustering: the technique is described in SWA (2019). The clustering methodology uses distance measures from sequence analysis, which is a collection of techniques for working with ordered series of data. The use of algorithmic clusters helps us identify the most common types of experience.
    - 3.2 Manual clustering: timelines of populations of specific interest to Tokona Te Raki were also analysed and representative timelines were created for them, e.g. rangatahi who attended Kura Kaupapa at any point during their last three years of high school, and students in mainstream schools, or females and males who became parents and who did not. The use of manual clusters helps us understand the experience of specific groups of interest.

[5] Detailed employment trajectories for 5 years (Fuller, 2014), for 7 years (Struffolino, 2019), at least for 14 years (from the age of 22 to 36-60 years, Blair-Loy, 1999) and for 20 years (Halpin et al, 2004 and Biemann et al, 2012).

[6] Examples include time intervals of 3 months (Halpin et al, 2004) to 12 months (Biemann et al, 2012).

[7] More explanation on subpopulations are provided in the next chapter, Population of interest.

The research focusses on young Māori in the Ngāi Tahu takiwā

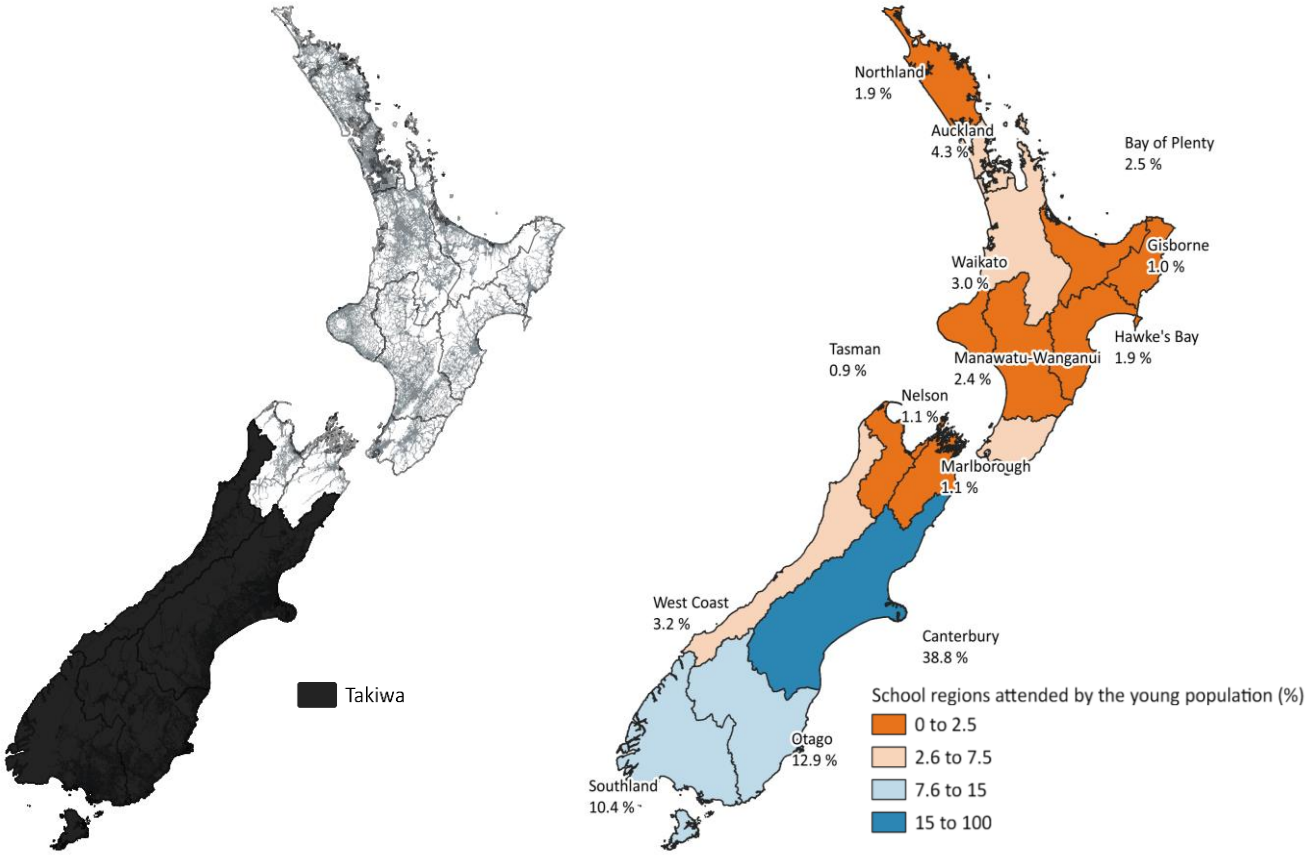
This section describes components of the research, from determining the target population in the IDI to displaying the research outputs.

Population of interest

Tokona Te Raki wanted to analyse the lives of rangatahi in the Ngāi Tahu takiwā from the last years of high-school into their early careers. To ensure that consistent, comparable, and relevant data is analysed, two sub-populations were constructed:

1. Both populations include individuals who
    a. Are of Māori ethnicity, and
    b. Attended high school in the Ngāi Tahu takiwā, for simplicity defined as the South Island except Nelson, Marlborough and Tasman. In the absence of high school information, we identified rangatahi having an address in the takiwā between the ages of 15-18. Figure 1 shows that this definition allows the inclusion of rangatahi, who attended high school only partially in the takiwā and also in other regions of New Zealand (or overseas).

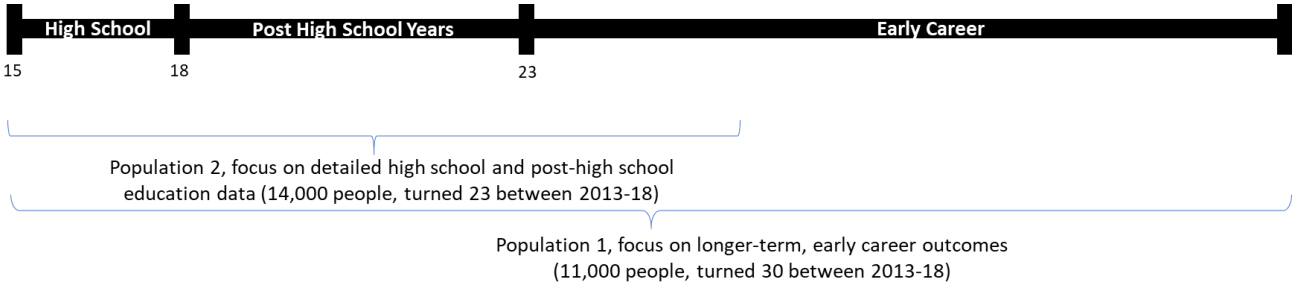
Figure 1. Approximate area of the Ngāi Tahu takiwā as determined in the IDI and high school attendance of Population 2 by regions



2. Population one (approximately 11,000 individuals) includes individuals who turned 30 years old in any of the years in 2013-2018. This group's experience is not as recent. A 15-year-long journey is available for them between the ages 15 and 30. Due to limitations in Ministry of Education (MoE) Data in the IDI limited information is available about their high-school performance and subject choices.
3. Population two (approximately 14,000 individuals) includes: individuals who turned 23 years old in any of the years in 2013-2018. This group's experience is more recent. An 8-year-long journey is available for them: between the ages 15 and 23. Rich data is available about their high-school studies due to MoE data availability.

Figure 2 displays the difference in age definitions and time windows of analysis for the two subpopulations.

Figure 2. Summary of sub-populations analysed



The location information was obtained differently for the two population groups as MoE secondary education data, including the location of high school attended, is only available for a subset of Population 1, as shown in Table 1.

Table 1. Relevant Ministry of Education data availability in the IDI (2000-2020)

Data Source	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	…	2020
Industry training		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Primary education								✓	✓	✓	✓	
Secondary education					✓	✓	✓	✓	✓	✓	✓	✓
Targeted training		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Tertiary Education	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Population 1 attending secondary school

When secondary school location was missing, address notifications in the takiwā between ages 15-18 were considered. This definition resulted a mix of location sources for individuals in the two sub populations (Figure 3 and Figure 4), which is to be considered when comparing analysis results of these two groups. The population displayed in these figures excludes individuals overseas, however subsequent analysis includes youth who spent time overseas during the analysis period.

Figure 3. Location source of Māori youth who turned 23 in reference years 2013-2018

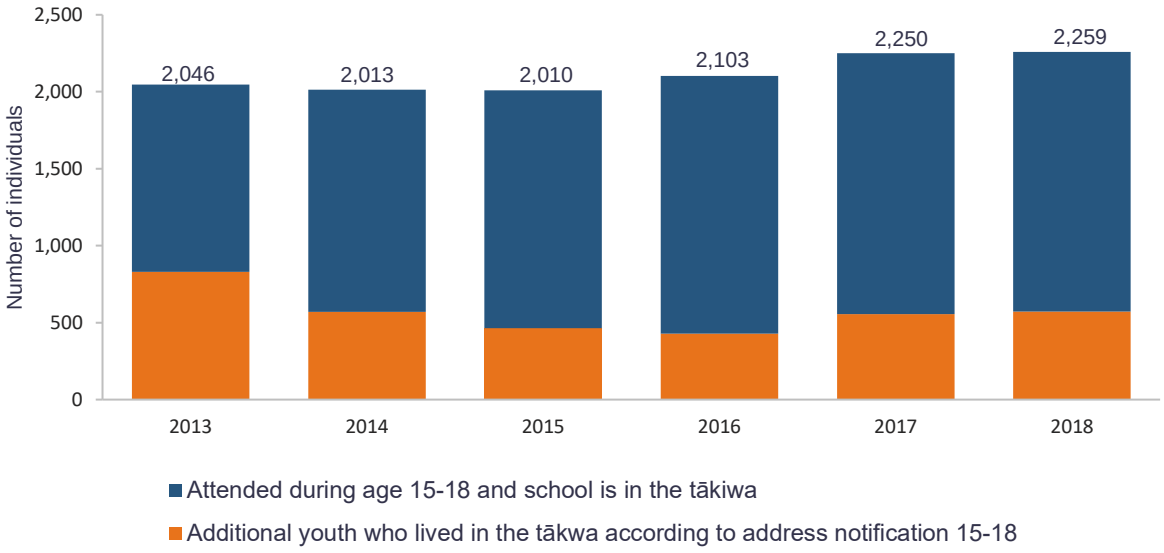
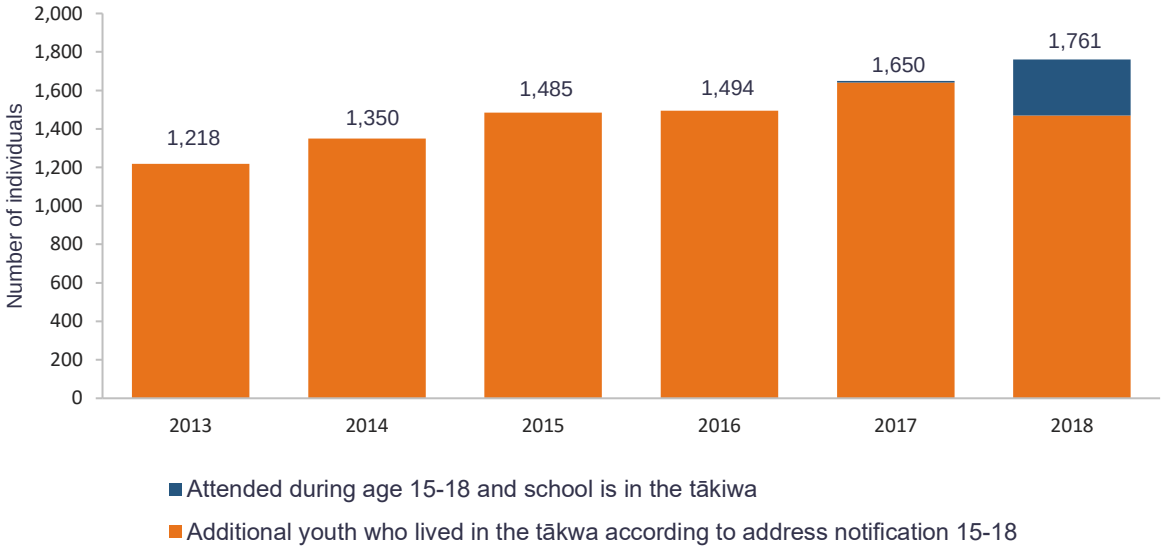


Figure 4. Location source of Māori youth who turned 30 in reference years 2013-2018



Measures of interest

The Agency and Tokona Te Raki agreed on a list of variables of interest during an iterative process, where data availability and reliability were considered. We decided to go broad and focus on a wide range of education and employment outcomes.

We have obtained more than 200 measures from numerous IDI sources, grouped into 34 broader topics. Some of these were self-explanatory, e.g. NCEA Level 1 Science achievement, while some took considerable effort, e.g. mental health service use. The measures are described in the Appendix. The definitions can be accessed via our GitHub page.

For repeated use, the measures can be extended or reduced to fit the purpose of future research. The results of algorithmic clustering changes depending on the measures included in the clustering algorithm. When new measures are added, we recommend checking whether the measures are available for the whole population otherwise algorithmic clustering will be affected by inconsistent availability of data for different cohorts. The availability of MoE primary and secondary school data was one of the reasons for dividing the study population into the two subgroups.

Visualising results

The Agency has shared timeline visualisation codes,[8] that help to understand the complexity of life for Māori youth in the Ngāi Tahu takiwā during the critical life stage of leaving secondary education.

Showing a story

Manual and algorithmic clusters can be selected. The next step is identifying the journey events of interest: the application displays only the groups and event measures selected.

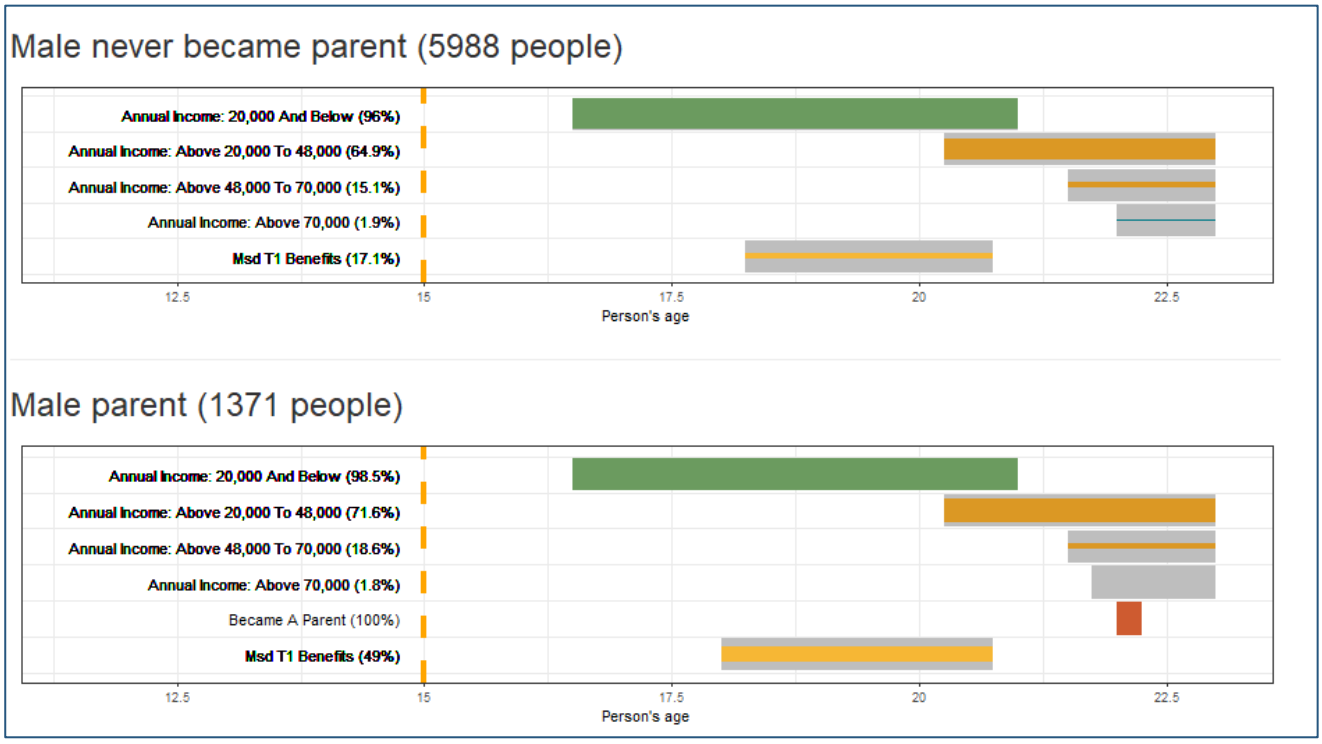
The screenshot below shows two subgroups to understand how earnings and benefit receipt might be influenced by parenthood: males who are not parents before the age of 23 and males who are. We selected six event measures: individual income categories, Ministry of Social Development tier 1 benefit receipt and becoming a parent.

The screenshot displayed in Figure 5 is interpreted as below:

- There are roughly five times more men in the group who did not have children by 23 (5,988) than those who had (1,371).
- The most common age to become a parent for the parent group is just after the age of 22 (22.25) and, on average, males in this group have one child.
- The parent group has worse unemployment outcomes:
    - A higher percentage of the parent group received tier one benefit (49% compared with 17.1%), the period of benefit receipt typically occurs before the child was born.
- Some of the disadvantages of the male parent group seem to decrease around the time of becoming parents:
    - A larger percentage of them enter higher earning income groups ($20-$48,000 and $48-$70,000).

[8] Available here: https://github.com/nz-social-wellbeing-agency/timeline_visualisation

Figure 5. Screenshot from the R shiny visualisation application



Interpreting results

While the visualisation application is easy to use, the explanation of timelines can be difficult for non-technical audiences. Caveats for interpretation are provided in SWA, 2019 (pp 19-22). Causal interpretation is not applicable, however interactions between pairs of events can be applied, e.g. “of people who have experience X, Q% of them had experience Y beforehand, and went on to have experience Z after.” (SWA, 2019, p 10). Contextual expertise is necessary to distil the meaningful information provided by this analysis. The Agency and Tokona Te Raki used participatory data analysis techniques to facilitate the knowledge transfer between the qualitative and quantitative teams.

The main quantitative findings from the overall research will form part of the comprehensive report published by Tokona Te Raki.

References

Abbott, A., & Hrycak, A. (1990). Measuring Resemblance in Sequence Data: An Optimal Matching Analysis of Musicians' Careers. *American Journal of Sociology*, 96(1), 144-185.

Abbott, Andrew & Tsay, Angela. (2000). Sequence Analysis and Optimal Matching Methods in Sociology: Review and Prospect. *Sociological Methods & Research*. 29. 3-33. 10.1177/0049124100029001001.

Biemann, Torsten & Zacher, Hannes & Feldman, Daniel. (2012). Career patterns: A twenty-year panel study. *Journal of Vocational Behavior*. 81. 159–170. 10.1016/j.jvb.2012.06.003.

Blair-Loy, Mary. (1999). Career Patterns of Executive Women in Finance: An Optimal Matching Analysis. *American Journal of Sociology*. 104. 1346-1345. 10.1086/210177.

Fuller, Sylvia & Stecy-Hildebrandt, Natasha. (2014). Career Pathways For Temporary Workers: Exploring Heterogeneous Mobility Dynamics With Sequence Analysis. *Social Science Research*. 50. 10.1016/j.ssresearch.2014.11.003.

Halpin, Brendan. (2010). Optimal Matching Analysis and Life-Course Data: The Importance of Duration. *Sociological Methods & Research*. 38. 365-388. 10.1177/0049124110363590.

Halpin, Brendan & T, Wing. (2004). Class Careers as Sequences: An Optimal Matching Analysis of Work-Life Histories. *Institute for Social and Economic Research, ISER working papers*. 14. 10.1093/oxfordjournals.esr.a018230.

Smith, L.T. (1999). Decolonizing Methodologies: Research and Indigenous Peoples. Dunedin: *Otago University Press*; ZedBooks Ltd.

Social Investment Agency.[9] (2019). *Representative timelines – modelling people's life experiences*. Wellington, New Zealand.

Stats NZ. (2020). *Microdata output guide (Fifth edition)*. Retrieved from www.stats.govt.nz.

Struffolino, Emanuela. (2019). Navigating the early career: The social stratification of young workers' employment trajectories in Italy. *Research in Social Stratification and Mobility*. 63. 100421. 10.1016/j.rssm.2019.100421.

[9] Renamed Social Wellbeing Agency in 2020.

Appendix – List and description of the variables and population segments

Table 2. Timeline events observable in the visualisation tool

Variable name	Description
ACC claim	The ACC data within IDI includes all claims made due to work-related or non-work-related injury and whether the injury occurred in New Zealand or not.

Achievement at various levels and subjects in high school. Either "taken" or if achievement is indicated: "At most achieved" or "Achieved with merit or excellence"	Achievement levels have the following dimensions:
	• The levels indicate levels of NCEA certificate. In general, students work through levels 1 to 3 in years 11 to 13 at school.
	• Subjects of interest have been pre-selected based on discussion with Tokona Te Raki.
	• Achievement can be
	  ◦ Simply taken, based on participation or
	  ◦ One of the following: fail, achieved, achieved with merit, achieved with excellence.

Address to or from	Moving in and out of the takiwā: defined as all South Island, except the regions of Nelson-Marlborough and Tasman.
Admitted to hospital	Every publicly funded hospital stay in New Zealand longer than 3 hours is required to be collected by DHBs to form a national dataset. This includes people treated in ED for more than 3 hours and/or admitted as inpatients. Both publicly and privately funded hospital events can be found in IDI.
Annual income	Annual income bands are determined as: below 20,000 NZD, 20-48,000 NZD, 48-70,000 NZD and above 70,000 NZD. Income includes income from different sources as total from IR which includes wages and salaries, withholding payment, and payments such as benefit payments, ACC payments, pension payments, paid parental leave, student allowance, sole trader income, rent income etc.
Applied for social housing, Housing New Zealand (HNZ) tenancy	Application for social housing requires a process including a general discussion, interviews, and if eligible, waiting for a property and accepting the public housing offer. As of early 2020 the wait list for public housing hit 14,500 households.
Attending specific years in high school	Attending year 9-13 in high school. This is usually denoted by “Attending: year x” which pertains to the spell or period when an individual attends the specific year level in high school.
Attending tertiary education, industry training, post-secondary no-tertiary training, targeted training etc.	Pertains to periods when an individual attends levels of post-secondary, industry training or tertiary education. This is based on the MOE levels of classification of levels in post high school attendance, e.g. post-secondary, 1st or 2nd stage tertiary and types of industry training.

Variable name	Description
Became a parent	Becoming a biological parent, according to Department of Internal Affairs (DIA) records. Only live births.
Chronic health conditions	The existence of one or more of the chronic conditions are recorded by Ministry of Health datasets: acute myocardial infraction, cancer, diabetes, gout, stroke, traumatic brain injury.
Combined annual income with partner, if applicable	Combined annual income bands are: below 40,000 NZD, 40-80,000 NZD, 80-100,000 NZD and above 100,000 NZD. We combined income from all sources to calculate the annual income of individuals with income defined as in the total annual income measure. We calculated income for both the individual and his or her spouse or de facto partner. For simplicity we banded the combined income measure, making sure that we have enough individuals in each band.
Community visit	We use community visits by hospital staff as recorded in out-patients. These visits may have only been consistently recorded recently. Events where the person Did Not Attend (DNA) are excluded.
Corrections experience	We identify periods and events of management within the justice system. This measure does not differentiate between the types of Corrections episodes, e.g. community sentence or detained sentence.
Court hearings and proceedings	The first and last hearings relating to a specific charge are recorded. If a charge has three or more hearings than the hearings other than the first and last will not appear in the dataset.
Employment related data	Employed: employed based on entry in internal revenue data, regardless if source is from wages and salaries or withholding income. Employed (W&S): employed based on receiving wages and salaries Employed (WHP): employed based on receiving withholding income
Exposed to family violence	There are multiple sources of family violence data. Family violence data has been obtained in IDI from the following sources: • NZ Police (victimisations from July 2014 onwards, proceedings from 2009 onwards) • Ministry of Justice (charges from 1992 onwards) • ACC (sensitive claims and injury claims based on keywords) SME from Ministry of Justice advised our team on how to determine offender codes relating to family violence. Oranga Tamariki (e.g. reports of concern) data was excluded. Exposure to family violence contains both perpetrators and victims. If required this information can be obtained. These codes may be revised in the future based on external review. This data has limitations. The New Zealand Crime and Victims Survey (NZCVS) shows that a large amount of family and sexual violence cases are not reported

Variable name	Description
	to Police.[10]
Health events	PHO (primary health organisations) enrolment and active use to ensure the provision of essential primary health care services, mostly through general practices. Publicly funded hospital events include inpatient and outpatient visits un public hospitals, and includes X-rays and laboratory tests carried out in public hospitals.
High deprivation	The NZDep is an area-based measure of socioeconomic deprivation in New Zealand. It measures the level of deprivation for people in each small area.[11] We used 2013 NZDep values, and selected the decile 9 and 10, most deprived scores. The indicators used for each NZDep may change from Census to Census.
High school decile	The decile of the provider/school as specified by MOE data. Decile 1 meaning that the school is zoned to the most deprived catchment area, decile 10 meaning that the school is zoned to the least deprived catchment area.
High school location (ever attended)	The region where the school is located. Note that a person who transfers schools can appear in different – multiple - groups.
Highest qualification level achieved	Where only year is available assumed qualification awarded 1st December (approximately end of calendar year). The codes identifying these events were guided by Population Explorer Highest Qualification code in SNZ Population Explorer by Peter Elis.[12] • 1 = Certificate or NCEA level 1 • 2 = Certificate or NCEA level 2 • 3 = Certificate or NCEA level 3 • 4 = Certificate level 4 • 5 = Certificate of diploma level 5 • 6 = Certificate or diploma level 6 • 7 = Bachelor's degree, graduate diploma, or certificate level 7 • 8 = Bachelors honours degree or postgraduate diploma or certificate level 8

[10] The second annual New Zealand Crime and Victims Survey (2019 NZCVS) is based on interviews with 8,000, people, reveals 94 per cent of sexual assaults were not reported to Police.

[11] The nine census variables can be found here: https://www.ehinz.ac.nz/indicators/population-vulnerability/socioeconomic-deprivation-profile/. Examples include people living in equivalised households with income below an income threshold, people aged 18-64 years receiving a means-tested benefit, people not living in their own home, people aged 18-64 years without any qualifications.

[12] https://github.com/StatisticsNZ/population-explorer/blob/master/build-db/01-int-tables/18-qualificiations.sql

Variable name	Description
	• 9 = Master's degree • 10 = Doctoral degree
Industry training by field	Training by industry field.
Industry: being employed by an employer within a specific industry	The industry of the employer for those individuals in the population who received wages and salaries. These categories are not mutually exclusive, an individual can work at any industry during the window analysis (15-22 and 15-29 years).
Intervention types at high school	Student interventions, as defined by Ministry of Education. We have focussed on stand downs and suspensions, based on conversations with Tokona Te Raki. A state or state integrated school principal may consider the formal removal of a student through a **stand-down** from school for a period of up to 5 school days. A stand-down, for any student, can total no more than 5 school days in a term, or a total of 10 days in a school year. Students return automatically to school following a stand-down. **A suspension** is a formal removal of a student from a school until a school Board of Trustees decides the outcome at a suspension meeting. Following a suspension, the Board of Trustees decides how to address the student's misbehaviour. The Board can either lift the suspension (with or without conditions), extend the suspension (with conditions), or terminate the student's enrolment at the school.
Kura Kaupapa Māori / Māori boarding school attendance, co-ed school	Secondary school types include e.g. co-educated secondary schools, Kura Kaupapa Māori – if attendance happened at least once during high school years. Therefore if someone moved high schools, one can belong to both groups: attended Kura Kaupapa. There is concern that this dataset in IDI may not be complete.
Major/large/medium urban area, rural area	Having address notifications relating to an address in different types of settlements: • Major urban areas are very large urban areas centred on a city or major urban centre, with a minimum population of 100,000 residents. • Large urban area: these urban areas are large urban areas centred on a city or major urban centre, with a population of 30,000–99,000 residents. • Medium urban areas are centred on the larger regional centres. They have a population between 10,000 and 29,999 residents • Small urban areas are urbanised settlements (outside major, large, and medium urban areas), centred around smaller towns with a population between 1,000 and 9,999 people. • Rural: those areas not specifically designated as 'urban'. They include towns of fewer than 1,000 population plus administrative district territory where this is not included in an urban area. Rural areas include offshore islands.
Medium in Te Reo	Indication of a student taking a standard that is taught in Te Reo. This is not

Variable name	Description
	mandatory information, hence there can be some unclassified standards.
Mental health service use, e.g. for anxiety, mood disorders or substance use	Based on the standardised definition of mental health and addictions (MHA) service access based on available data in the IDI (available on GitHub): • Prescription of pharmaceuticals deemed to be mental health conditions and addiction related (excluding potential mental health conditions and addiction category). • Any hospitalisation event with an associated mental health conditions and addiction diagnosis. • Any record with a mental health conditions and addiction related medical certificate for benefit support.
MSD t1 and t2 benefits	Tier 1 benefit is the main benefit, including Jobseeker Support (JS), Sole Parent Support (SPS), Supported Living Payment (SLP), and other main benefits. Tier 2 benefit includes accommodation supplement, child disability allowance and disability allowance.
NEET	This measure is calculated (as observable in the IDI) among individuals who were not studying, employed or overseas as a composite of (1) not having an overseas spell, (2) not found in employment according to IRD information and (3) not in study or training. Constructed from IRD, MOE and border movements data. We estimate the number of days employed, studying, and not studying, employed or overseas (these categories are not mutually exclusive). A person is assigned a main activity of NEET if they were not studying, employed or overseas for at least 180 days during 2018 and were not studying, employed or overseas for more days than they were employed and more days than they were studying.
Parents qualifications and employment status	Employment and (highest) qualification of parents. Parental links are primarily DIA births-based.
Person overseas	Spells when a person has been recorded to be overseas.
Post-HS study	Post-high school study subjects, tertiary or industry training.
Post-secondary, non-tertiary	Studying subjects after high school which are not intended for tertiary level.
Program with alcohol and drug team	Gives time periods (often single days) that an individual was recorded as at a drug and alcohol program. Reports are provided by DHBs and NGOs on services rendered (hence we assume non-attendance is not captured). Reporting by NGOs has been progressive from 2008 to 2014. Hence counts over time will vary with increased reporting/connection of NGO reports to PRIMHD.
Program with maternal mental health team	Gives time periods (often single days) that an individual was recorded as at a maternity mental health program. Reports are provided by DHBs and NGOs on

Variable name	Description
	services rendered (hence we assume non-attendance is not captured). Reporting by NGOs has been progressive from 2008 to 2014. Hence counts over time will vary with increased reporting/connection of NGO reports to PRIMHD.
Qualification levels awarded	NZQF level of qualification and indicating when these qualifications were awarded: • 00 No Qualification • 01 Level 1 Certificate • 02 Level 2 Certificate • 03 Level 3 Certificate • 04 Level 4 Certificate • 05 Level 5 Diploma • 06 Level 6 Diploma • 07 Bachelor's degree / Level 7 qualification • 08 Post-graduate and honours degree • 09 Master's degree • 10 Doctorate Degree
Student loan debt	The duration while one has student loan debt exceeding $40.
Taking standard levels 1-6	Students, taking any subjects belonging to standard levels 1-6.
Targeted training enrolments	The individuals here are "people placed in targeted training programmes funded by government through tertiary education commission. These training may include gateway, skill enhancement, training opportunities, foundation focused training opportunities and youth training."
Tertiary (1st and 2nd stage)	Attending tertiary education. First stage pertains to non-research tertiary education that can lead to either academically or vocationally based education. Second stage of tertiary can lead to advanced research.
UE Eligible: Yes	This measure indicates whether New Zealand university eligibility requirements have been met. The requirements are include completing NCEA Level 3 exams, completing numeracy and literacy credits and additional credits of approved subjects. The list of required and approved subjects may vary year-to-year.[13]
Unit levels taken/at most achieved	Unit standards taken by the student. Each standard describes what a student needs to know, or what they must be able to achieve, in order to meet the

[13] For more information, visit New Zealand Qualifications Authority's (NZQA) website: https://www.nzqa.govt.nz/qualifications-standards/awards/university-entrance/

Variable name	Description
	standard. One category of standards is unit standards that are competency based.[14]
Upper secondary	These are levels of education designed to lead to either higher academic or vocational education or directly to labour market.
Vehicle registration, driver licensing	The transport data in IDI includes information on driver licence holders, motor vehicles, and motor vehicle ownership. We have not distinguished in this study between different types of vehicles. The dataset only contains receiving the full driving license.

Table 3. Comparable population segments (manual segments) in the visualisation tool

Segments	Description
ACC claim	Individuals who have claimed any benefit from ACC, due to work-related or non-work-related injury and whether the injury occurred in New Zealand or not during the study period.
All groups	Contains all the individuals in the population.
Annual income and combined annual income	Individuals were assigned to six groups based on their average annual income in the last three years of the study period. The income bands for these groups were the same bands used in the variable definition: • Annual income bands are determined as bands: below 20,000 NZD, 20-48,000 NZD, 48-70,000 NZD and above 70,000 NZD. • If combined annual income with partner is applicable, the bands are: below 40,000 NZD, 40-80,000 NZD, 80-100,000 NZD and above 100,000 NZD.
Attended tertiary education after age 20	All individuals who enrolled in tertiary after age 20. Sometimes referred to as second chance learners because university eligibility requirements no longer apply. New Zealand or Australian citizens (or permanent residents) who do not have a university entrance qualification and are above the age of 20 can apply for special admission.[15]
Attending achievement level	Individuals by type of achievement level taken as well as grade obtained, i.e. at most achieved or with merit /excellence. Achievement levels correspond to NCEA levels 1, 2 and 3.
Attending unit level	Individuals by type of unit level standards taken. These standards are available for a number of learning areas, but not all of them contribute to university entrance qualifications.

[14] https://www.nzqa.govt.nz/ncea/understanding-ncea/how-ncea-works/standards/

[15] Universities New Zealand, Discretionary Entrance Regulations, Guidelines and Procedure, May 2020.

Segments	Description
Becoming parents (male/female parents vs non-parents)	Individuals who became and did not become parents and their sex (male or female), based on DIA births during the timeline. These categories help the viewer compare journeys of males and females who did or did not become parents by their 23rd birthday.
Child youth and family events	SEC15, abuse and placement events. SEC15 refers on report of concern to CYF that meets Sec 15 criterial (ill-treatment or neglect of child or young person).
Chronic health conditions	The population with one or more of the chronic conditions diagnosed by their 23rd/30th birthdays as recorded by Ministry of Health datasets: acute myocardial infraction, cancer, diabetes, gout, stroke, traumatic brain injury. Not all the conditions in this dataset in the IDI are being kept up to date.
Ethnicity	This groups are not mutually exclusive groups, i.e. individuals can have multiple ethnicities. All study participants are of Māori ethnicity.
Geographic area	Individuals who reside in urban and rural areas by size. As these are journeys, some people may appear in two or more geographic areas across the timeline.
Health attributes	Admitted to hospital, MHA (mental health and addiction) service use, attended program with alcohol and drug team. Mental health service use is defined in the variables list.
Highest qualification	Individuals who gained specific highest level of qualifications based on the New Zealand Qualification Authority classification : • Level 0 – no qualifications • Level 1-3 – NCEA Level 1-3 • Level 4-6 – Certificate or diploma level 4-6 • Level 7+ – Bachelor's degree and above
Industry of employment	Individuals who work in major sectors or industries. Similarly, some workers may appear in different sectors. The primary aim of this grouping is to see individuals' journey characteristics given their employment in an industry.
Māori schooling attributes	Grouping by enrolment in Kura Kaupapa Māori/ Māori Boarding School as well as attending classes that are taught in Te Reo.
Parent and older sibling highest qualification	Individuals' birth parents and biological siblings and half siblings in the study population who gained specific qualifications based on generic groups of highest qualification as described in the variable list.

Segments	Description
Post HS attendance	Includes individuals who attended post high school subjects by major field.
Sex	Female or male.
Student intervention	Groups of individuals who received a specific education intervention. There is also distinction among those who were never stood-up/suspended.
Synthetic clusters	Machine defined clusters computed after sequence analysis using optimal matching.[16] These are groups whose members are expected to have similar journeys in terms of the measures defined.

[16] Using TraMiner package in R: https://cran.r-project.org/web/packages/TraMineR/index.html