IDI research for Workforce Development Councils

Data dictionary and guide to queries

September 2022

SCARLATTI
Evaluation | Analytics | Insights

Contents

Introduction

Overview

This document provides a technical overview of research work conducted for the six workforce development councils in 2022. This includes the methodologies used to complete each section of research as well as any considerations taken into account.

Research objectives

The wider research project aims to identify the specific transferable skills that enabled the service sector workforce to switch careers, before and during the Covid-19 Pandemic and will discuss further opportunities to strengthen the resilience of the service sector workforce. And the key objectives of the quantitative piece would be:

1. Provide quantitative estimates regarding the **mobility of the workforce** in the services sector, pre and during the Covid-19 pandemic – including estimates of the flow of workers/employers into and out of the service industries.
2. Understand **Māori employment in the service sector** – particularly the impacts of the Covid-19 pandemic on Māori workers and businesses in the service industries.
3. Identify the **skill levels** of those in the service industries, and whether there is any association between skill level and mobility.
4. **Provide recommendations** (and contribute to discussions) around how to further lift workforce resilience for the services industries from a skills (and quantitative research) perspective.

Analyses in this work

The analyses to address these research objectives include:

1. Workforce composition
2. Māori business composition
3. New entrant volumes and origins
4. Tenure and retention
5. Mobility and destinations
6. Skills
7. Resilience
8. Workforce specific queries

Data tools

Two key data tools have been used in this work[1]:

[1] For more information, please see the Stats NZ websites https://www.stats.govt.nz/integrated-data/integrated-data-infrastructure/ and https://www.stats.govt.nz/integrated-data/longitudinal-business-database/.

1. **The Integrated Data Infrastructure (IDI)**

   The most appropriate data tool to address these research objectives is the Statistics NZ Integrated Data Infrastructure (IDI): a large world-leading research database containing microdata about people and households, specifically detailing life events such as education, income, and employment.

2. **The Longitudinal Business Database (LBD)**
   A tool closely related to the Integrated Data Infrastructure is the Longitudinal Business Database (LBD): an additional large research database containing de-identified microdata relating to businesses. The LBD is linked to the IDI through tax data.

Please see Appendix A for a discussion of the limitations and caveats of these data sources.

1. Workforce composition

What do we mean by workforce composition?

Workforce composition describes the population of people working in a workforce at a given point in time, including the demographic information of various groups..

Why is this of interest to WDCs?

Describing the workforce is foundational for further investigation into workforce mobility and resilience.

What types of questions are we interested in answering?

- What is the size of the workforce at any given point in time?
- How does the size of the workforce change over time?
- What is the demographic breakdown of the workforce at a point in time and how does it vary over time?

We will create an analysis with the following breakdowns:

- Age
- Gender
- Ethnicity
- Visa status
- Regional council
- Disability status
- Monthly income level
- Industry tenure
- New entrants vs existing workforce
- Training experience

What issues / points are worth consideration?

Populations can be defined by industry or occupation.

Workforce populations can be defined by the industry that people work in and / or their occupation. Both views of a population are useful and worth consideration.

Currently WDC workforces are defined using a list of relevant industries, which are identified using *Industry* codes. However, most vocational training is designed with specific occupations in mind, which are identified by *Occupation* codes within the IDI and LBD.

There is a 'many-to-many' relationship between industry codes and occupation codes. That is, one industry can include many occupations. For example, design firms employ bookkeepers as well as artists/designers. One occupation can also be present across many industries, for example, a designer could work for organisations such as consultancies, or manufacturers, as well as design businesses.

Neither industry nor occupation codes are strongly aligned with the industry and occupation groups that WDCs work with

Neither set of codes do a perfect job of defining the industries and occupations of interest to WDCs, with the degree of imperfection varying with different WDCs. Various WDCs are currently still in the process of working through the industry and occupation codes that apply to them.

One of the outputs of this work is a visualisation of a preliminary mapping of Industry to Occupation workforce populations created using 2018 Census data.

Privacy and rounding constrain how many ways workforces can be 'sliced and diced'

Statistics NZ's rounding rules mean that data sets about the workforce populations can typically only by characterised using 2-3 attributes of interest at one time. Therefore, multiple queries and datasets are needed to drill down on different attributes of the workforce.

Data sources

Within the IDI, we can identify a workforce using two different data sources:

1. Census data
   Within the Census data tables, codes associated with occupations (ANZSCO) and industries (ANZSIC) are available. For a given list of agreed occupations and industries, a workforce may be defined simply and by using both occupation and industry codes ensures that any individuals associated with the industry code, but in an unrelated occupation, remain excluded.

2. IRD tax return data
   Within the IRD data tables, both income and industry codes (ANZSIC) are available on a monthly basis and the size of the workforce may be defined for a given ANZSIC code. This immediately provides a time advantage over the Census data with the potential for a longitudinal analysis, with a monthly time step, for anyone earning taxable income within a given industry. If desired, an income threshold can also be set to exclude any individuals from the workforce i.e., those workers earning less than half-time minimum wage per month. This approach will capture anyone entering or leaving a workforce over a given time period of interest.

Each source has its pros and cons (see below), but together, can be used to form a complete picture of the workforce size and nature over time. For further information on the differences between the two approaches described above, see our webpage here https://sweetanalytics.co.nz/content/defining-workforce/.

Query definitions

A. Occupation-defined workforce

The intent of this query is to give a snapshot of the workforce in agreed occupations by using 2018 Census data and the comparisons in different attributes breakdowns (listed below).

Attributes

Demographic	Definitions
Work type	Signify if the employee is full time, part time or not in labour force
Age	Five-year age band group
Gender	Gender type
Region	Council of location that respondents are in
Ethnicity	One column for each ethnicity, so people who have multiple ethnicities could all be recorded.
Hours	Hours the person worked per week. Five categories: 1. 0-10 hours 2. 11-20 hours 3. 21-30 hours 4. 31-40 hours 5. Over 40 hours
Employment type	Signify if the group are employees, employers or self-employed
Qualification	Signify respondents' highest qualifications
Income level	Yearly income band

B. Industry-defined workforce

Identifying workforces of relevant industries from monthly IRD tables provides a detailed picture of variations in the workforce each month, and a more up-to-date time series than Census data.

Total workforce (including seasonal / part-time)

Cohort	Data sources	Definitions
Employees	IRD, Business Register	• Had a tax return filed for Wages & Salaries received for at least **1 month in the given calendar year** • For each of those months, earning **any income** (i.e., no income threshold)
Self-employed / employers	IRD, Business Register	• Had a specific income source code • No income threshold

Demographics

Demographic	Data sources	Definitions
Age	IRD	From 15 to 65+, ten-year age bands
Work type	IRD	Way to distinguish between casual worker, part-timer and full-timers by comparing total monthly income with 100 hours x minimum wage and 150 hours x minimum wage
Gender	IRD	Gender type
Ethnicity	IRD	Ethnicity type. People with multiple ethnicity type are combined into one by using the order of Māori, Pacific, Asian, MELAA, European, Other (Further improvements could be added to allow recording all ethnicities for people with more than one ethnicity).
Disability	Census 2018	Three categories, 'Disabled', 'Not disabled' and 'Not stated'
Region	IRD	Location of the workforce group (Regional council level)

Visa status	DOL decision table	Type of visa the group was holding in each month
Training status	MOE enrolment table, TEC-it learner table	Regard tertiary training as pre-employment training from MOE enrolment table and TEC training as on-job training from TEC it-learner table. Signify person's training status (Had trained before/In training/Never trained) in each tax return month.
Industry tenure	IRD	Total length of time people have had spent in an industry till the given month

What approaches did we take in the IDI?

For an occupation-defined workforce, we use agreed occupation codes (ANZSCO) to retrieve data from the 2018 Census table and decode useful variables by matching with metadata tables.

For an industry-defined workforce, we use the IRD tax return table to retrieve a monthly in agreed industries and time periods. By using unique snz_uids to match the DOL_desicion table, Census 2018 table and so on, demographic information could be added into monthly workforce composition tables to give a complete picture of the workforce characteristics.

How did we structure the query?

For an occupation-defined workforce:

1. Retrieve the workforce from 2018 Census table by using occupation codes (ANZSCO)
2. Decode variables of interest (such as age, gender, region, work type and so on) by matching with metadata tables

For an industry-defined monthly workforce, we use the latest version of IDI data (202206) and do not apply a minimum income threshold, we only look at people with income source code 'W&S' (Wage and Salary) in the IRD table. The logic of the query is as followed:

1. Select target population:

   In the IRD tax return table, using industry codes (ANZSIC) to select the monthly workforce at each month from 2015 to 2022 and calculate the individual's monthly income (for a person that has more than one job at the same time, sum all incomes from all jobs in a month as their total monthly income)

2. Add a 'work type' column:

   In order to help signify the type of employees (casual, part-time, full-time), we compare monthly income with 100 hours x minimum wage and 150 hours x minimum wage. There are 3 categories currently: '<100 hours','100-150 hours', 'More than 150 hours'.

3. Add demographics columns:

   Match the monthly workforce composition table with metadata tables to decode demographics columns (Age, region, ethnicity, gender)

4. Add training breakdown:

   In order to get the training status for individuals each month, we use the *MOE enrolment table* (to signify pre-employment training) and *TEC learner table* (to signify on-job training) to match with our monthly workforce composition table. There is total of 16 types of training flags to capture all possible scenarios and signify all the types of trainings. The detailed definitions of main training statuses are described as followed:

- If a tax return month falls between the start and end date of pre-employment training and on-job training, we regard the training status of that person in that month as 'In training (pre-employment training)' or 'In training (on-job training)' or 'In training (pre-employment training, on-job training)' depending on which training they are enrolled in
- If a tax return month falls out of some of the end dates of the training programmes that the person is enrolled in, then we summarize all the training he/she had completed and group into three categories ('Had trained before (pre-employment training)','Had trained before (on-job training)','Had trained before (pre-employment training, on-job training)'
- If there is no record for that given individual or the tax return month is before the start date of the training programmes that the person enrolled, we regard them as 'never trained'
- Given the complexity of training in real-life, a person's monthly training status could also be combinations of the types discussed above which are all recorded into our 16 unique types of training status.

5. Add industry tenure column:

   Match with IRD tables to sum up all the months so far that these people have worked in current industry. For example, Person A was working in the Cafés industry in Jan 2015, then this step calculates the total number of months that person A has worked in the Cafés industry up to Jan 2015. The industry tenure is grouped into 4 categories: 'Less than 1 year','1-3 Year', '3-5 Year' and '5 years and over'

6. Add visa type column:

   In DOL decision table, we select all the visa information for each person in the monthly workforce table. We compare the visa decision date and expiry date with each tax return month to determine the current visa type of each individual held in that month as their visa status. We regard the workforce that don't have visa information in DOL_decision table as New Zealand citizens (NZ citizens).

7. Add disability column:

   Match the monthly workforce table with the 2018 Census table to determine the disability status of each individual. This might not be accurate currently, as the 2018 census data is out of date but it is the only available dataset that has disability information.

8. Specify new entrant and existing worker:

   If an individual has an industry tenure of 1 month, then they are tagged as a new entrant, as they have entered the workforce that month. Everyone else is classed as an existing worker.

2. Māori business composition

What do we mean by Māori business composition?

By Māori business composition, we mean the proportion of businesses in the sector which are involved with the Māori community in one or more of the following ways: a Māori-owned business, a Māori sole trader and / or a significant employer of Māori.

Why is this of interest to WDCs?

This analysis provides insights into the needs of Māori within the sector, and context around the development of Māori businesses within the sector in terms of both ownership and employment.

What types of questions are we interested in answering?

- How many Māori-owned businesses or Māori sole-traders are there in these industries?
- How many businesses in the sector are significant employers of Māori?
- How have these businesses changed over time (pre- and post- covid)?

We will create an analysis with the following breakdowns:

- Business type (self-employed vs employer)
- Business size (number of employees)
- Business subsector
- Business profitability

What issues / points are worth consideration?

The thresholds for Māori-owned businesses and significant employers of Māori are somewhat arbitrary. The definition we have used is outlined below, but it is worth noting that we could adjust the proportion of active shareholders, directors or partners, or the proportion of Māori employees that we consider indicates a Māori-owned business or significant employer of Māori.

There are limitations on identifying business ownership within the IDI, for many businesses the IDI simply does not contain the information needed to identify the ownership of the business. In these cases, it is not possible to ascertain if the business is Māori-owned under our current criteria.

What approaches did we take in the IDI?

As part of this research, we spent some time working with Nicholson Consulting, who have previously completed some IDI research within the scope of identifying Māori businesses and significant employers of Māori as part of the Te Matapaeroa 2020 report, an analysis of Māori business across New Zealand commissioned by Te Puni Kokiri, the Ministry of Māori Development. We used the same logic to approach this analysis as Nicholson used. The first step of which is to understand what classification to use to define a business, and subsequently a Māori-owned business or significant employer of Māori.

Stats NZ define a business as a significant economic enterprise. That is, it has an enterprise number in the LBD and an IRD number and meets at least one of the criteria listed in Table 1.

Table 1: Stats NZ criteria for a business to be classified as an economically significant enterprise

An economically significant enterprise meets one or more of the following criteria:
Greater than $30,000 annual GST expenses or sales
More than three paid employees
In a GST exempt industry, other than residential property leasing and rental
Part of a Business Register (BR) group
Has a new GST registration and has registered for salaries and wages PAYE but has not yet started filing GST returns or has a new GST registration and part of an IRD GST group return. The former has a12-month window of being considered an enterprise before the other criteria in this table are applied
Has a live GEO classified to agriculture
IRD10 income is greater than $40,000

1. **Māori-owned businesses**
   A Māori-owned business is defined as a business which meets one or more of the following criteria[2]:
   - A business that pays at least 50% of wages[3] to active shareholders, directors or partners of Māori ethnicity or descent.
   - A business identified as Māori by Stats NZ.
   - A sole trader identified as Māori by IDI ethnicity data or NZ Census 2013 data, earning a non-zero income.
2. **Significant employers of Māori**
   A significant employer of Māori is defined as a business whose employees of Māori ethnicity or descent (as identified by IDI data or the NZ Census 2018) make up an arbitrary proportion of the business' employees. Te Puni Kōkiri, the Ministry of Māori Development have used a 75% threshold in their research4, which we will also use in this project. This definition excludes any business identified as sole traders.

How did we structure the query?

1. Define the business population within New Zealand

[2] This approach is in alignment with the Te Matapaeroa 2020 report, an analysis of Māori business across New Zealand commissioned by Te Puni Kokiri, the Ministry of Māori Development.
[3] This threshold can be changed depending on the research being undertaken, this threshold is the one set by Te Puni Kokiri, The Ministry of Māori Development in the Te Matapaeroa 2020 report.
[4] This threshold is used in the Te Matapaeroa 2020 report commissioned by Te Puni Kokiri, the Ministry of Māori Development.

- These are all enterprises as defined by Stats NZ according to Table 1 that are active in each month / year

2. Find the average number of employees per year for each of these businesses, as well as the average number of employees identifying as Māori per year
   - Calculate the income of the above businesses by calculating the total sales and purchase expenses at the enterprise level
3. Define ownership of these businesses based on any Director / Partnership / Shareholding incomes paid out by the business
4. Find the sole trader population, that is all enterprises with only one employee
   - Calculate the yearly income of sole traders
   - Identify ethnicity of sole traders
5. Calculate the proportion of Māori employees for each business and ownership payments made to people identifying as Māori
6. Identify relevant businesses using the following criteria:
   - Māori business: a business making more than 50% of its director / partnership / shareholding payments to people identifying as Māori, or identified as a Māori business by Stats NZ
   - Significant employer of Māori: a business with more than 75% of its employees identifying as Māori
   - A Māori sole trader: a sole trader identifying as Māori

3. New entrant volumes and origins

What do we mean by origins?

By the origins of a population, we mean where workers come from before they enter the workforce or start training. Specifically, we look at what cohorts of people were doing in the year before they joined the workforce or learner population.

Why is this of interest to WDCs?

This analysis provides context around the experience people bring and the backgrounds they have prior to joining the workforce or starting training. This is relevant to attracting new talent into the sector in the future. By understanding where people have come from historically, we can predict where they will come from in the future, and identify any talent pools that could be targeted to increase the number of workers from that source.

What types of questions are we interested in answering?

- How many new entrants have joined the workforce over time?
- What talent pool did any given person belong to at a specific time period before joining the workforce or starting training? The talent pools we consider are:
  - Career changers, i.e., those who were previously working in a different sector
  - Secondary school leavers
  - Tertiary education leavers
  - Migrants
  - Returning Kiwis
  - Beneficiaries/NEETs (Not Employment, Education or Training)
  - Other not elsewhere classified.

We are interested in the attributes of each talent pool at any given time. This includes:

- A breakdown of the highest qualification of members
- An age profile
- An ethnicity profile
- A gender profile

We also want to explore the career changers group in particular:

- Are they from relevant industries or non-relevant industries?
- Which industries are they coming into?

What issues / points are worth consideration?

In order to breakdown the origins in the manner outlined here we need to define a hierarchy of mutually exclusive talent pools. We also need to decide how far back we are interested in analysing. We would also like to explore if the breakdown of talent pools that new entrants come from changes if they are not treated as mutually exclusive.

What approaches did we take in the IDI?

We take cohorts of interest at each time step and determine their primary activity in each year prior to joining the workforce or starting training. This is done by analysing specific criteria for each activity, for example, if they were enrolled in secondary school, or receiving a benefit for the majority of the year etc.

How did we structure the query?

1. Define thresholds for workforce population, currently sit at:
   - Earning at least minimum wage for:
     i. Less than 100 hours per month
     ii. 100 – 150 hours per month
     iii. More than 150 hours per month
   - That is, their earnings fall into one of the three columns in Table 2
   - Combine income from multiple employers to compare against threshold

Table 2: Monthly earnings thresholds for new entrants

Tax year	Less than 100 hours per month	100 – 150 hours per month	More than 150 hours per month
2015	$1 - $1475	$1476 - $2212.50	>$2212.50
2016	$1 - $1525	$1526 - $2287.50	>$2287.50
2017	$1 - $1575	$1576 - $2362.50	>$2362.50
2018	$1 - $1650	$1650 - $2475	>$2475
2019	$1 - $1770	$1770 - $2655	>$2655
2020	$1 - $1890	$1890 - $2835	>$2835
2021	$1 - $2000	$2000 - $3000	>$300
2022	$1 - $2120	$2120 - $3180	>$3180

2. Collect all employees working in relevant industries with no income thresholds
3. Link the employees to personal information tables and evaluate their relevant demographic information, including their income grouping as defined in Table 2
4. Collect all employers in relevant industries with no income threshold set, including demographic information
5. Remove employers from employee dataset

6. Define new entrants as people entering workforce for the first time in any month in the given year, and collect these people from the overarching workforce dataset
7. Define the source of new entrants using the following definitions:
   - Visa table: Contains all new entrants who start work between issue and expiry date of visa and hold a Work or Student visa
   - First arrival table: New entrants who's first arrival into NZ is in year of starting work
   - Secondary students table: New entrants who finished secondary education less than a year before joining the workforce
   - Tertiary leavers table: New entrants who complete a tertiary education programme less than a year before joining the workforce
   - Career changers table: New entrants who meet previously defined income thresholds for fulltime work in year prior to joining workforce (i.e. would be considered a workforce member in a different industry) N.B. if they belonged to multiple industries, it defines previous industry as the highest source of income in the previous year
   - New employers table: All new entrants who have an income from a business i.e. shareholder etc. that started in the last year
   - The query gives priority in classifying employers over employees i.e. if someone is both it counts them as an employer
   - Returning Kiwis table: New entrants that joined the workforce in the past year who arrived into NZ within the past year and are not entering NZ for the first time
   - Beneficiaries table: New entrants who received income from a beneficiary in the year prior to joining the workforce
8. If someone meets multiple criteria the main query uses the following hierarchy of sources:
   i. Secondary
   ii. Tertiary
   iii. Migrants as visa holders
   iv. Career changers
   v. Beneficiaries
   vi. Migrants as first arrivals
   vii. Returning Kiwis
   viii. Other
   - That is, if someone meets the criteria for more than one source then they are counted as belonging to the first talent pool they appear in as ordered above. If they belong to no talent pools they are classified as "Other"
   - We also included a table that uses non-mutually exclusive sources, that is, someone can belong to multiple talent pools.

4. Tenure and retention

What do we mean by tenure and retention?

Tenure, as we define it here, is a measure of the cumulative work experience individuals in a workforce have at a point in time. Tenure here mainly focuses on *industry* tenure, i.e., the total number of months that a person's income meets an income threshold (if any) from any combination of employers in an industry in a time period. This is different to *job* tenure – the time spent working for a single employer. Tenure is a 'stocks' measure – that is, an attribute of the workforce at a point in time.

Retention, as we define it here, is a 'flows' measure. This measures how long a cohort of new entrants will remain working in an industry before leaving.

Both measures are useful to explain how long people stay in a workforce before they leave to pursue other avenues. These concepts are relatively simple at an intuitive level; however, they are not particularly easy to implement in the IDI, which we will discuss below.

Why is this of interest to WDCs

The industry tenure profile, and cohort retention profile, illustrate how long people are likely stay working in an industry. This is relevant to, for example, designing appropriate training packages, i.e., how specialist it is, when it is offered, how long it lasts. This measure also allows us to estimate the quantity of new entrant training needed. Replacement demand is typically a more important source of demand for training than demand created by industry growth – a fact often overlooked in workforce plans.

What types of questions are we interested in answering?

We would like to develop an understanding of the "tenure profile" of the CCRT industries. More specifically we want to track what proportion of the industry has a given level of experience. Another measure we are interested in is how long people stay in the industry, out of a given starting cohort. These will be covered by research questions as follows:

- How long have the workers at a given point in time worked with their current employer?
- How long have the workers stayed in the industry (from a given starting cohort)?
- What does the retention curve of new entrants look like?
- How have these changed over time (pre- and post- COVID)?
- We will also create an analysis with the following breakdowns:
  - Age
  - Gender
  - Ethnicity
  - Region
  - Training experiences
  - Qualification

Furthermore, we would like to be able to predict what percentage of the workforce will leave in any given year. However, this is much harder to define than first appears, see below.

What issues / points are worth consideration?

We must consider that a lot of turnovers are planned. People will often enter a workforce planning to only remain for a brief period. It is difficult to distinguish between this planned turnover and any unplanned turnover in the IDI. Typically, people do not just leave the workforce, many people enter and leave the workforce repeatedly. We need methodologies that allow for this. The IRD datasets we use to measure retention are not perfect, we are relying on employers having accurate PAYE records. These issues become more complex for part-time, casual, or seasonal workers.

What approaches did we take in the IDI?

There are multiple ways to approach retention in IDI research, we can develop tenure profiles, retention profiles and definitions of turnover. Our definition (one of them) of turnover is to only count people who have been in the workforce for a year e.g., turnover of people with > 1 year experience. It is important to look back through the work history of members to figure out if they are likely to return to the workforce or are truly leaving. We can triangulate around Census variables where it is flagged if people are part-time, full-time or volunteers, but it is of note that concordance from Census defined variables and IRD defined workforces means triangulation is not 100% reconcilable.

How did we structure the query?

The queries below are using the annual workforce that generated from workforce composition query, which currently uses 202206 data version, with no income thresholds and only contains people that have wage and salary income from certain industries..

Query 1: Tenure

1. Select the targeted workforce from the IRD tax return table: from selected industries in a given period, calculate the number of months that individuals satisfy income thresholds (if any) in each year. For each year, sum up the months in that year and years before as variable months_todate variable to signify the industry experience of that person until that month, sum up the months after that year until the individual leaves as the variable months_togo to signify how many months left ahead for that person in that industry. Here, months_todate column is regarded as the industry tenure (Example as below). For example, In the end of 2018, people with id '1' in the below table has been working in Café industry for 5 months and he/she will still work in this industry for another 7 months end of 2018 onwards).

Id	Industry	Start year	Months_todate	Months_togo
1	Café	2018	5	7
2	Airport	2019	4	9

2. By referring to metadata table and MOE tables, demographics and training information are added into the table in order to breakdown retention rates for different groups of people.

SCARLATTI

Query 2: Retention

1. Select new entrants to the target workforce from IRD tax return tables in selected industries and a given time period. Calculate the number of months that satisfy income thresholds (if any) for a person in each year and sum up as total industry tenure for that person in that industry (A short example is given as below).

Id	Industry	Start year	Tenure
1	Café	2018	5
2	Airport	2019	4

2. Add one column to signify each month of the person's tenure (Example as below).

Id	Industry	Start year	Tenure	Tenure_month
1	Café	2018	5	1
1	Café	2018	5	2
1	Café	2018	5	3
1	Café	2018	5	4
1	Café	2018	5	5
2	Airport	2019	4	1
2	Airport	2019	4	2
2	Airport	2019	4	3
2	Airport	2019	4	4

3. Once the above table is generated, we calculate the number of people that start in a certain year in an industry and the number of changes overtime, then calculate retention rates based on that.
4. By referring to metadata tables and MOE tables, demographics and training information are added into the table in order to breakdown retention rates for different groups.

5. Destinations

What do we mean by destinations?

Destinations, as we define it here, is where people go to after they complete training or leave the sector. We would like to depict a broad career pathway of people who a) leave the industry, and b) complete the sector-relevant training.

Why is this of interest to WDCs?

By investigating into the destinations of people that left the workforce, it could help answer which sector leavers have the intention of doing further training after leaving the workforce and how to design training packages based on their characteristics (demographics, region…).

By looking into the destinations of a cohort of people that have completed training, it can give us a sense of if qualifications are meeting the needs of the sector. For example, we can get an idea of whether short courses with a high transferability of skills are more useful than longer courses with a higher level of technicality. It also allows us to track the trainees that enter a workforce in the same sector as their completed training. This can give us a measure of if training is being offered at a suitable time. For example, if people complete training but leave the industry a brief time after, it may be more suitable to offer the training after being in the industry for a set period.

What types of questions are we interested in answering?

For people who leave the workforce, what is their main destinations at a given point (e.g., immediately or 1/3/5 years) after leaving the workforce? The destinations we are tracking are:

- Working in a different sector (Career changer)
    - For the career changers, which industries do they go to (relate to previous sector or not)?
- In further training
- Overseas
- Unemployed
- Beneficiaries/NEETs (Not Employed, Education or Training)
- Other not elsewhere classified

For people that complete sector-relevant training, we want to investigate:

- What length of time do people work in the relevant industries after completion of training?
- How many industries do people generally work in within a certain period (e.g., 5 years) after the completion of the sector-relevant training?

How do these vary for cohorts with different attributes? We will create an analysis with the following breakdowns:

- The highest qualification of members
- An age profile

- An ethnicity profile
- A gender profile
- A region profile

What issues / points are worth consideration?

In a similar fashion to workforce origins, we need to create a mutually exclusive hierarchy of destinations of workers/learners. In addition, we need to outline how far after leaving the workforce or completing training we are interested in. Also, it should be noticed that a person could be categorised into different destination group depend on how the exclusive hierarchy is ordered. For example, a person could be beneficiary (monthly income source code is 'BEN') but happen to be travelling overseas on the date that we select.

What approaches did we take in the IDI?

The approach when tracking destinations of workers in the IDI is like that of tracking origins of population members. We take cohorts of people and analyse their primary activity after leaving the workforce or completing training. For example, did they enrol in any further training, or were they paid wages from an enterprise in a different industry.

How did we structure the query?

These queries below are currently uses the annual workforce that generated from annual workforce composition query (202206 data version, no income threshold and only look at people that got wage and salary from certain industries).

For immediate destinations of workforce leavers:

1. Define workforce leavers in certain industries that we are interested in a specified year (i.e., people that became no IRD tax return records from wage and salary for certain industries in that given year)
2. Exclude leavers that came back to previous industries within a timeframe (within 1 year currently) and get completely workforce leavers
3. Define destinations of completely workforce leavers using the following tables:
   - IRD table: Contains people PAYE history which includes time, industry code and type of income etc.
   - MOE enrolment table & TEC table: Contain people who enrol in tertiary program or join in industry training (e.g., IT training)
   - DOL decision table: Contains passengers' travel information which includes departure and arrival airport, country and time etc.
4. For a given time after completely workforce leavers stepped out current industries, if someone meets certain criteria the query uses the following hierarchy of destinations:
   - Career changer: Leavers that entered new industries with wage and salary IRD records
   - Training: Leavers enrol in tertiary program or vocational training

    - Overseas: People that left New Zealand and didn't come back within a given time
    - Beneficiaries: People whose income source code in IRD table is tagged as 'BEN'
    - Other: Leavers don't fall into above categories (e.g., people look after family members at home etc.)
  - That is, if someone meets the criteria for more than one source then they are counted as belonging to the first destinations they appear in as ordered above. If they belong to no above categories, they are classified as "Other".

For destinations of workforce leavers in 1,3,5 years:

1. Define the time range that we want to investigate the mobility (currently from 2015-2022)
2. Select leavers from industries of interest during the time range by using existing workforce table generated from monthly IRD dataset
3. Match with IRD table to get their last IRD return dates and monthly incomes when left the previous industries
4. Match with IRD table, MOE table, TEC-it table, DOL decision table at the time of 1 year after their last dates and categories them into following hierarchy of destinations:
    - If people have income tax returns and the income source codes are 'W&S', regard as career changer
    - If the date is within a program from MOE and TEC table that the person enrolled, regard as tertiary
    - If the times of flying out of New Zealand is larger than the times of flying into New Zealand within the 1-year timeframe, regard as overseas
    - If the income source code of the person is 'BEN', regard as beneficiary
5. Repeat step 4 for 3-year, 5-year timeframe and record the destinations

For destination of training leavers:

This query is still in development.

6. Mobility

What do we mean by mobility?

We define mobility as the movement of workers within the sector. We would like to develop a broad picture of the pathways people take within the sector. We aim to evaluate the number of employers and industries that workers work with within several timeframes from a chosen starting point.

Why is this of interest to WDCs?

By exploring how frequently people move between different employers and industry groups within the sector, and how this varies for different demographic groups, it helps us to understand how vocational training should look like for those groups of people in the future. For example, providing small packets of learning supporting the development of more transferable skills might be beneficial for certain groups of people.

By looking into the frequency with which people who have completed training move between employers within an industry and between industries, we can begin to evaluate the value that employers place on vocational training within the sector. By exploring this we can also hope to gather a sense of whether vocational training is currently providing suitable skills for the sector.

What types of questions are we interested in answering?

- For people employed in the sector at a given point in time, how many employers/industries do the have over a certain period of time (e.g. 1/3/5 years)?
- How many months do people spend with the same employer/industry?
- How do these vary for different groups of people? We will break down the population by:
    - Industry groups
    - Age
    - Ethnicity
    - Gender
    - Region
    - Training status of workers

What issues / points are worth consideration?

We need to establish a point in time to evaluate mobility from. Initially we have used March 2015, although this could be changed for future versions of this query. This is not an analysis that we can evaluate over time, so picking a relevant starting point that will either give us a suitably long period of time to evaluate, or an interesting time in the workforce to consider (such as just before the Covid-19 lockdowns) can help to give us more valuable insights.

What approaches did we take in the IDI?

The approach in the IDI is similar to that taken for the skills analysis. We identify an initial cohort at a specific point in time (March 2015 for this analysis) and track them over time. We want to identify flows of people between employers and industries. We also identify relevant demographic details of the cohort, so we can evaluate whether mobility varies for different groups.

How did we structure the query?

This query uses 202206 data version, with no income threshold and only look at "Wage and Salary" income source code.

1. Select targeted population:

   Select people that worked in industries of interest in March 2015 (this date can be changed at the beginning of the query) from the IRD table with income source code equals to 'W&S'.

2. Calculate industry tenure:

   Match workforce population with IRD tables to calculate the total number of months that people have earnt income from their current industries as industry tenure

3. Add training status:

   Match workforce population with MOE completion and TEC-it tables to select the training that people have completed in the specified time period of interest (could be changed in the beginning of query, currently it is set to 2000, which means all months since 2000 are considered).

   Here, we distinguish between job relevant training and non-job relevant training by using a general mapping from NZSCED to ANZSIC, the mapping used is as followed:

1st digit of anzsic code	Industry	first 2 digits of NZSCED	Qualification field
A	Agriculture, Forestry and Fishing	05	Agriculture, Environmental and Related Studies
B	Mining	03	Engineering and Related Technologies
C	Manufacturing	03	Engineering and Related Technologies
D	Electricity, Gas, Water and Waste Services	03	Engineering and Related Technologies
E	Construction	04	Architecture and Building
F	Wholesale Trade	08	Management and Commerce
G	Retail Trade	08	Management and Commerce
H	Accommodation and Food Services	11	Food, Hospitality and Personal Services
I	Transport, Postal and Warehousing	08	Management and Commerce
J	Information Media and Telecommunications	02	Information Technology
K	Financial and Insurance Services	08	Management and Commerce
L	Rental, Hiring and Real Estate Services	08	Management and Commerce
M	Professional, Scientific and Technical Services	01	Natural and Physical Sciences
M	Professional, Scientific and Technical Services	08	Management and Commerce
N	Administrative and Support Services	08	Management and Commerce
O	Public Administration and Safety	08	Management and Commerce
P	Education and Training	07	Education
Q	Health Care and Social Assistance	06	Health
R	Arts and Recreation Services	10	Creative Arts
S	Other Services	11	Food, Hospitality and Personal Services

   For example, if people A works in 'Agriculture, Forestry and Fishing' industry has completed a training with first 2 digits of the training's NZSCED code equals to '05', then this training that he/she has completed is regarded as job relevant training.

   If people are found to have completed job relevant training from MOE or TEC IT tables, it is regarded as 'Completed industry relevant training' first, then if people completed non relevant training, it is regarded as 'Complete non-industry relevant training' or 'No training'.

4. Calculate job tenure and industry tenure after complete job relevant training:

For people that completed job relevant training from either MOE or TEC IT table, record the completion date and the first date of having wage and/or salary incomes after completion. Then calculate the first job tenure and first industry tenure.

5. Add gender, age, ethnicity, and region attributes

6. Calculate number of employers and industries within 1, 3 and 5 years:

7. Skills

What do we mean by skills?

In the scope of this research, skills are measured by the highest qualification individuals have completed.

Why is this of interest to WDCs?

Tracking the skill level of the workforce is fundamental to the work of WDCs. By understanding the stocks and flows of skills within the workforce, WDCs can gather insights into whether or not the training they currently offer are having the impacts that they were designed to do.

What types of questions are we interested in answering?

We aim to develop an indication of the stocks and flows of skills. Stocks describes the level of formal qualifications of a workforce while flows are a measure of the quantity of qualifications being provided to a workforce. These are determined directly from the training provider, and we can use completions as the measure we track. Relevant questions include:

- What is the highest qualification of the workforce?
- What are the qualifications the workers have gained prior to entering workforce?
- What are the qualifications that the workers have undertaken on the job?
- What are the associations between highest qualification and
    - Time in the industry?
    - Time with same employer?
    - Income level?

What issues / points are worth consideration?

This is remarkably hard from an IDI analysis perspective, the only source we have is the 2018 Census which is not the most robust. We can construct a historical training record of someone who has done the majority of their training in New Zealand over the last 15 years, where IDI records provide a more complete record of the training provided. For older training records or tracking people who have done training overseas, we do not have a complete record. We also do not have a complete dataset of qualification completions in the IDI.

Also, we want to use the occupation role as another way to measure skill, but this information is only available for people that participated in Household Labour Force Survey (HLFS), which could only give us a cross-sectional snapshot of small number of people.

What approaches did we take in the IDI?

This is fundamental to this work, but more work is needed to discuss which queries we want to construct around stocks and flows of training. Ideally, we want to create insights about correlations between qualifications and measures of mobility and measures of resilience defined above. How we define that is something we still need to work on.

How did we structure the query?

This query is using the 202206 data version, with no income threshold and only looks at the wage and salary income source code.

1. Select the workforce in industries of interest from IRD tables in the latest month (currently we use February 2022) and calculate their total monthly incomes
2. Add and decode demographics attributes (gender, ethnicity, age and region) by matching with IR personal information tables
3. Match with IRD tables to calculate the industry tenure and job tenure of individuals up to the latest month (Industry/job tenure here means the total length of time the person has worked in current industry/ for current employer)
4. Use the highest qualification from 2018 Census data, training history in MOE completion table and TEC-it-table to generate the highest qualification of workforce.
5. Group monthly income as followed:
    - Less than 2k
    - 2k-5k
    - 5k-10k
    - 10k-20k
    - More than 20k
6. Group tenure as followed:
    - Less than 1 year
    - 1-3 years
    - 3-6 years
    - More than 6 years
7. Generate output tables:
    - By industry x Highest qualification x monthly income
    - By industry x Highest qualification x job tenure
    - By industry x Highest qualification x industry tenure
    - By industry x Highest qualification x age
    - By industry x Highest qualification x gender
    - By industry x Highest qualification x ethnicity
    - By industry x Highest qualification x region

8. Resilience

What do we mean by resilience?

For the resilience of the workforce, we want to find out the ability of individuals in the workforce to be re-employed within a certain time after losing a job, especially during transitional events like the Covid-19 pandemic, and their monthly income change before and after moving between employers.

Why is this of interest to WDCs?

Understanding the different group of people in the workforce performed and adapted when facing external challenges helps us to identify some key factors in the ability of the workforce to gain re-employment. This will contribute to the discussion on how to further lift the workforce resilience.

What types of questions are we interested in answering?

- How long did it take for those who lost their jobs to be re-employed?
- Is the income from their new job similar to their previous job?
- How different attributes of individuals affect their ability to be re-employed, this includes:
  - Breakdown by ethnicity
  - Breakdown by education level
  - Breakdown by other demographics

What issues / points are worth consideration?

We define people that become unemployed as the people that have no IRD tax record in IRD tables at a point in time, and it should be noted that it's hard to tell whether this status change is planned or due to redundancy. Also, when unemployed people are looking for jobs, it is likely that they do some transitional job to cover life expenses until they get career-relevant employment. Based on these considerations, and in order to evaluate mobility more comprehensively, two evaluation criteria (time of looking for job and the monthly income change) should be combined.

What approaches did we take in the IDI?

The query is quite like the destination query with a slight change on the income filter to make comparisons consistent. Leavers of certain industries are retrieved by using the destination query, then we select career changers as the main targeted group to analyse.

How did we structure the query?

This query is currently using the 202206 data version, with no income threshold and includes income source codes only from 'W&S'.

1. Define the time range that we want to investigate mobility over (currently from 2015-2022)
2. Select leavers from industries of interest during this time range by using existing workforce tables generated from monthly IRD dataset

3. Match with IRD tables to get their last IRD tax return date and monthly income from when individuals left their previous industries
4. Add completed highest qualifications information to each year, when they left the industry (for people left before 2018, use MOE completion table and TEC-it table to get the highest domestic qualification; for people left 2018 onwards, use 2018 Census table, MOE completion table and TEC-it table)
5. Re-match to IRD table to get the next IRD tax return date and monthly income after the date of individuals leaving the previous industry, then calculate the duration between new tax return dates and last date of leaving as time spent looking for jobs for that individual
6. Calculate monthly income changes and group as followed:
    - Income increases more than 20%: 'Increase >20%'
    - Income increases between 10% and 20%:'Increase 10%-20%'
    - Income increases between 5% and 10%:'Increase 5%-10%'
    - Income changes between -5% and +5%: 'Almost the same'
    - Income decreases between 5% and 10%:'Decrease 5%-10%'
    - Income decreases between 10% and 20%:'Decrease 5%-10%'
    - Income decreases more than 20%:'Decrease >20%'
7. Group tenure as followed:
    - Less than 1 year
    - 1-3 years
    - 3-6 years
    - More than 6 years
8. Generate output tables:
    - By industry x year x months looking job x income change
    - By industry x year x age x months looking job x income change
    - By industry x year x gender x months looking job x income change
    - By industry x year x ethnicity x months looking job x income change
    - By industry x year x region x months looking job x income change
    - By industry x year x qualification x months looking job x income change
    - By industry x year x new entrant source x months looking job x income change
    - By industry x year x employment status x months looking job x income change
    - By industry x year x workforce status x months looking job x income change

9. Workforce specific queries

How many people have worked in Services at some point in their career to date?

Why is this of interest to Ringa Hora?

The services sector workforce has often been described as a more transient workforce compared to other sectors. Investigating this question helps to explore whether the service sector has been acting as a gateway for workers to enter the workforce and whether people are able to gain transferrable skills to benefit them in their future careers.

What issues / points are worth consideration?

What population of people are we interested in evaluating? Are we interested in everyone currently working in New Zealand, or only those in certain sectors? If we are interested in the current NZ workforce as a whole, we will need to allow for a long time for the query to run and potentially a lot of memory for the resulting tables. We also need to consider if we are interested in locating every instance of an individual working in Services, and their tenure, or if we are only interested in a binary yes / no label.

How could we structure the query?

1. Identify the initial population (i.e., current NZ workforce or sectors of interest)
2. Identify the income threshold to count someone as working in a particular month
3. Track back in time year-by-year until the start of reliable income records in the IDI (circa 1999)
4. For each year identify the ANZSIC codes of businesses making income/owner payments to each person each month and compare to ANZSIC codes of Services sector
5. At this point there are a few directions to take the query:
    - If we are only interested in a yes / no tag on whether people have worked in services, then for each person who is identified as working in the services moving backwards in time, we no longer need to consider them
    - If we are interested in each instance in someone's career they work in the Services sector, then we will not be able to refrain from counting every worker at every point in time moving back
    - If we are interested in the length of time that everyone has worked in services at this point, then the easiest way is probably to extend the tenure query backwards in time

Ethnicity as a focus

Evaluation of Māori in the sector has been a huge focus in the work we have done for Ringa Hora. And this could be a starting point to where we could potentially extend the research focus to other ethnicity groups, e.g., Pacific, Asian etc.

Gender and other demographic attributes as a focus

Building on our current research framework and IDI query structure, it is also possible to explore opportunities to extend our focus on other demographic attributes in the workforce, such as gender.

Stability

What do we mean by stability?

When we are evaluating stability in the scope of this work, we want to develop a breakdown of how likely people are to be changing sectors, leaving the workforce to train, or leaving the workforce as unemployed, along with other possible outcomes. This analysis will look at people at specific stages in life and is a measure of the resilience of a workforce.

Why is this of interest to WDCs?

The stability of a workforce is an indicator of the suitability of training programmes. For example, if there is a high churn of staff, it may not be ideal to offer training to new entrants immediately, but more appropriate to offer it after they have stayed in the workforce for a certain length of time.

What types of questions are we interested in answering?

There are several metrics we wish to track:

- How often do people in the sector change employer?
- How often do members of the sector workforce change industries within the sector?
- How often do people move out of the workforce?

Whilst researching these questions we will develop breakdowns by various attributes including (but not limited to):

- Age
- Ethnicity
- Gender
- Visa status

What issues/points are worth consideration?

We must decide how granular we wish to be with our analysis. That is, to what extent do we consider stability: movements in/out of the workforce, movements between industries or movements between employers. We are not researching this to pass judgement, only to provide insights into the workforce.

How could we structure the query?

We have not performed this analysis in the IDI previously, but we envisage it as being something we should pursue as part of this research.

Income

What do we mean by income?

When discussing income, we are referring to the total income of a person in a given period of time (typically a year).

Why is this of interest to WDCs?

This analysis allows us to measure the impact of training on learner outcomes. Income is a key learner outcome, and positive outcomes can justify investment from taxpayers, learners, and employers.

What types of questions are we interested in answering?

What happens to the income of learners in the years after training? We are interested in breaking this down for distinct groups of people and for several types of training.

What issues/points are worth consideration?

Ideally, we wish to make assertions of causality that doing training results in different income outcomes, however in practice this is difficult to do as we cannot create control groups. The best we can do is establish correlation.

What approach will we take in the IDI?

Our method is to track cohorts of learners in the IDI and calculate their total income each year from employment and self-employment. This is tracked yearly prior to and after completing their training.

Appendices

Appendix A - Data limitations and caveats

IDI confidentiality rules

In order to meet the requirements of Statistics New Zealand to release data from the IDI, counts must be suppressed below a certain threshold and randomly rounded to certain bases depending on the datasets used to derive the results. This process is in place to prevent individuals or entities from being able to be identified from results, but it does present some limitations to the way we manage the datasets.

The implications of these rules are that in most cases, after aggregating results across two demographics, counts will fall below suppression thresholds. This means that in order to be able to 'slice-and-dice' data in the large number of ways we aim to in this research, many outputs need to be created for each query.

Due to the nature of the random rounding element of the submission process, this means that counts across different aggregations will not necessarily sum to the same. The suppression aspect of the data release process also impacts this aspect, as different suppressions across aggregation splits can contribute to this as well.

Availability of data

As part of the research into Māori business composition, the ownership of businesses was identified within the IDI. Due to the definition of Māori-owned businesses we used in this process, ownership data is not available for every enterprise. This means that we cannot identify whether businesses are Māori-owned or not in every case (this is identifiable in approximately 20% of cases). As ethnicity is also self-reported, this is also not necessarily available for every individual.

Appendix B - Industry codes

Industry description (L2 ANZSIC)	Specific industry description (L4 ANZSIC)	Code	Toi Mai groupings
Computer Systems Design and Related Services	Computer System Design and Related Services	M700000	Toi Whānui – Enabling Technologies
Library and Other Information Services	Libraries and Archives	J601000	Toi Ora – Sport and Recreation
Library and Other Information Services	Other Information Services	J602000	
Heritage Activities	Museum Operation	R891000	
Heritage Activities	Zoological and Botanical Gardens Operation	R892100	
Heritage Activities	Nature Reserves and Conservation Parks Operation	R892200	
Sport and Recreation Activities	Health and Fitness Centres and Gymnasia Operation	R911100	
Sport and Recreation Activities	Sports and Physical Recreation Clubs and Sports Professionals	R911200	
Sport and Recreation Activities	Sports and Physical Recreation Venues, Grounds and Facilities Operation	R911300	
Sport and Recreation Activities	Sports and Physical Recreation Administrative Service	R911400	
Sport and Recreation Activities	Horse and Dog Racing Administration and Track Operation	R912100	
Sport and Recreation Activities	Other Horse and Dog Racing Activities	R912900	
Sport and Recreation Activities	Amusement Parks and Centres Operation	R913100	
Sport and Recreation Activities	Amusement and Other Recreation Activities n.e.c.	R913900	
Gambling Activities	Casino Operation	R920100	
Gambling Activities	Lottery Operation	R920200	
Gambling Activities	Other Gambling Activities	R920900	
Furniture and Other Manufacturing	Jewellery and Silverware Manufacturing	C259100	Toi-A-Ringa – Art and Design
Other Store-Based Retailing	Flower Retailing	G427400	
Publishing (except Internet and Music Publishing)	Newspaper Publishing	J541100	
Publishing (except Internet and Music Publishing)	Magazine and Other Periodical Publishing	J541200	
Publishing (except Internet and Music Publishing)	Book Publishing	J541300	Toi-A-Ringa – Art and Design

Publishing (except Internet and Music Publishing)	Directory and Mailing List Publishing	J541400	
Publishing (except Internet and Music Publishing)	Other Publishing (except Software, Music and Internet)	J541900	
Publishing (except Internet and Music Publishing)	Software Publishing	J542000	
Personal and Other Services	Hairdressing and Beauty Services	S951100	
Motion Picture and Sound Recording Activities	Motion Picture and Video Production	J551100	Toi Pāho — Broadcast and Screen
Motion Picture and Sound Recording Activities	Motion Picture and Video Distribution	J551200	
Motion Picture and Sound Recording Activities	Motion Picture Exhibition	J551300	
Motion Picture and Sound Recording Activities	Post-production Services and Other Motion Picture and Video Activities	J551400	
Motion Picture and Sound Recording Activities	Music Publishing	J552100	
Motion Picture and Sound Recording Activities	Music and Other Sound Recording Activities	J552200	
Broadcasting (except Internet)	Radio Broadcasting	J561000	
Broadcasting (except Internet)	Free-to-Air Television Broadcasting	J562100	
Broadcasting (except Internet)	Cable and Other Subscription Broadcasting	J562200	
Internet Publishing and Broadcasting	Internet Publishing and Broadcasting	J570000	
Artistic Activities	Performing Arts Operation	R900100	Toi Puaki — Expressive Arts
Artistic Activities	Creative Artists, Musicians, Writers and Performers	R900200	
Artistic Activities	Performing Arts Venue Operation	R900300	