MINISTRY OF SOCIAL DEVELOPMENT
TE MANATŪ WHAKAHIATO ORA

Effectiveness of MSD employment assistance

Technical report for 2023/2024 financial year

May 2025

Table of Contents

Author

Marc de Boer, Principal Analyst, Insights MSD, Strategy and Insights

Acknowledgements

I would like to thank the following people for their contributions and comments in preparing this report and previous versions of this:. Any omissions and errors remain the responsibility of the author.

Report disclaimer

The views and interpretations in this report are those of the researcher.

Integrated Data Infrastructure (IDI)

Some of the information contained in this report comes from the SNZ IDI. Below are the standard SNZ, IRD and NZDF disclaimers for this information.

Statistics New Zealand IDI disclaimer

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

Inland Revenue IDI disclaimer

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.

New Zealand Defence Force IDI disclaimer

The New Zealand Defence Force has consented to the release of IDI results for the Limited Services Volunteer programme to Stats NZ as part of this report.

Creative Commons

Published

Ministry of Social Development

PO Box 1556

Wellington

www.msd.govt.nz

May 2025

ISBN: 978-1-99-110583-7

Overview

This technical report explains the approach taken to evaluate the effectiveness of MSD's employment assistance (EA) expenditure for the 2023/2024 financial year.

In this report, we describe:

- what we define as EA interventions
- how we estimate the cost of EA interventions
- the outcome measures used to determine the effectiveness of EA interventions
- methods used to estimate the impact of EA interventions
- the method for estimating unobserved future impacts
- the process used to rate the effectiveness of individual interventions.

What are employment assistance interventions?

Employment assistance (EA) covers employment and training programmes, and services designed to help people prepare for, move into, and sustain employment. EA interventions funded by MSD are primarily targeted at people eligible for income support assistance. Table 1 below provides a broad classification of EA interventions.

Table 1: Classification of employment assistance interventions

Type	Description	Example interventions
Activation measures	Activation measures cover programmes and case management techniques designed to maintain job search activities for people expected to move into employment (eg people receiving unemployment related benefits). If people are judged not to be sufficiently engaged in job search then they can have their income support payments reduced or even cancelled.	Jobseeker Work Ready 52-week benefit reapplication, Work Obligations, Pre-employment drug testing
Work confidence	Programmes designed to encourage and motivate people to have the confidence to begin to move into either employment or further study.	Limited Services Volunteer, Outward Bound
Case management	One-to-one meetings with a case manager to discuss and plan on how they will move back into employment. Case management can be either in-house or contracted to an external provider.	Phone-based case management, Work to Wellness
Career advice /Information Services	Career advice and counselling is a standard service provided by public employment services to help jobseekers make informed decisions about their current and future employment choices.	Careers Guidance and Counselling
Health Interventions	Providing employment-focused health interventions (including integrated employment support and health care) to support people to improve their health and wellbeing and enable them to remain in, prepare for, or move into suitable employment.	Individual Placement and Support (IPS), Oranga Mahi services
Training (contracted)	Contracted training programmes aim to increase the foundational and vocational skills of participants to enable them to compete in the labour market.	Training for Work, Driver licence programmes
Training (financial)	Financial assistance to help people access education and training programmes.	Training Incentive Allowance
Work experience	Provide people with work experience either with a private sector employer or through placements with not-for-profit organisations to help in social or environmental projects.	Mainstream Employment Programme, Activity in the Community, Flexi-wage Project in the Community
Job search assistance	Seminars and job clubs designed to provide jobseekers with the skills to look for work (eg searching for job leads, CV and applications and interview skills) and to provide peer support to maintain motivation.	Work Search Support
Job placement services	In-house or contracted-out services to place people into paid employment. For contracted-out services,	Employment Placement or Assistance Initiative, Vacancy Placement

Type	Description	Example interventions
	payments are often based on a fee-for-outcome contracting model.	
Hiring wage subsidies	A temporary subsidy to compensate employers who take on disadvantaged jobseekers (ie they would not have been hired by the employer in the absence of the subsidy).	Flexi-wage
On the job training	Assistance to employees to help them gain skills whilst in work.	Mana in Mahi
Training for predetermined employment	Programmes that involve matching jobseekers to vacancies by providing short-term training to meet the specific needs of an employer (eg industry specific certificates or licences).	Skills for Industry
Self-employment assistance	Assistance to help people set up their own business. Self-employment assistance can involve a combination of training, mentoring, capital grants, and a temporary subsidy to cover living costs until business cash flow is sufficient to support the participant.	Be Your Own Boss, Business Training And Advice Grant, Flexi-wage Self Employment (subsidy)
Relocation payments	Provision of payments to help people move to take up employment outside of their local area.	5K to Work
Transition to work financial support	Financial assistance to help cover initial costs of moving into employment (eg work clothes and equipment) or to cover the period until the person is paid by the employer.	Transition to Work Grant
In-work support (Financial)	Financial assistance to help people with disruptions to employment or pay to ensure they can continue in employment and avoid returning to main benefit.	Seasonal Work Assistance, In-work tax credit
In-work support (Pastoral)	Programmes that contact people once they are in work to see how things are progressing and to help with any issues that might arise.	In-Work Support
Childcare assistance	Financial payments to low-income families to help cover the cost of childcare services.	Flexible Childcare Assistance, Childcare Subsidy, OSCAR subsidy
Incentive payments	Payments to people who remain in employment for set periods (eg after 3, 6 and 12 months).	In Work Payment, Work Bonus
Vocational Services	Contracted services to support disabled people to participate in employment or in their communities .	Vocational Services Employment
Youth Programmes	Assistance targeted at teens (usually under the age of 18) to help them remain in education, training or employment.	Youth Services
Migrant assistance	Assistance targeted at new migrants and refugees.	Migrant Employment Assistance
New Initiatives	Locally developed initiatives that cannot be easily categorised.	New Initiatives

Definition of an EA intervention spell

For our analysis we define EA interventions as discrete events that have the following attributes:

- Person identifier: system-id that determines who received the intervention
- Intervention name: name of the intervention
- Start date: calendar date the person started the intervention.

These elements are the minimum necessary to evaluate the effectiveness of the intervention. Additional attributes that we also try to collect include:

- End date: the date the person finished the intervention
- Referral date: when they were referred to the intervention
- Provider: who delivered the intervention, especially when it is contracted out
- Cost: how much it cost.

Treatment of short duration spells or non-completers

In the analysis of EA interventions, we include all participant starts and do not exclude any short participation spells or those recorded as not having completed the programme or course. There are two reasons for taking this position. The first is the difficulty in having reliable participation end dates or information on who completed the intervention. As discussed below, this information may or may not be recorded; it depends on the source system or the diligence of staff in recording these types of outcomes in the administrative system.

The second reason is that we consider early exits or non-completion as a core feature of the intervention which should be included in any assessment of its effectiveness. For example, if an intervention is being run such that many participants either exit soon after starting or fail to complete the programme, then this should be reflected in its performance.

Common issues with EA intervention data

Because EA information exists in more than seven MSD administrative systems compiling information about EA interventions is not always straightforward. As with most administrative data, there are several issues with how well EA intervention data is recorded.

Duplicate participation events

Participation events are defined as any recorded participation spell. In some cases, a person may participate in two different interventions on the same day. This occurs where a person may receive different forms of assistance (eg a Job Plus and Work Start Grant or Enterprise Allowance and Enterprise Capitalisation). However, there are also duplicate participation spells for the

same person, intervention and start date. When these occur, we select only one event.

In a small number of instances, we also take an intervention's participation from only one IT administrative source. These occur where business processes indicate that this is the primary system for recording intervention participation spells.

Inconsistent system information

In a substantial number of cases, there is more than one source of information for a variable for a given EA participation record. Important examples include the name of the intervention, provider name and participation start and end dates. We identify those records where these inconsistencies occur. The general approach to resolving these inconsistencies is to favour the source that is most associated with the event itself. For example, if a contract system end date differs from the front-line system recorded end date, we take the front-line system end date.

Participation end dates

One difficult area of EA participation is an accurate recording of participation end dates. Either end dates are missing, or they are miss-keyed, giving either implausibly long participation spells or end dates that are earlier than start dates. In many instances, it is necessary to impute end dates where they are currently null or implausible (eg a seminar lasting eight years). If there is information about the expected end date, then we use this when the actual end date is missing. If no suitable end date is available, we estimate the end date based on the observed duration that people spend on the intervention or a similar intervention, if required.

Referral dates

From an evaluation perspective, we are interested in when people are referred to interventions to identify who might have been approached about participating, as well as to identify likely effects of being referred to an intervention. For example, do we see a lock-in effect or the reverse, people exiting benefit in the period between referral and intervention start?

Referral information is generally unreliable for this type of analysis. In many instances, the formal recording of a referral occurs after an informal discussion and conversation with the intervention providers. Under these conditions, referral captures the point when a person is confirmed as intending to participate in the intervention. What is missed are those people where the case manager may have discussed the opportunity with the

participant and the individual turned it down or where a provider had screened the potential applicant out.

Estimating the cost of EA interventions

We use the individual Cost Allocation Model (iCAM) to estimate the cost of EA interventions for each financial year (MSD, 2017). Insights MSD created iCAM to provide a view of how spending to date has been allocated to outputs at the individual level. Here we define outputs as activities that MSD does to assist people such as a face-to-face meeting, a main benefit application, or an EA intervention.

Principles behind the cost allocation model

The cost allocation model works on the following principles:

- **Include all financial costs for Service Delivery (the operational arm of MSD)**: the model starts with appropriation[1] expenditure for all outputs delivered by Service Delivery. The reason behind this principle is to make sure we do not exclude any costs that are already recorded in the Ministry's financial systems. Having said this, income support payments designed to reduce income inadequacy are currently excluded, but we plan to include this information in later updates.
- **Reconcile allocated expenditure to financial totals**: for each appropriation, the model reconciles (as far as possible) the allocated expenditure back to the appropriation amount in each financial year. At the very least, the sum of the allocated expenditure in each financial year should not exceed the appropriation amount.
- **Disaggregate costs down to the individual output level**: to provide the highest level of accuracy and flexibility, the model disaggregates costs down to outputs (see the Cost allocation framework section below) at the person-event level. By doing so, we can accurately assess the amount of expenditure for individuals as well as retain the flexibility to summarise costs for any group of people. By building the model this way, we can also estimate the variability in the cost of delivering specific types of outputs.
- **Estimate the distribution of costs across outputs**: when possible the iCAM uses metrics that try to align with the actual variation in the

[1] We use the term here to refer to how public money is spent, see: https://treasury.govt.nz/publications/guide/guide-appropriations-html#section-1

cost of delivering a given output, rather than relying on simple averages.

- **Apply the same approach over all financial years**: by applying the same approach across financial years (from 2001/2002 onwards) it is possible to identify trends in the cost of Service Delivery outputs across groups of people. However, this also means it is not possible to compare results across different versions of reports or cost models.

Cost allocation framework

In this report, we briefly describe how the cost model works by using an example of an in-house seminar delivered by MSD. For a more detailed description, please refer to the iCAM technical report (MSD, 2017).

The process of delivering a seminar can be broken down into several components as listed in Table 2. For example, one component would be the time taken to book an appointment, alongside the seminar cost itself in the form of staff running the seminar. We first determine the total expenditure (see the Financial inputs section below) for each of these components by financial year.

Table 2: Cost components and their metrics

Component	Definition	Metric
Appointment	Scheduling an appointment	Staff time
Benefit administration	Assessing and maintaining entitlement to income support assistance	Staff time
Benefit payments	Bank fees for payment of income support benefits	Pay weeks
Client contact	Contact with individuals to help them plan and move into employment or time spent updating their records	Staff time
Contract Administration	Administration of contracts, including tendering, negotiation, payment and managing the performance of contracted providers	Contract amount
Contract payment	Payment of contracts	Contract amount
Grant	Financial transfer to people to assist them with further training or with transitioning into employment	Grant amount
Grant Administration	Assessing and administering grant applications	Staff time
Integrity (fraud and debt)	Identification of benefit fraud and the collection of outstanding debt	Staff time
Placement opportunity	Time spent by contact centre staff and work brokers to identify and establish vacancies with employers	Starts
Referral	Time spent by case managers in referring people to employment vacancies, employment programmes, or training programmes	Staff time

Component	Definition	Metric
Seminar	Staff time in administering and running seminars	Staff time
Study Assistance	Time in assessing and maintaining entitlement to student loans and allowances	Staff time
Wage Subsidy	Payments made to employers or sponsors in relation to wage subsidy, work experience, or self-employment programmes	Subsidy payments
Wage Subsidy Administration	Cost of administering wage subsidy assistance	Starts
Provider management	Staff time in managing service provider information and relationships.	Staff time
Unallocated Service Delivery	Unallocated frontline staff time costs for Service Delivery	Duration on income support or student allowance

The next step is to find a metric related to each component so that we can assign a dollar value to that component. We define metrics as quantitative information about each component of output. For example, for the appointment component, we can use the number of minutes that staff spent on booking participants for each seminar. Multiplying the number of minutes spent by staff cost-per-minute rate will give us the appointment cost for each seminar attendee.

Finally, we add the cost of each component to arrive at a total cost for the seminar. The variation in the cost of each seminar output for the financial year will depend on the variability in the cost of each of its components.

Financial inputs

Having identified the outputs, their cost components, and how to assign costs to them, the next question is where we source the financial costs for Service Delivery. We can access records of Service Delivery expenditure through the Ministry's financial accounting system. These records capture expenditure information down to the cost centre and general ledger (GL) nominal level.

With monthly financial data the next step is to link expenditure to cost components. For some cost components there is a relatively straightforward link to the financial inputs. For example, the wage subsidy payments for a wage subsidy programme have their own GL nominal code. For others the relationship is less clear. In particular, for those cost components that involve staff time, the component costs are a subset of the overall expenditure on staff costs recorded in the financial systems. In these instances, we need to apportion staff costs to components based on the estimated time it took to undertake each component task.

How do we estimate staff time?

Table 2 above shows that staff time is a commonly used metric in the model. However, obtaining this data is not straightforward. In this section, we summarise how we estimate the time spent on different activities. The source of this information is system transactions on MSD's various IT administrative systems combined with appointments, seminars and task management data. The key information for these transactions is:

- A unique ID for a staff member
- A unique ID for an individual
- A start time
- An end time
- What the action was.

This allows us to construct a transaction-based view of a staff member's day. Table 3 below shows an example for a staff member from the start of their day. For each period, the model identifies the type of action they are undertaking and measures the time until the next action based on the Time (end) value. If there is more than one action, then the elapsed time is split evenly between each action as shown in the Minutes column. Where client ID is missing, these represent periods where either the staff member is undertaking an action unrelated to a client (eg a lunch break) or the action exceeded the expected time it would have taken to complete the action. We have set the threshold of excessively long tasks at the 95th percentile for that activity over all staff on the same day. In cases where the activity exceeds the 95th percentile, the activity is split into two records, with the excess time is allocated to non-contact time in the model.

Table 3: Example of a staff member's actions from the start of their day

Time (end)	Action type	Action	Client id	Minutes
9:12:00	Case management	Search for client	10	5.52
9:16:00	Case management	Case Management	25	2.00
9:16:00	Case management	Scan Document	25	2.00
9:19:00	Income Support Administration	Third tier assistance	6	3.00
9:20:00	Income Support Administration	Third tier assistance	6	0.50
9:20:00	Case management	Case Management	33	0.50
9:21:00	Case management	Search for client	33	1.00
9:22:00	Income Support Administration	Maintenance	33	0.50
9:22:00	Income Support Administration	Third tier assistance	33	0.50
9:23:00	Income Support Administration	Third tier assistance	33	1.00

Time (end)	Action type	Action	Client id	Minutes
9:24:00	Case management	Scan Document	33	1.00
9:29:00	Income Support Administration	Maintenance	33	3.50
9:29:00	Non contact time	Non contact time	-	1.50
9:30:00	Income Support Administration	Third tier assistance	33	1.00
9:31:00	Case management	Case Management	14	1.00
9:37:00	Case management	Search for client	14	6.00
9:38:00	Case management	Search for client	14	1.00
9:47:00	Case management	Case Management	14	3.50
9:47:00	Non contact time	Non contact time	-	5.50
9:48:00	Case management	Search for client	14	1.00

We then link transactions to outputs that have components with staff time as a metric. These transactions should occur around the start date of the output, or within the start date and end date of the output, depending on the type of cost component. Also, staff transactions need to be of the same type. For example, staff time spent on income support administration is not linked to the management or delivery of employment programmes or services.

Outcome measures

When reporting on effectiveness, we measured the impacts of EA interventions across a range of outcome domains. We focus on those domains that we expect employment assistance to have a direct impact on as shown in Table 4 below.

However, we acknowledge that we do not have all outcomes that interventions could reasonably be expected to impact (eg outcomes such as children's short- and long-term outcomes, health status and household income). The absence of an outcome measure is often because the Statistics New Zealand Integrated Data Infrastructure (SNZ IDI) either lacks this information or there are issues with the data that need to be resolved. In subsequent reports, we plan to further expand the range of outcomes included in the analysis, to include those such as mortality and periods spent overseas.

Table 4: Outcome domains that employment interventions can be expected to have an impact on

Outcome domain	Included	Comment
Employment	Yes, Inferred from tax data	By definition, all employment interventions have a long- to medium-term goal of increasing time in employment.
Income	Yes, Labour market income and transfers to individuals.	While employment is important, interventions should not result in a reduction in overall income. Currently, we have not developed a measure of household income because of the difficulty of defining households within the IDI.
Education and training	Yes, Government funded training and education.	Many interventions have the goal of helping participants take up further training or education.
Qualifications gained	Yes, Only includes formal qualifications	In general, returns from education and training depend on the achievement of qualifications, especially higher-level qualifications. The gain in qualifications assumes an increase in human capital.
Confidence and motivation	No, Difficult to measure from current data sources	A common objective of case management and related programmes is to increase participant's confidence and motivation. The assumption is that increases in confidence and motivation will move people towards employment or further study.
Health care use	No, Data is available and we plan to include this outcome in later reports	A number of employment interventions involve the purchase of health care or help participants to access health care. The assumption is that resolving health needs will enable participants to move back into employment.
Health status	No, Limited data that can be used for impact evaluation	A number of employment interventions involve the purchase of health care or help participants to access health care. The assumption is that supporting improved health outcomes will assist participants to move back into or sustain employment. We may be able to look at the use of acute health care as an indicator of poor health.

Outcome domain	Included	Comment
Justice/offending	Yes	An expected indirect impact of employment is a reduction in criminal behaviour.
Children's immediate outcomes	No	EA interventions may have impacts on the immediate outcomes of participants' children. For example, improvement in income may result in better health.
Children's long-term outcomes	No	A long-term impact of EA interventions targeted at sole parents may be seen in the adult outcomes of their children. We are beginning to reach follow up periods (eg 18 years) where this analysis may be feasible.
Mortality	No, Include in next update	While most interventions are not intended to impact on mortality directly, this could be a long-term impact.
Time overseas	No, Include in next update	One consequence of increased employment may be a reduction in the probability of moving overseas (eg to Australia to find work).

Income

Total income is an important measure of a family's overall wellbeing. In the current analysis, we are restricted to looking at the income of individuals only. Ideally, we would like to measure household income to better account for the overall material wellbeing of individuals (eg supporting children or non-working household members). However, we have not yet developed a suitable measure of household income that can be used for the evaluation of EA interventions.

Net income from all sources

Net income from all sources is the main outcome measure. It includes all sources of income but excludes the drawdown of student loans. Income is net of tax. The measure was based on Inland Revenue (IR) and MSD data provided to the SNZ IDI.

The current income measures include:

- **Employer Month Schedule (EMS)**: New Zealand operates a Pay As You Earn tax system. Accordingly, all employers provide IR with monthly schedules of the earnings of all their employees. In addition to employee earnings, the EMS also includes taxable income support (main benefit), Accident Compensation Corporation (ACC) and pension payments.
- **Self-employment and company earnings**: people who run their own business or company are also required to file annual tax returns. In the analysis, these annual returns are converted into monthly spells with annual totals split equally across the 12 months of the tax year. There can be considerable lags in the lodging of self-

employment earnings that can mean measures of income for the most recent periods underestimate actual income. Note, however, because we update the analysis regularly the results incorporate these lags in reported earnings in later updates.

- **Non-taxable income support payments**: not all income support payments are subject to tax. Second-tier assistance, such as the Accommodation Supplement and third tier or ad hoc assistance such as Emergency Food Grants are not taxed. For hardship payments, we exclude recoverable assistance, as these are advances on main benefits. Recoverable payments will either be reflected in lower main benefit payments or, if the person moves off a main benefit, in the form of an income support debt. At present, we do not have reliable data on income support debt.

Income sources not covered by the current measure:

- **Tax credits**: in the current analysis we have not included tax credits. IR has recently supplied tax-credit datasets to the IDI and we are in the process of developing business rules to extract this information.
- **Child Support**: transfer payments between custodial and non-custodial parents that are administered by IR.
- **Non-taxable income support payments to people over 65**: current income support data supplied by MSD to the IDI exclude payments to people on New Zealand Superannuation or Veteran's benefit.
- **Illegal and undeclared income**: the IDI data does not cover income from informal or illegal activities, including tax evasion.

Income support received after tax

Income support payments are both taxable and non-taxable. For consistency, we calculate the total amount of income support a person receives after tax. Because of data limitations, income support only includes second and third-tier income support payments for working-age people. By using IR data, we can include income support payments for people receiving New Zealand Superannuation payments. As noted above, we have not yet included tax credits into this measure.

Employment

Any time in employment

Employment is based on the period that people declare income from employment or self-employment. Note that employment spells are based on either monthly or annual periods so we may be overstating or understating the actual time a person is in employment depending on where in the month or tax year they started or ended employment. At present, we have not attempted to adjust for this (eg by looking at the following or subsequent month to identify the likely start and end periods).

There are also lags in lodging tax returns, with these most pronounced for annual returns. We choose not to censor our analysis period to accommodate these lags and instead rely on regular updates to the analysis to incorporate delayed tax data into the results.

Time in employment while on main benefit

Here we include periods where a person is both on main benefit and receiving employment income. Sole parent benefits are designed to allow people to remain entitled while earning relatively high levels of income from employment. Similarly, people on health and disability-related benefits may only be able to work part-time.

Time in employment and independent of Work and Income

Here we are interested in employment without support from main benefits or employment assistance. This measure is particularly useful when looking at subsidy-based interventions that mean participants are in employment but are supported indirectly through a subsidy.

Justice

We have two sources of information on justice outcomes: police offending and periods under Corrections supervision. These data are also obtained from the SNZ IDI.

Any offence

This measure is based on Police data of people who are arrested for an offence (but may not result in a prosecution). Note that offending data is only available from 2009.

Time spent in any Corrections spell

This includes any spell under Corrections supervision and covers periods of custodial and non-custodial supervision (such as prison, Community Service, home detention, remand, parole, and Periodic Detention).

Time spent in prison

An important subset of Corrections supervision spells is time spent in prison. Accordingly, we also include time spent in prison as a separate outcome.

Education qualifications

Educational achievement information is based on secondary and tertiary qualifications achieved. We include school, tertiary, industry training and targeted training qualifications data. There is a considerable reporting lag for qualifications data in the SNZ IDI, and normally qualifications data are out of date by over 12 months. Also, qualification data only provides the year the qualification was attained. In our analysis, we assume that the qualification was attained at the end of the year (ie 31 December). For analysis of intervention impacts, we exclude qualifications gained in the year the participant started the intervention as we cannot know whether they achieved the qualification before or after starting the intervention.

Here we assume that gains in education qualifications reflect improvements in human capital. This may not always be the case. For example, analysis by Crichton (2013) found people on income support who achieved low-level qualifications (NQF3 and below) appeared not to gain any benefit in terms of subsequent employment or income. Similarly, this measure ignores any human capital gained through informal means.

Qualifications achieved at NZQCF Levels 2, 3, 4

For each person, we construct spells when they have achieved a specified minimum NZQCF level. NZQCF levels start at 1 (first national school assessment) through to 9 (doctorate). For each individual, we identify the date they first achieved the specific NQF level.

Highest NZQCF level

The highest NZQCF level is the highest NZQCF level achieved by a person at a specific date. From this measure, we can calculate the average NZQCF level achieved by the participant and comparison group of an EA intervention.

Education participation

Participation in further education and training provides an early indication of whether people are engaged in developing their human capital. The unit of measurement for this outcome is the number of days enrolled. However, people may not be attending training even when they are enrolled.

Time spent in any education participation

For any education participation, we combine all education spells in school, tertiary, industry training and targeted training.

Time spent in education while off benefit

Education participation spells where a person is also off a main benefit (based on benefit entitlement spells).

Education participation NZQCF4

Time spent participating in education courses at NZQCF Level 4 or above (broadly equivalent to post-school qualifications).

Welfare

Welfare covers the cost of income support payments. In this analysis income support payments include main benefits, supplementary assistance and one-off payments (non-recoverable), but excludes tax-credits paid by IR as well as study assistance such as Student Allowance.

Income support payments

All income support payments for main benefits as well as supplementary and one-off payments. For one-off payments only non-recoverable payments are included. Income support payments are net of tax.

Driver licence status

We can measure progress through the New Zealand driver licence graduation system. The system started in 1984 and has been through several changes. People go through learner and restricted licence stages before getting a full licence.

Time spent while holding a learner's, restricted and full licence

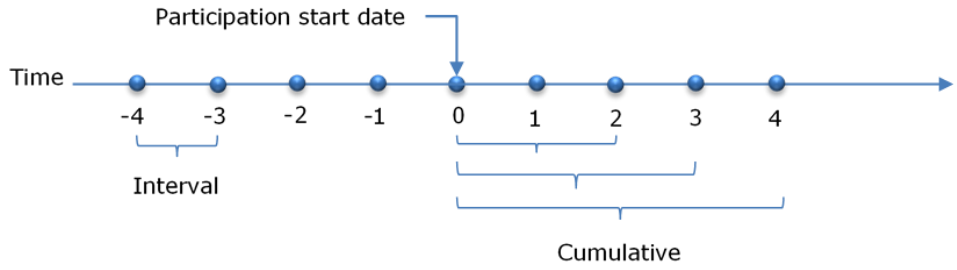
We can track the number of days spent at each driver licence stage that enables us to accurately track people's progression towards a full driver licence over time.

Tracking outcomes longitudinally

It is useful at this point to explain how we analyse the outcomes relative to participation in EA interventions. The outcomes described in the previous section are all longitudinal. Therefore, we can measure outcomes at multiple points in time rather than being limited to a small number of measurement periods as would be the case for survey-based outcome measures.

This flexibility allows us to track outcomes relative to participation start dates as shown in Figure 1. The first point to make is that we measure outcomes from when people start an intervention, and this is defined as zero on our timeline (we explain why below). From the zero point, we can then create a series of lapse periods that represent the periods before and after the participation start date. Based on this timeline, we can measure outcomes in two ways: interval and cumulative.

Figure 1: Tracking EA intervention outcomes using administrative data



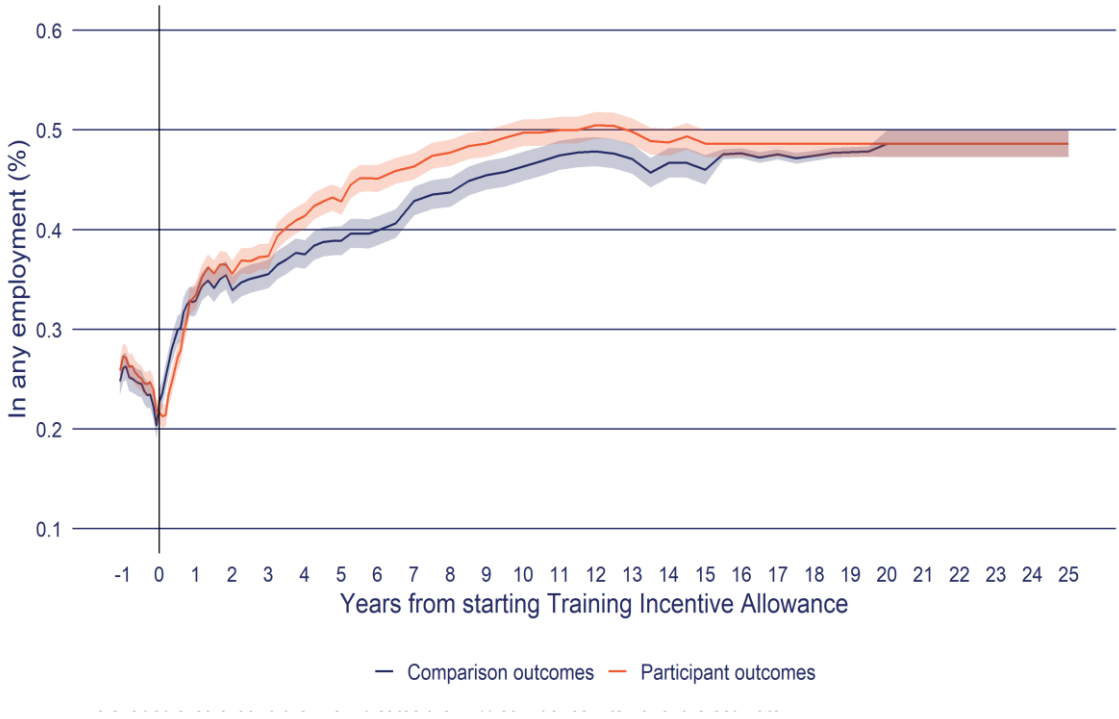
Interval outcomes

Interval outcomes are measured within a discrete lapse period, for example, the amount of income a person earned in the 12th month after starting an intervention. These intervals can vary in duration from one day to any period, but for EA interventions we usually use 30-day intervals. Figure 2 below shows, for Training Incentive Allowance (TIA), the percentage of each lapse interval that the participant and comparison group spend in employment. For example, at one year before starting TIA, the participants spent 27±1.0% of the period in any employment, while this proportion was 26±1.0% for the comparison group.

Tracking interval outcomes is most useful in understanding the dynamic relationship between the intervention and the outcome in question. The purpose of EA interventions is to change the outcome trajectories of participants. Looking at how outcomes change in each lapse interval before and after commencing an intervention provides important information on the likely behavioural responses to the intervention.

To return to the TIA example in Figure 2, we can see that the employment outcomes of participants are less than that of the comparison over the initial six months after starting TIA (lock-in effect). However, over later intervals, the outcomes of the participants exceed that of the comparison group. In other words, after completing the intervention participants are more likely to be in employment than the comparison group.

Figure 2: Interval employment outcomes for Training Incentive Allowance 2007 participants and matched comparison group



2: In any employment: Employment is based on tax data (PAYE and annual tax returns). Periods with less than $100 of real (at report year) employment income per month are excluded. Annual returns are left censored to lapse period 0 if they start before the lapse period 0 calendar date.

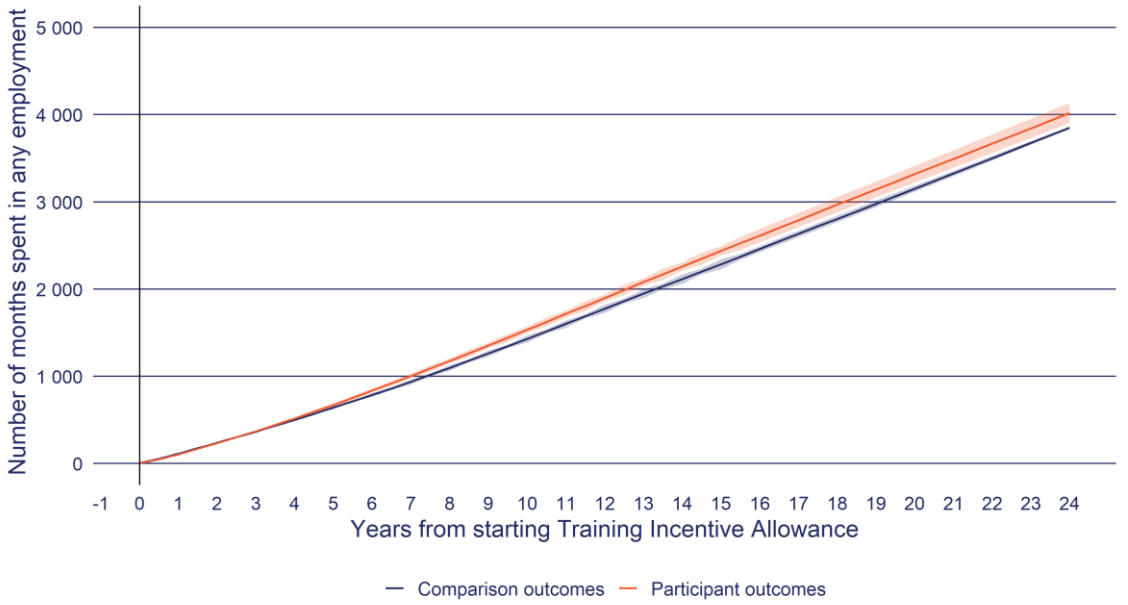
Cumulative outcomes

While interval outcomes are useful to understand how outcomes and impacts change relative to when people start an intervention, they do not allow us to quantify the overall impact of an intervention. To make summative judgements we use cumulative outcomes. As Figure 1 shows, cumulative outcomes start from period zero to each subsequent follow-up

period. For example, how much income did participants receive over the 12 months after starting the intervention?

Figure 3 shows the cumulative outcomes for the same participant and comparison groups in TIA as illustrated in Figure 2 above. Cumulative outcome measures only cover the period after participation start and not before. Figure 3 shows the average months over each successive period after starting the intervention that participants and comparison spent in employment. At the end of two years, participants spent an average of 7.80±0.20 mths, increasing to 63.0±1.40 mths after twelve years. In other words, the cumulative outcomes are simply the sum of the outcomes achieved in each successive interval after the intervention started.

Figure 3: Cumulative employment outcomes for Training Incentive Allowance 2007 participants and matched comparison group



1: The shaded area around the line indicates the 95% confidence interval of the estimate.
2: In any employment: Employment is based on tax data (PAYE and annual tax returns). Periods with less than $100 of real (at rep income per month are excluded. Annual returns are left censored to lapse period 0 if they start before the lapse period 0 calendar

Source: Statistics New Zealand. Integrated Data Infrastructure. June 2024.

Using cumulative outcomes, we can conclude that participants spent longer in employment than the comparison group a difference of 18.0±8.90 wks (63.0±1.40 mths compared to 59.0±1.50 mths).

Why measure outcomes from participation start?

A common question is why we measure outcomes from when people start an intervention, rather than when they finish. There are two reasons. The first is practical, in that when people finish an intervention is often poorly

recorded. Therefore, the date when people finished participating in an intervention is much less certain than the date they started.

The second reason is the importance of capturing the full impact of an intervention. As Figure 2 above shows, the period while a person is on a programme can have an impact on their outcomes. The most common impact is referred to as the lock-in effect. As the name suggests, while people are participating in an intervention, they are less likely to achieve an outcome, such as moving into employment. This can occur for several reasons. One is simply the reduction in time participants have available to look for work. And, for training programmes, the need to complete the course to gain a qualification provides an incentive to turn down job opportunities if they do arise. If we did not include these effects, we would run the risk of overstating the positive impact of interventions.

Migration and mortality

In the current analysis, we have not adjusted for people moving out of New Zealand or dying. These events would, over longer follow-up periods, reduce the denominator for each of the above outcome measures (ie right censoring). We plan to adjust for right censoring in subsequent updates to this analysis.

Estimating the observed impact of EA interventions

To rate the effectiveness of EA interventions we need to determine their impact on outcomes to date. In this analysis, we estimate effectiveness using counterfactual designs.[2] The term counterfactual refers to the question: what would have happened in the absence of the intervention? The problem posed by this question is that it is not possible to observe the counterfactual outcomes of participants. The solution is to identify a proxy for the counterfactual, usually a group of non-participants whose outcomes represent the counterfactual scenario. The challenge is to ensure that the comparison outcomes are an accurate representation of participants' counterfactual outcomes. Specifically, other than programme participation, are there other reasons for any differences between the outcomes of participants and those of the comparison group (ie selection bias)?

Various methods can control for selection bias to a greater or lesser degree. To assist readers in judging the robustness of a counterfactual design, we categorise methods according to the Scientific Maryland Scale (SMS). The SMS scale ranks counterfactual designs from 1 (least robust) to 5 (most robust). Robust in this context refers to the level of confidence we have that the impact estimate accurately quantifies the causal effect of the intervention on the outcome.

In the current report, we have four designs: randomised control trial (SMS 5), propensity-matched comparison group (SMS 3), propensity-matched historical comparison group (SMS 3(-)) and natural experiments (SMS 3) designs. We outline each in turn.

Randomised control trial designs

Randomised control trial (RCT) designs are the most robust counterfactual designs as they require the fewest assumptions and therefore can make the strongest quantitative statements about the causal relationship between participation in an intervention and later outcomes. RCTs in the context of MSD EA interventions have been used extensively to evaluate the impact of

[2] It is important to emphasise that quantitative counterfactual designs are not the only or primary evaluation method. To fully understand the effect of an intervention requires a mixed-method approach. Specifically, we need additional information to help understand the context and operation of the intervention itself to fully explain why the intervention has the impacts that it does. Similarly, not all outcomes are always quantified in a way suitable for impact evaluation.

case management services, such as Work Focused Case Management or Investment Approach Trials (MSD, 2018).

Propensity score matching (PSM)

Propensity score matching (PSM) is the main method we use to estimate the impact of EA interventions.

PSM is a common alternative to randomisation. It estimates the counterfactual by constructing a matched group of non-participants who have the same (or similar) characteristics as the participants. These non-participants are drawn from the same population as the participants. For MSD funded EA interventions, this is primarily people receiving income support. PSM is one of a group of methods referred to as quasi-experimental designs that attempt to replicate the same conditions as a randomised control trial. However, in all instances, quasi-experimental designs rely on additional assumptions that make them less robust than RCTs.

Before outlining PSM, it is useful to think of an intuitively appealing alternative of exact matching. Exact matching, as the term suggests, is to match a participant to a comparison with the same characteristics (eg same age, gender, benefit history and so on). However, exact matching is limited by the probability that two people share the same set of observable characteristics (and is also unnecessarily restrictive)[3]. The more characteristics included in the exact match, the less likely it is to find a comparison person with the same characteristics for each participant. As a result, these methods require the arbitrary selection of only a few matching variables.

Propensity matching overcomes this problem by using a logistic regression model to relate observable characteristics to programme participation. The logistic regression produces an estimate of how likely a given individual is to be a participant in a programme. It is possible to use this likelihood (called 'the propensity score') to match participants and non-participants based on the similarity of their propensity scores. In effect we match each participant on the date they start an intervention with a non-participant with the same likelihood of participating in the intervention. If the propensity score is properly specified, the participants and matched comparison groups will have a similar observable characteristic profile (eg similar duration, benefit type, age, number of children).

[3] Within a randomised control trial, the treatment and control groups share the same statistical profile, rather than each treatment group member having an identical twin in the control group.

PSM for EA interventions

The PSM for EA interventions involves first identifying EA interventions suitable for PSM. We then split interventions by cohort intakes, for example a large programme will have cohorts by calendar year (ie undertake separate PSM for participants starting in 2008 and then 2009). For smaller interventions, years need to be combined to have enough participants (eg participants starting between 2009-2012). In the following discussion, each PSM is applied at the level of a participant cohort (ie intervention by start year(s)).

For PSM EA interventions we use nearest neighbour matching with replacement. In other words, for each participant we select the non-participant with a propensity score closest to the participant's score (ie nearest neighbour). "With replacement" means we allow the same non-participant to be matched to more than one participant. This approach ensures the closest match in scores between the two group, but at the expense of having fewer unique comparison group members.

Finally, we do not exclude any participants from the matching because of large differences in propensity score with the nearest non-participant. Therefore, if there are common support issues (ie the distribution of propensity scores are different between participants and non-participants) these will result in poor balance test results (see next section) and will not be used for analysis of impact.

Conditional Independence Assumption

A key assumption for propensity score matching is the Conditional Independence Assumption (CIA), which states that controlling for differences in observable characteristics between the participant and comparison groups also controls for unobserved differences between the two groups. To be more precise, this requires that the two groups are equivalent on those characteristics that influence both the outcomes of interest as well as the probability of participating in the intervention, and this is true for observed as well as unobserved factors.

Estimating impact by controlling for observable characteristics requires that the CIA holds. If it holds, the only statistically significant difference between the participant and comparison groups will be their participation in the programme. Any resulting estimates would be unbiased. In other words, the only explanation for differences in subsequent outcomes between the two groups would be whether they had participated in the programme or not. If the CIA fails, the estimates will be biased. Here differences in subsequent outcomes could be due to unobserved differences between participants and their comparisons, as well as the impact of the programme.

The main limitation of the propensity matching method is that it relies on available and measurable information about people likely to participate in the EA intervention. Comprehensive information rarely exists about the types of people who participate in the programme or those who could form part of the comparison group (see the next section on using the SNZ IDI). The second limitation of the CIA is that it is not possible to determine whether it has been violated or, if it has, to what extent. Instead, we make a judgement as to whether the profile information is sufficient to accept the CIA. For a number of interventions such as those that involve people on long term health condition and disability benefits we do not think we have enough information to identify why a given individual would have chosen participate or not. In such instances, we say it is not feasible to estimate the programme's impact using non-experimental methods.

Eligible population and profile variables

The PSM analysis was undertaken in the Statistics New Zealand Integrated Data Infrastructure (IDI), which is a data platform for researchers that links anonymised individual-level information across several domains ranging from health care through to driver licence status. While researchers have access to individual-level data, all outputs are aggregated with measures in place to protect the privacy of individuals, firms, and institutions. Statistics New Zealand reviews all IDI output to ensure that these measures have been implemented.[4]

The IDI is well suited to undertaking PSM for two reasons.

1. The IDI has information on the entire New Zealand population, allowing the selection of a potential comparison group from the largest pool of potential matches possible.
2. The credibility of PSM is based on the inclusion of a rich set of profile variables and using linked data from a wide range of administrative, census and survey sources enables the creation of such a profile across a wide range of domains.

Profile variables

Table 5 summarises the domains of the variables included in the PSM for EA interventions.

[4] For more detail on the SNZ IDI, please visit https://www.stats.govt.nz/integrated-data/integrated-data-infrastructure/

Table 5: Summary of profile variables used in propensity matching

Area	Description
Demographics	
Age	Age group
Gender	Gender identity, only includes male and female.
Ethnicity	Total response, SNZ level one ethnic identity.
Education	
School	Information on the type of school (state or private), the decile of the school, the number of schools attended, suspensions, standdowns, truancy and special education support.
Tertiary study	Time enrolled in tertiary study by NZQF level and enrolled in study at set months before participation profile date.
Qualifications	Highest qualification based on education, census, or MSD data sources. Highest qualification is measured a set lapse periods before profile date to account for any changes in qualification status before starting a programme. This control is most important for younger people whose qualification level can change over relatively short periods.
Health and disability	
Incapacity information	Recorded incapacity information for people who have applied for Health Condition or Disability related benefits. A person can have up to four recorded incapacities at any one time. There are two measures, one for current incapacity status and one for incapacity in the last 5 years.
Mental health	Indicators of mental health care access including use of pharmaceuticals.
Location	
Deprivation index decile	The NZDep is an area-based measure of socioeconomic deprivation in Aotearoa New Zealand, it measures deprivation at SA2 level with decile 1 representing least deprived areas and 10 the most deprived. SA2 geographies aim to reflect communities that interact together socially and economically (eg at the level of a suburb or small town).
Urbanisation of location	SNZ classification of the person's location from major urban area through to rural as well as overseas.
Local labour market	Labour market information on the location a person lives (SNZ SA2 geographies), including average income, employment or study rate, average qualification level, working age population on main benefit and the dependency ratio.
Housing	
Number of address changes	Number of changes in recorded address over the last two years.
Employment	
Duration in employment	If currently employed the duration in their current spell of employment.
Duration since last employment	If not employed, the time since last employment.

Area	Description
Working life in employment	Proportion of working life (16-64) spent in employment, excluding time living outside New Zealand or before the year 2000.
Employment history	Employment status at set months before profile date.
Income Support	
Current benefit status	Current main benefit information.
Benefit duration	Duration on current main benefit.
Recent benefit history	Previous main benefit received.
Total benefit contact	Proportion of adult life spent on different types of main benefit.
First benefit information	Age and which benefit a person was first granted.
Childhood benefit receipt	Time that care givers where receiving a main benefit split by age group.
Income support history	Total income support payments at set months before profile date.
Justice	
Police offences	Includes number of offences, the time since last offence, the most serious offence and age of first arrest.
Corrections spells	Total time spent in different Corrections services, age of first Correction contact and time since last Correction involvement.
Youth Justice	Number of youth justice referrals and time spent in youth justice placements.
Corrections history	If in a correction service at set months before profile date. Correction service is split between prison and non-prison service.
Income	
Income history	Total net income from all sources, labour market income and child support payments at set months before profile date.
Residency	
Migrant status	Identifies time spent living in New Zealand, age of first arrival in New Zealand, Migrant's first arrival visa, including if arrived as a refugee, region of origin.
Overseas	
Overseas history	Whether a person is overseas at set lapse periods before profile date.
Employment assistance	
Participation in employment assistance	Expenditure on MSD funded employment assistance programmes and services at set months before profile date.
Care and Protection	
Care notifications	Notifications to child protection agencies, split by age group.
Care placements in childhood	Time spent in child protection placements, split by age group.
Transport	

Area	Description
Private driver licence	Private motor vehicle status at set lapse periods before profile date.
Commercial driver licence	Commercial driver licence status..

One strategy to ensure participants and matched comparison group have similar expected future outcomes is to include key measures of those outcomes in the profile. In particular a number of profile variables related to outcomes such as employment and education and training are measured at set periods before the profile date. The current periods are 1 to 12, 15, 18, 21, 24, 30, 36 and 42 months before profile date. The purpose of measuring profile variables at set periods before profile date is to account for trend in outcomes leading up to participation in an intervention. For example, it is important to account for the often-observed downward trend in employment and increased benefit receipt by participants in the months before starting an intervention.

Quality of the matching, the balance test

While we cannot test if the conditional independence assumption (CIA) has been violated, we can check to see if the comparison group has a similar average profile to the participants. This is referred to as the balance test, with balance referring to whether the profiles of the participants and comparison group are similar to each other. The balance condition can be expressed as,

$$P(D) \perp X$$

where $P(D)$ is the probability of participating in the programme, X is a set of observable characteristics, the '$\perp$' indicates that $P(D)$ is independent of X. One way to test this condition is to predict D based on X, using a logistic model,

$$\frac{D}{1-D} = exp(\alpha + x_1\beta_1 + x_2\beta_2 + \cdots + x_n\beta_n)$$

where, the target is membership of the participant group (D=1) or the matched comparison group (D=0), and X is the set of all the profile variables available for matching (see Table ??). Balance is achieved when the logistic model cannot predict D and the model fit is poor. In other words, the regression model cannot identify if a given individual is in the participant or matched comparison group based on the available observed characteristics.

To test model fit, we use the area under a receiver operating characteristic (ROC) curve, abbreviated as AUC. The closer the AUC is to 1 the better the model is at predicting whether a given observation is in the participant or

comparison group (ie a low false prediction rate). The lower bound of the AUC scale is 0.5, where the model cannot predict whether a given observation belongs to the participant or matched comparison group.

The next question is determining how high an AUC would need to be before we consider the profiles are unbalanced (ie the profiles of the participant and matched comparison group are not the same). To set this cut-off, we determine the expected AUC based on randomising an equivalent set of individuals into a control and treatment group. We achieve this by combining the participant and matched comparison group into a pooled sample. From this pooled sample, we randomly allocate half to treatment and the other half to a control group. In other words, we replicate an RCT where membership to the control or treatment is, by definition, independent of X (ie $P(D) \perp X$) and then proceed to calculate the AUC.

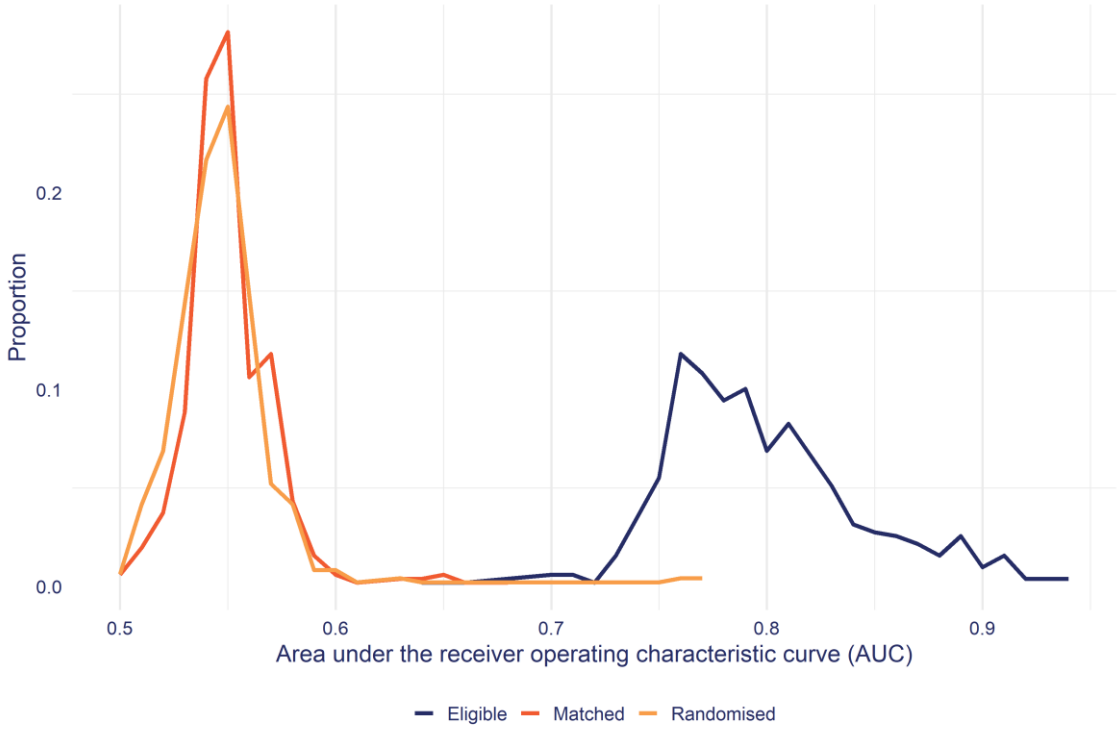
We repeated this process 100 times to generate an expected distribution of AUC for randomly allocated control and treatments drawn from the pooled matched participant and comparison group and using the same profile information. Figure 4 shows the results for randomised, matched and eligible AUC for all the individual EA interventions where PSM has been used. The Matched line shows the AUC for PSM matched, while the Randomised line shows the AUC distribution if these PSM had been randomly assigned to a treatment and control instead. The Eligible shows the AUC for participants and a sample non-participant group with a greater than zero probability of participating in the intervention.

From Figure 4 we can make the following observations:

- The average AUC for Eligible is 0.8, in other words, a regression model can identify to a high degree of accuracy whether a person is a participant or non-participant based on their observed characteristics. This result provides compelling evidence that participants differ in important ways from the eligible population. Such differences occur through a combination of institutional practices and guidelines, case manager preferences and assessments as well as self-selection decisions by participants themselves.
- The Randomised AUC distribution, by contrast, is close to, but not centred on 0.5. Instead the AUC of the randomised simulations averages to 0.55 and 95 percentile value of 0.58. This distribution simply reflects that, for any given random draw, there will be spurious associations between X and D and therefore even when $P(D) \perp X$ is known to be true, the AUC is normally greater than 0.5.
- Of most importance is the Matched AUC that represents the performance of the PSM in selecting a comparison group that is observationally the same as the participant group. Reassuringly, the

distribution of Matched AUC closely matches that of the Randomised baseline, with the Matched AUC mean being similar to the RCT AUC at 0.55.

Figure 4: AUC distribution for randomised, matched, and eligible groups for all EA interventions with a PSM comparison group



Using a classical hypothesis testing approach. For each individual PSM cohort, we define that the balance test fails if the PSM AUC is greater than the 95th percentile of the equivalent RCT AUC distribution for each PSM cohort. In other words, if the PSM AUC is less than the 95th percentile, we conclude it lies within the expected distribution of AUC where $P(D) \perp X$ is true. In the CBA, we only include interventions that have passed this balance test.

Propensity score-matched historical comparison group

Interventions covered: Youth Service (YP), Youth Service (YPP).

For two EA interventions (Youth Service Youth Payment and Youth Service Young Parent Payment) there was no contemporary non-participant population. Instead, the analysis constructed a propensity-matched comparison group based on a similar population who received the Independent Youth Benefit in the past (McLeod, Dixon, & Crichton, 2016). The comparison group resembled the Youth Service participants on average but were exposed to different policies and labour market conditions.

Natural experiments

Interventions covered: Jobseeker Work Ready 52-week benefit reapplication, WRK4U.

Natural experiments are instances where an EA intervention is introduced in such a way that we have a natural comparison group. The key assumption of natural experiments is that the introduction of an EA intervention is unrelated to differences in future outcomes between participants and comparisons in the absence of the intervention or, if any differences do exist, they can be controlled for. For example, in the current EA report, we used a natural experiment to evaluate the impact of the 52-week reapplication process on exits from a benefit and how soon affected people returned to a benefit. We used information on the behaviour of job seekers in the years before the introduction of the 52-week reapplication process to provide a baseline comparison for those affected by the new policy. Because the policy was introduced nationally, we had to include labour market measures into the analysis to help control for changes in labour market conditions before and after the introduction of the 52-week reapplication process (MSD, 2013).

Likewise, we evaluated the impact of the WRK4U seminar by comparing the behaviour of job seekers in three trial sites before and after the intervention as well as the behaviour of job seekers in non-trial sites before and after the intervention (de Boer, 2003).

Other considerations

Making multiple statistical inferences

When presenting summative statements about the effects of many EA interventions on these outcomes on different subgroups of participants, we are making hundreds of statistical inferences at a time. There is a chance that some of these inferences are incorrect. Specifically, we are worried about claiming that an impact exists when there is none.

For example, imagine an EA intervention that had no impacts on any outcomes in the general population. If we took a sample of participants and comparison group members and analysed 100 impact results on different outcomes for that intervention, we would expect some of these to be statistically significant due to sampling variability. This is known in the statistical literature as the multiple comparisons problem.

At this stage we have not made an adjustment to the impact estimates for making multiple impact estimates but plan to include this adjustment in subsequent reports.

Interpretation of EA impacts in the context of multiple interventions

As the analysis makes clear, Service Delivery runs many different types of employment assistance interventions. Moreover, an individual may receive one or more interventions over time. Therefore, it is important to understand what an impact estimate for an individual EA intervention is telling you.

We are estimating the impact of participating in an EA intervention at a point in time. Participations in EA interventions occur over time, and a person may participate in only one or a series of either the same or different interventions. When we estimate the impact of an EA intervention, we are looking at a single event, namely the date a person starts an EA intervention. We are comparing this to a similar person who did not participate in the intervention on that date. Anything that happens after this date is regarded as an outcome, including subsequent participation in EA interventions. For example, an EA intervention may well increase the probability of participants receiving additional assistance relative to the comparison group. This is interpreted as an impact of the initial EA intervention. But this also means that the impact on longer-term outcomes, such as employment or income is a combination of the initial EA intervention as well as subsequent assistance. Currently, we do not have reliable techniques to try and disentangle these effects.

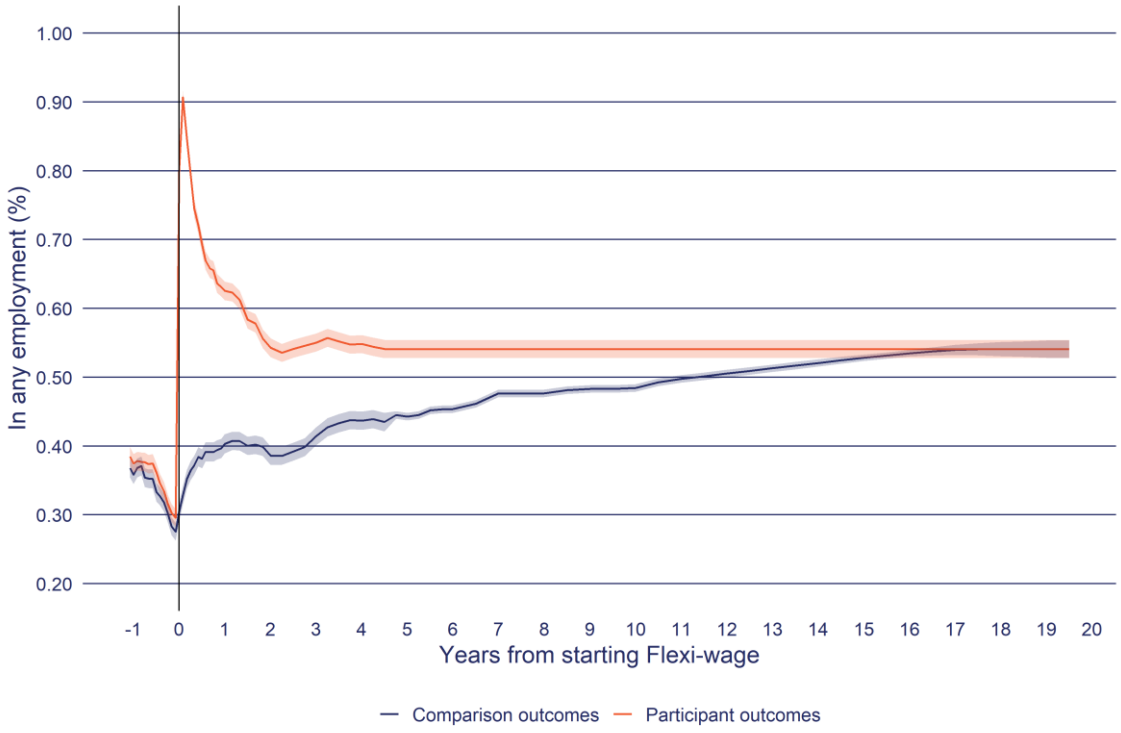
A further point to make about the comparison group, other than the participation selection period, is that we do not exclude any comparison group member who subsequently goes on to receive the EA intervention being evaluated after the selection period.[5] More generally, the comparison group will also receive other types of EA assistance over the outcome period. Therefore, when we report an impact, it is the marginal effect of the EA intervention relative to the average level of assistance received by the comparison group. This is an important point to keep in mind, as in some instances, a specific intervention may appear to have no impact because the comparison group is receiving similar levels of assistance or near proxies. These issues point to the need to carefully examine the relative experience of the comparison group against the participants to properly interpret the observed impacts.

[5] This is usually a calendar year. So, we identify all participants in an EA intervention in a given year (say 2018) and define everyone who did not participate in the intervention in 2018 as non-participants.

Estimating future impact from observed impact

In general, the period that we can observe outcomes over is shorter than the period that an intervention has an impact on participants' outcomes. Also, EA interventions often have negative short-term impacts, such as lock-in effects,[6] while positive impacts tend to occur over the medium to long term. Taken together, if we judge EA intervention effectiveness over a too short follow-up period, we are more likely to rate the intervention as ineffective by including short-term negative impacts and failing to include potential long-term positive impacts.

Figure 5: Observed outcomes for participants of Flexi-wage 2018 and the matched comparison group



a. The shaded area around the line indicates the 95% confidence interval of the outcome estimate.
b. In any employment: Employment is based on tax data (PAYE and annual tax returns). Periods with less than $100 of real (at report year) employment income per month are excluded. Annual returns are left censored to lapse period 0 if they start before the lapse period 0 calendar date.

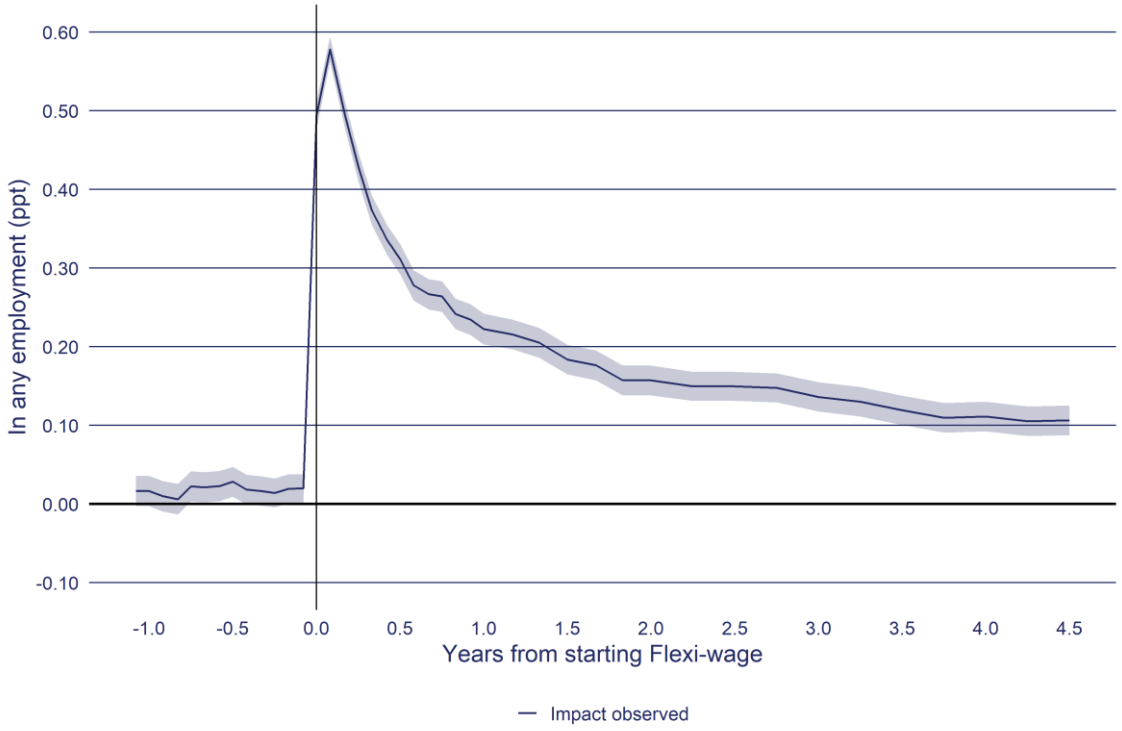
[6] Lock-in refers to the phenomenon that, while on the intervention, participants are less likely to move into employment than the comparison group. As a result, when participants finish an intervention, their average time on the benefit is longer than that of the comparison group. Therefore, if the intervention increases their employment prospects at completion it still takes time after completion before the intervention shows a cumulative positive impact.

c. Dollar values are CPI adjusted to June 2024 values.

Source: Statistics New Zealand, Integrated Data Infrastructure, June 2024.

Figure 5 shows this problem of unobserved future impacts for a hiring wage subsidy programme called Flexi-wage on the time spent in any employment. In Figure 5 we can see that nearly all participants are in employment at one month after starting Flexi-wage. But over time this proportion begins to fall. At the same time, we can see the proportion of comparison group members in employment increase over the follow up period. The difference between the two groups' outcomes is the estimated impact of Flexi-wage 2018 on participants' outcomes and is shown in Figure 6.

Figure 6: Observed impact of Flexi-wage 2018



a. Impact is the difference in outcomes between the participant and comparison group.
b. the shaded area around the line indicates the 95% confidence interval of the impact estimate.
c. In any employment: Employment is based on tax data (PAYE and annual tax returns). Periods with less than $100 of real (at report year) employment income per month are excluded. Annual returns are left censored to lapse period 0 if they start before the lapse period 0 calendar date.
d. Dollar values are CPI adjusted to June 2024 values.

Source: Statistics New Zealand, Integrated Data Infrastructure, June 2024.

Looking at the last data point in Figure 6 (4.5 years), we can see that the interval impact remains greater than zero (impact: 11±2.0 ppt). What this tells us is that we have not seen the full impact of the intervention on time spent in any employment. The full impact of the programme will not occur until the interval impact converges to zero and remains at zero indefinitely. In other words, when the proportion of participants and matched comparison group in employment is the same.

The challenge in this analysis is to estimate the unobserved interval impact to be able to get an estimate of the full cumulative impact on participants' outcomes. We do this using a three-step process:

1. Based on the earliest cohort for the intervention, we project the interval impact until it converges on zero. If the natural trend is away from zero or constant, we force the trend towards zero over the long term.[7] This referred to as the archetype impact projection.
2. Using the archetype impact projection, we estimate the expected future interval impact for later cohorts by adjusting for any differences between the later cohorts' interval impact and that of the archetype.
3. Using the observed and projected interval impacts (calculated in step 2), we accumulate the interval impacts to arrive at our total cumulative impact estimates.

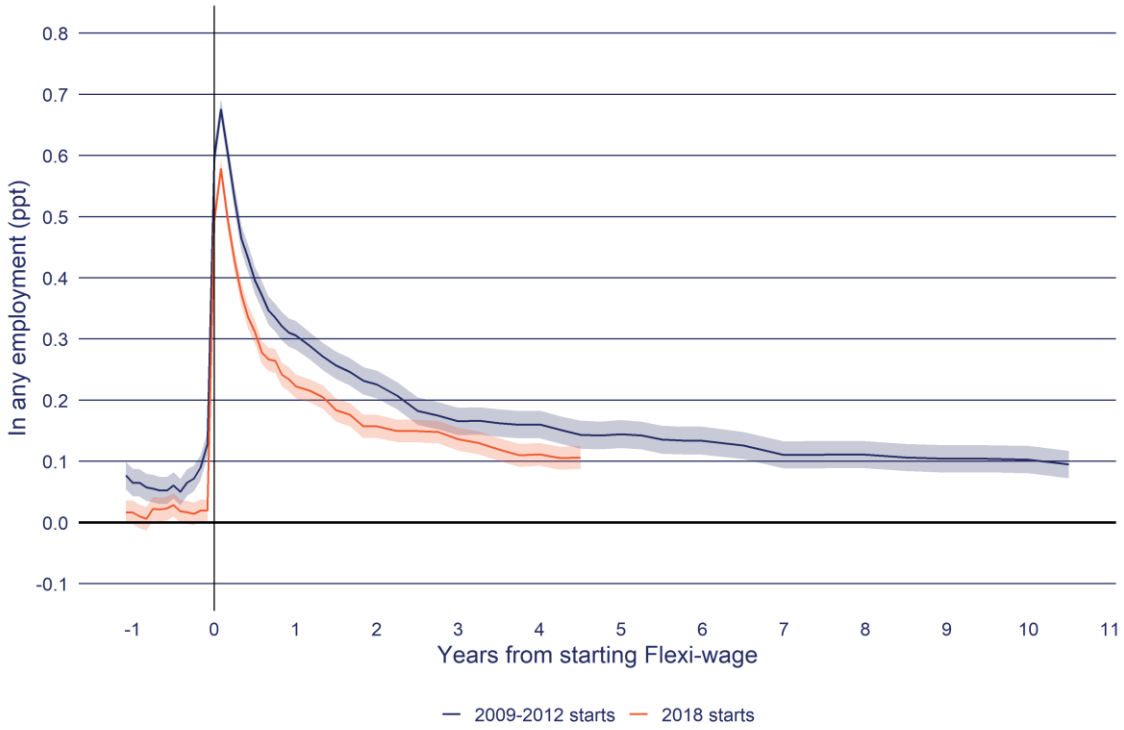
Below is a more detailed outline of each of the above steps.

Step 1: Archetype interval impact

The first step is to estimate the trend in the interval impact for the earliest cohort available and for whom we can observe the longest outcome period. This cohort is usually the first cohort for a given intervention. In the case of Flexi-wage, the intervention started in 2012 so we can follow the impact on participants for 10.5 years (Figure 7). [Could consider a weighted average for all cohorts, but this can produce odd shifts in the trend as earlier cohorts drop out at longer lapse periods]

[7] The interval impact will always converge to zero over the very long term as all members of both groups will eventually die (ignoring potential intergenerational effects). The policy question is whether this conversion occurs earlier than this point. Interventions that have persistent long-term impacts (eg continue for more than 10 years) would, on balance, generate a larger overall cumulative impact than those with short term impacts (eg impacts lasting less than one year). At present we have not included second order effects such as intergenerational impacts in our impact estimates, once these have been estimated, we may need to examine how to project long term effects for these people. However, because of the relatively high discount rate used (discussed later), any long or very-long term impacts have little bearing on the overall CBA result.

Figure 7: Observed interval impact of Flexi-wage (Basic/Plus) 2012 and 2017



a. The shaded area around the line indicates the 95% confidence interval of the impact estimate.
b. In any employment: Employment is based on tax data (PAYE and annual tax returns). Periods with less than $100 of real (at report year) employment income per month are excluded. Annual returns are left censored to lapse period 0 if they start before the lapse period 0 calendar date.

Source: Statistics New Zealand, Integrated Data Infrastructure, June 2024.

However, even using earlier cohorts as a guide to future impact trend does not resolve the problem of unobserved impacts for these earlier cohorts. In the case of Flexi-wage, there continues to be a residual interval impact at the end of the observation period of 10.5 years. Here we need to project the expected trend in interval impacts beyond the observed window. We make this projection based on the trend in the interval impact over the previous 12 observations using an Ordinary Least Squares (OLS) regression model:

$$I_t = \alpha + \beta\ t + \varepsilon$$

Where I_t is the interval impact at a given lapse period (t), usually measured in months. The $\beta.t$ provides the linear trend of the interval impact and α is the intercept term. If the trend ($\beta.t$) is towards zero then the linear trend is allowed to reach zero and the projection ends at this lapse period. If the trend is away from zero an arbitrary decelerator is applied to force the trend to zero over time. The decelerator is applied as follows:

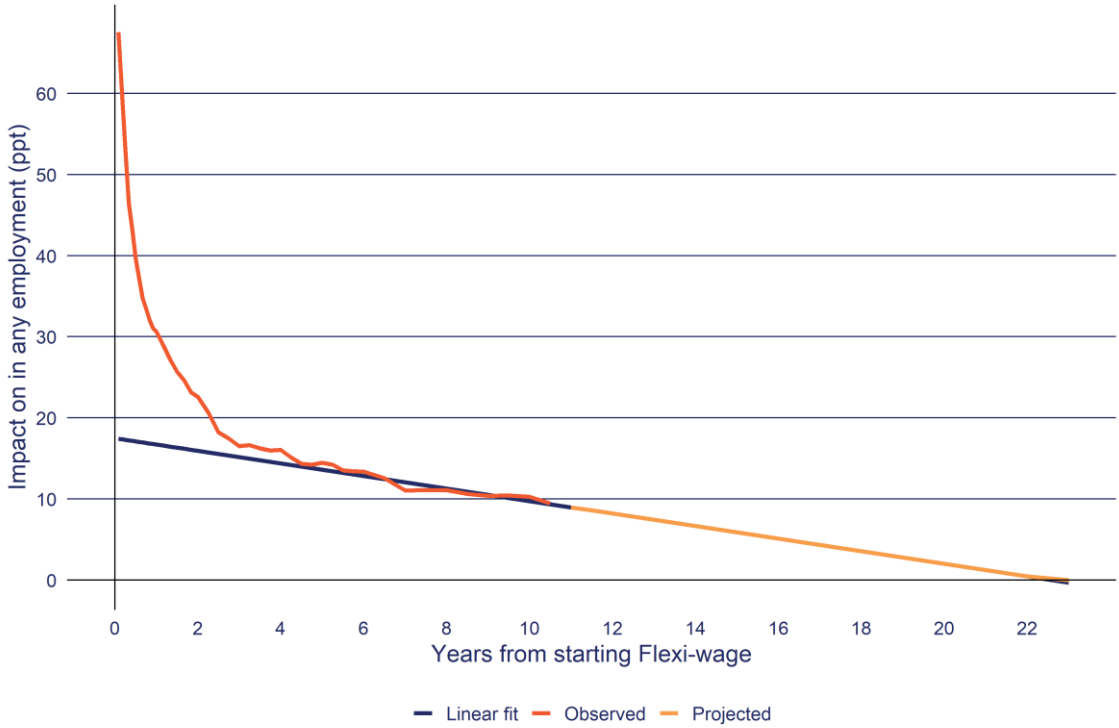
$$\hat{I}_t = (\hat{\alpha} + \hat{\beta}.t)\ (max(0, (1 - 0.005.t_{projection})$$

Where $\hat{I}_t$ is the projected interval impact based on the fitted parameter estimates $(\hat{\alpha}, \hat{\beta})$ from the OLS regression model, the decelerator is the duration of the projection period $(t_{projection})$ multiplied by a constant of 0.005. This constant is a judgement of ensuring that the projected interval impact does not overwhelm the observed interval impacts, balanced with the decelerator not being so strong that the projection period becomes too short to be meaningful.

Using an arbitrary constant is not ideal and it would be preferable to use an empirically based value. We plan to look at alternative approaches in later updates to this framework.

In the case of Flexi-wage the linear trend in the interval impact was negative (Figure 8), and therefore the projected impact is simply a linear trend to zero.

Figure 8: Projected impact for Flexi-wage (Basic/Plus) on time in employment



a. The interval impacts have been converted into daily rates.
b. In any employment: Employment is based on tax data (PAYE and annual tax returns). Periods with less than $100 of real (at report year) employment income per month are excluded. Annual returns are left censored to lapse period 0 if they start before the lapse period 0 calendar date.

Source: Statistics New Zealand, Integrated Data Infrastructure, June 2024.

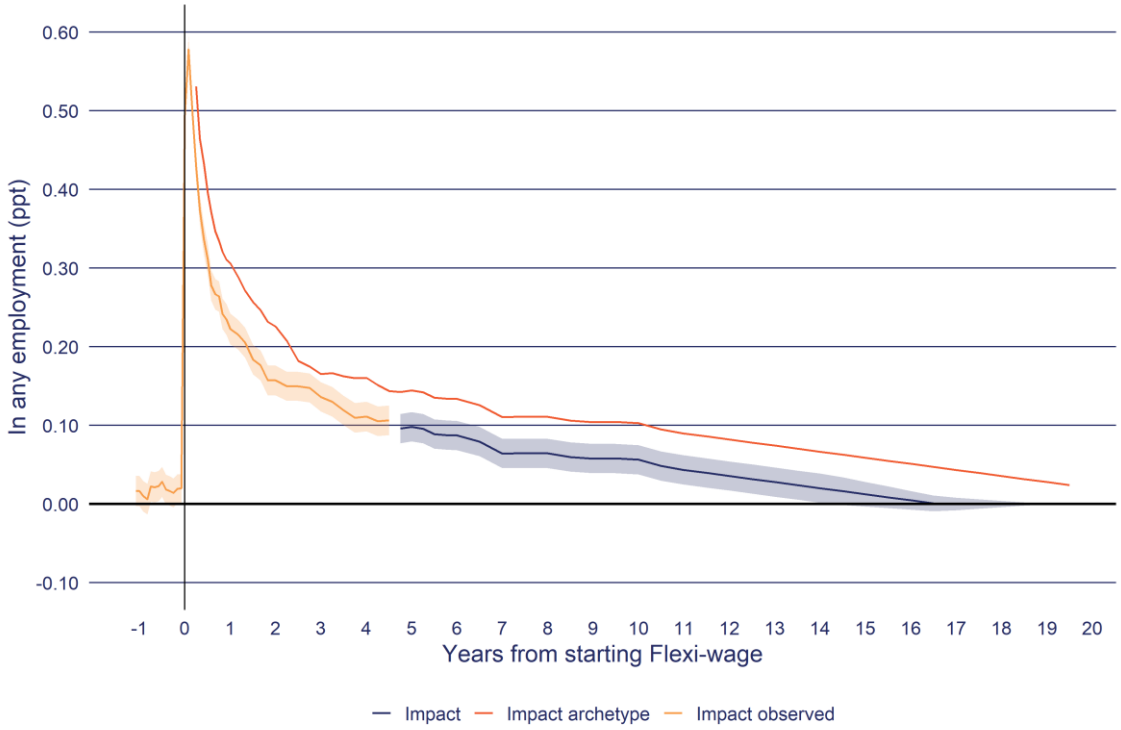
Step 2: Project observed interval impact

Based on the archetype impact described in the previous step, the next stage is to use the archetype impact as the basis for projecting the impact of each of the more recent cohorts for the intervention. Figure 9 shows the observed and projected impact for Flexi-wage 2017 and the archetype impact based on the impact for the 2012 participants as shown in Figure 7.

Scaling interval impacts

For each EA intervention cohort, we compare the last 12 observed interval impacts to archetype interval impacts, calculate the difference between the two and adjust the archetype impact accordingly. For example, if a particular EA intervention cohort is showing lower observed impacts than the archetype then the adjustment would be negative as shown in Figure 9. Currently the adjustment is a simple mean of the difference between the observed impact and the archetype impact for last 12 observed lapse intervals. The adjustment value is added to the archetype impact to shift the archetype higher or lower to match the observed impact over the last 12 lapse periods.

Figure 9: Observed and projected interval impact of Flexi-wage 2012



a. The shaded area around the line indicates the 95% confidence interval of the impact estimate.
b. In any employment: Employment is based on tax data (PAYE and annual tax returns). Periods with less than $100 of real (at report year) employment income per month are excluded. Annual returns are left censored to lapse period 0 if they start before the lapse period 0 calendar date.

Source: Statistics New Zealand, Integrated Data Infrastructure, June 2024.

The confidence interval for projected impact

To provide an estimate of the confidence interval for the projected cumulative impact requires accounting for two sources of uncertainty:

- the observed impact has a given intrinsic level of uncertainty
- the projected interval impact is itself also an estimate with some level of uncertainty.

In the current analysis, we only include the uncertainty from the first source. We plan to look at including the uncertainty introduced through the projection process itself in later updates. Therefore, the confidence intervals for the projected impact currently understate the true uncertainty for these estimates.

To reflect the confidence intervals for the observed impact on the projected impact we used Monte Carlo simulations by taking random draws from the observed impact distribution and running the projected impact calculation for each draw. We repeated these simulations 1,000 times and took the 2.5 and 97.5 percentiles as the 95th confidence interval for the projected impacts.

Rating the effectiveness of interventions

This section outlines how we systematically rate the effectiveness of interventions based on their impacts on outcomes. The aim of providing a rating is to qualitatively summarise the effectiveness of an EA intervention separately from its quantitative impacts. The goal here is to ensure that all EA interventions are rated in the same way and that the rating process is transparent.

Rating by outcome domain

For each EA intervention, we have one outcome measure grouped under each broad outcome domain. In the current effectiveness report, we focus on five outcome domains: income, employment, justice, educational qualifications, and independence from welfare.

At present, we select one outcome measure to provide the summative assessment for the impact of each EA intervention on that domain. In the current analysis:

- income effectiveness is based on the EA intervention's impact on net income from all sources
- employment effectiveness is based on the impact on any time in employment
- justice is the time that participants spend in correctional services
- qualifications is the increase in average NQF level
- independence from welfare assistance is based on time spent independent from Work and Income Assistance (ie not on main benefit or participating in EA interventions).

Translating impact to an effectiveness rating

For each outcome, we examine the observed and projected cumulative impact and categorise intervention effectiveness as shown in Table 6. In our analysis, we start with an initial assessment based on the observed impact and then adjust this assessment based on the projected impact. The higher weight given to the observed period is because it has an empirical basis, while the projected impact is sensitive to the most recent trend in the observed impact. The projected impact serves to moderate the observed impact in those instances where the two differ (ie in the off-diagonal cells in Table 6). For example, if an intervention has a significant negative observed

impact and a significant positive projected impact, we only increase the rating from negative to likely negative, rather than to promising. In practice, the majority of observed and projected impacts are consistent with each other in terms of sign (ie they are either both positive or both negative).

Table 6: Rating of outcome domain by the impact on outcomes

		Projected impact		
		Significant positive	**Zero**	**Significant negative**
Observed impact	**Significant positive**	Effective	Effective	Promising
	Zero	Promising	No difference	Likely negative
	Significant negative	Likely negative	Negative	Negative

Rating the overall effectiveness of an intervention

Once we have an effectiveness rating for each outcome domain we then combine these ratings to arrive at an overall rating of a programme. Because we are combining five outcome domains, the number of combinations of results becomes much greater. We use the following steps to determine an intervention's overall effectiveness.

1. Convert outcome domain impacts into numerical values: positive = 1, likely positive = 0.5, no difference = 0, likely negative = 0.5 and negative = 1.
2. Sum positive and negative impacts separately. If sum of negative effects equals zero then rate effective if sum of positive effects is greater or equal to 1, else if equal to 0.5 then promising. Conversely if sum of positive effects equals zero then rate negative if sum of negative effects is greater or equal to 1, else if equal to 0.5 then likely negative. If both sum to zero, then the rating is 'no difference'.
3. If sum of positive and negative effects both exceed zero, then assigned mixed if both negative and positive effect sums exceed one. Else adjust to promising if sum of negative equals 0.5 or likely negative if sum of positive equals 0.5.
4. If the rating is negative, likely negative or mixed and the outcome period is less than two years, then the rating is 'Too soon to rate'. If

the outcome period is less than one year, then the rating is always 'Too soon to rate'.

5. Not feasible is a manual override where the assessment is that the current method is not sufficiently robust to rate the intervention's effectiveness.

Based on these rules, the definition of each of the effectiveness ratings is as follows.

- **Effective**: EA interventions are rated effective only if they are effective against the majority of outcome domains and they show no sign of having a negative impact on any other outcome domain. We do not wait two years before rating a programme as effective.
- **Promising**: promising interventions are those that are effective or likely effective for at least one outcome and show no negative effects.
- **Mixed**: mixed covers interventions that show both positive and negative effects across outcome domains. We wait until we have two years of outcome data before rating a programme as mixed.
- **Makes no difference**: includes all EA interventions that have no effect on any outcome domain. We wait until we have two years of outcome data before rating an intervention as making no difference.
- **Likely negative**: interventions are in this group because either a minority of outcome domains are rated as negative with the remainder having no impact. Or, the majority are negative, with a minority having the possibility of being positive. We wait until we have two years of outcome data before rating an intervention as likely negative.
- **Negative**: interventions where most outcome domains are rated as negative. We wait until we have two years of outcome data before rating an intervention negatively.
- **Too soon to rate**: except for interventions rated as effective or promising, interventions with less than two years of observed impacts are rated as too soon to rate. The reason for waiting at least two years is that most EA interventions have negative effects in the short-term (eg lock-in effects) and it is necessary to wait sometime after commencement before positive effects are potentially observed.
- **Not feasible**: several interventions have been identified as not currently feasible to estimate their effectiveness.

References

Crichton, S. (2013) The Impact of Further Education on the Employment Outcomes of Beneficiaries, Ministry of Business, Innovation & Employment, Wellington (File ref:A7337408)link

de Boer, M. & Ku, B. (2018) Effectiveness of the Limited Service Volunteer programme: Financial Year 2014/2015, Ministry of Social Development, Wellington (File ref:A8915021)link

de Boer, M. & Ku, B. (2019) Cost-effectiveness of intensive case management services (from October 2012 to July 2017): Evaluation report, Ministry of Social Development, Wellingtonlink

de Boer, M. & Ku, B. (2021) Effectiveness of MSD employment assistance: Technical report for 2019/2020 financial year, Ministry of Social Development, Wellington

de Boer, M. & Ku, B., (2017) Service Delivery Cost Allocation Model for Individual Outputs: 2017 version, Ministry of Social Development, Wellington (File ref:A9317887)

de Boer, M. (2003) Impact of Jump Start seminar on uptake of the unemployment benefit, Ministry of Social Development, Wellington (File ref:A177933)

Dixon, S. & Crichton, S. (2016) Evaluation of the Impact of the Youth Service: NEET programme, The Treasury, Wellington (File ref:A9299141)link

Fyfe, C., Mar D., and Taptiklis, P. (2023) COVID-19 Wage Subsidy: Outcome Evaluation - Value for Money, Motu Working Paper 23-04, Motu, Wellingtonlink

McLeod, K., Dixon, S. & Crichton, S. (2016) Evaluation of the Impact of the Youth Service: Youth Payment and Young Parent Payment, The Treasury, Wellington (File ref:A9299143)link

MSD (2013) Impact of the 52-week Unemployment Benefit Reapplication Process Update 2: Technical Report, Ministry of Social Development, Wellington (File ref:A6516526)link

MSD (2018) In-work Support service and In-Work Payment trial, Ministry of Social Development, Wellington (File ref:A10173601)