August 2023

Measuring early intervention effectiveness: Principles, methods and examples

Paper 3

Prepared for the Victorian Department of Treasury & Finance

Centre for Evidence and Implementation

Authors

Dr Vanessa Rose, Director, Centre for Evidence and Implementation // Dr Robyn Mildon, Executive Director, Centre for Evidence and Implementation

Suggested citation

Rose, V. & Mildon, R. (2022). *Measuring early intervention effectiveness: principles, methods and examples.* Paper 3 prepared for the Victorian Department of Treasury & Finance. Centre for Evidence and Implementation.

About CEI

The Centre for Evidence and Implementation (CEI) is a global, not-for-profit evidence intermediary dedicated to using the best evidence in practice and policy to improve the lives of children, families, and communities facing adversity. Established in Australia in late 2015, CEI is a multi-disciplinary team across four offices in Singapore, Melbourne, Sydney and London. We work with our clients, including policymakers, governments, practitioners, program providers, organization leaders, philanthropists and funders in three key areas of work:

- Understand the evidence base
- Develop methods and processes to put the evidence into practice
- Trial, test and evaluate policies and programs to drive more effective decisions and deliver better outcomes

Table of Contents

Executive Summary

This paper outlines different methodological designs appropriate to evaluating the outcomes of children, adults, and families participating in early interventions funded through the Early Intervention Investment Framework (EIIF). In brief, the paper outlines four groups of methodological designs that measure causal impact and can be used for evaluating outcomes through the EIIF.

Experimental designs have randomisation (i.e., the random assignment of individuals or service sites to intervention or control groups) as their defining characteristic. These designs, such as Randomised Controlled Trials, are methodologically strong for determining causation and are a good option for departments where existing data systems are available, technical support is available in-house and there is a 'standard care' service which can act as a 'control', ensuring that participants randomised to that condition receive a service that meets their needs.

Quasi-experimental designs, often used in evaluation of government services because of the availability of administrative data, also examine causation but they do not involve randomisation. These designs involve the construction of comparison groups using routinely collected administrative data, through for example, the statistical matching of individuals receiving a service to individuals who are eligible but not receiving the service. Quasi-experimental designs are well-matched to measuring the impact of EIIF initiatives because they are pragmatic, feasible, and make good use of existing resources. They are best used by departments who have large, reliable data systems.

Non-experimental designs (e.g., pre-post designs, repeated measures, or longitudinal studies) are commonly used by departments to measure and report on service outcomes. These designs are less desirable than experimental designs for measuring EIIF impact because attribution (i.e., whether the change in outcomes can be attributed to the EIIF project) is limited. They are however the default when an initiative is offered to all who are eligible at the same time (and therefore no randomisation can occur) and high-quality data

is not readily available to establish a counterfactual. While non-experimental designs are less rigorous, and overall confidence in attributing service user outcomes to the initiative is reduced, some assessment of causal attribution should still occur.

Hybrid designs are becoming more popular as the role of context is better understood in impact measurement. These designs enable departments to measure, and learn about, initiative implementation and participant outcomes at the same time. Hybrid designs are agnostic in terms of what kinds of designs are used to measure effectiveness (i.e., they may be randomised, quasi-randomised or non-randomised), and the approach used to measure implementation. They are a good option for departments who want to speed up the transfer of research insights into practice, and have a view toward scaling the initiative if successful.

Measuring early intervention effectiveness

1.1. Background

The Early Intervention Investment Framework (EIIF) is Victoria's key mechanism to achieve the shift to a more balanced service system, by intervening when people first encounter services, and addressing government spending on late intervention and acute services. The EIIF, administered by the Victorian Department of Treasury and Finance (DTF), channels investment into initiatives that offer timely intervention to Victorians to reduce and prevent acute service usage, with the aims of improving outcomes for vulnerable people and avoiding costs to the state government.

Quantifying impact and measuring the effectiveness of early intervention initiatives is central to the EIIF. Victorian departments seeking EIIF funding are required to set outcome measures and annual targets and estimate avoided costs for their initiatives in budget proposals and should factor in methods for data collection and evaluation at this early stage. EIIF funded initiatives are required to report annually internally to the government on progress against targets, which are measures designed to capture the expected impact for service users receiving the intervention (they are not accountability and performance measures in the manner reported on through other government processes).

There is more than one way to robustly quantify impact and measure the effectiveness of EIIF initiatives. Guidance on best practices in the measurement of early intervention effectiveness including practical examples of different methodological designs will support Victorian departments and service delivery partners to better understand the impact of their programs and provide robust insights to Government as part of annual EIIF reporting. This evidence base contributes to re-balancing the service system through effective early intervention.

1.1.1. This paper

DTF commissioned the Centre for Evidence and Implementation (CEI) to prepare three brief discussion papers to explore how the EIIF could be leveraged and enhanced to support reorientation of Victoria's service system so that early intervention forms a larger proportion of the system. The papers were informed by material on the EIIF supplied by DTF, key reports, papers, and journal articles identified through desktop search and expert recommendation, and consultations with a select group of senior executives in the Victorian government across finance, health, and social service portfolios.

The primary audience for the papers is policymakers, including those involved in preparing budget bids that seek funding under the EIIF. In Paper 1 we explored the features of a successful early intervention system and ways in which the EIIF could be leveraged and strengthened to support achievement of this goal in Victoria. In Paper 2 we explored how embedded data systems that enable quality outcome measurement and evaluation are a key feature of successful early intervention systems. The purpose of this paper is to provide a guide for measuring early intervention program effectiveness in the Victorian service system through:

- outlining key principles for measuring the effectiveness of early intervention programs funded by the EIIF
- identifying rigorous methods to measure program effectiveness within early intervention systems, and
- illustrating examples of methodological designs that have been used nationally to measure the effectiveness of a social service.

We note all lapsing programs that have received state funding are required to undertake an evaluation within 12 months of when the funding is due to lapse. Victoria's Resource Management Framework provides guidance for this and emphasises that evaluation planning should start at the initial stages, when developing a logical model or causal relationship between the program and outcomes to be achieved can help to determine what information will be needed for the evaluation. It is hoped this paper will prove useful in thinking about how both the impact measurement, and evaluation, of EIIF initiatives may be considered at the same time.

1.2. Key principles in measuring the effectiveness of EIIF initiatives

1.2.1. Understanding evidence

In general, evidence can be categorised within three broad domains:[1]

- evidence on aetiology and burden (i.e., the causes and nature of the problem)
- evidence on effectiveness and interventions, and
- evidence on implementation within context.

Evidence on aetiology forms part of EIIF budget proposals – analysis of the cause, nature and burden of a problem is essential to describing the problem and the potential impact of earlier intervention on people (through improved outcomes) and the system (through

[1] Brownson, R.C., Shelton, R.C., Geng, E.H. & Glasgow, R.E. (2022). Revisiting concepts of evidence in implementation science. *Implementation Science*, 17:26.

avoided costs). The evidence used to describe a problem is, however, often different to that required for measuring the effectiveness of EIIF initiatives (or how understanding these initiatives should be implemented for that matter). For example, a department interested in addressing violence against people with a disability, may first source existing quantitative evidence of the prevalence of the issue in Victoria across different groups, and qualitative evidence on the nature of the issues, or how violence is experienced by people with disability (i.e., in what situations, forms and across different groups of people with disability).

Evidence on effectiveness and interventions, or services, programs, or other innovations (i.e., evidence on 'what works') is different from that required to understand aetiology and burden. This evidence relates to evidence of the change for service users receiving the intervention, ideally compared to what would have happened had they not participated in the service. For example, the difference (if any) observed in young people's school attendance and educational outcomes following receipt of a school scholarship program that distributes funds for electronic devices, books, and clothes. Demonstrating evidence of the effectiveness of initiatives by identifying available data, supplementing data, and establishing data collection, analysis, and reporting processes, is central to the EIIF. Some of the most commonly used and applicable methodologies for evaluating early interventions are discussed later in this paper.

Evidence on implementation within context is critical to ensure the service, program, intervention, or innovation is received and used by the target group (e.g., Do young people know about the scholarship or how to access it? Are the vouchers provided linked to local stores? Do electronic devices integrate with the school environment?). Poorly implemented initiatives are less likely to be effective and demonstrate the desired impact than well implemented initiatives. We include an example of a methodological design, a hybrid design, that integrates measurement of effectiveness and implementation and generates insights during rollout of the innovation for course correction if required. Measuring implementation allows departments the ability to contextualise the outcomes achieved (e.g., demand has been slower than forecast because of the department's delayed roll-out of the new data reporting infrastructure, unrelated to the program). It is important to evaluate cost effectiveness and other fundamentals of broader government evaluations as part of an assessment of the impact of the initiative. Guidelines for evaluation fundamentals can be found in the Resource Management Framework.[2]

In short, different types of methodological designs and evidence are needed to answer different questions. Questions about the impact of EIIF initiatives on participant outcomes require evidence on effectiveness. This is best achieved through impact measurement approaches that enable an assessment of causal attribution. While evaluation designs vary – ranging from the most rigorous, best-practice designs (which are desirable but can be highly resource intensive) through to more commonly used designs – it is important that EIIF-funded initiatives consider causality as part of their approach to impact measurement and include this in their planning.

1.2.2. Causal attribution

Impact measurement is the process of measuring and describing the changes that occur in outcomes *as a result of the EIIF program*. Measuring impact, which is in essence a way of conceptualising cause and effect, sits at the core of EIIF reporting. That is, observed changes in parenting outcomes, from program intake to program exit, can be causally

[2] Victorian Department of Treasury and Finance. (2022). Resource Management Framework. Available at: https://www.dtf.vic.gov.au/planning-budgeting-and-financial-reporting-frameworks/resource-management-framework

linked (or attributed) to the parenting program. This is known as causal attribution. In general, there are three different ways of thinking about causal attribution:[3]

- Sole causal attribution - where a program is necessary and sufficient in itself to produce the desired change in outcomes, independent of contextual factors (i.e., the program works in the lab)
- Joint causal attribution – where a program produces a desired change in outcomes in conjunction with other interventions and/or contextual factors (i.e., the program works in the real world but only in certain contexts, such as when a housing stability program only demonstrates outcomes when people are also receiving mental health support)
- Alternative (or multiple) casual paths - where a program is just one of several ways to produce a desired change in outcomes (i.e., the program works in the real world, but it is unclear whether it is the program or other services or factors that achieved the impact). The outcome school attendance, for example, may be influenced by several means including service supports focussed on a child getting to school, feeling engaged at school, or even through housing stability.

Both joint and alternative path causal attribution are important in measuring the impact (and interpreting the measurement of impact) of EIIF funded initiatives. Strategies for inferring causal attribution include:

- Estimating a counterfactual, or what would have happened without the EIIF program
- Assessing the consistency of evidence for causal relationships with that predicted in the theory of change, and
- Systematically ruling out alternative explanations or causal pathways.

1.2.3. Counterfactuals

Causal attribution in impact measurement is often associated with a counterfactual – i.e., an estimate of what would have happened in the absence of a program. For example, would there have been the same observed reduction in court delays without implementation of the 12-month court support program? This can only be observed using a counterfactual design where program participants (e.g., people who attend court and receive specialist legal support and assistance as part of the program) are compared to non-participants (e.g., people who attend court and receive the standard support and services available to them but not the specialist program).

Counterfactuals represent the most rigorous approach to impact measurement for demonstrating improvement in EIIF initiative outcomes. They produce the strongest evidence that an initiative has been effective. For this reason, a counterfactual should be considered during EIIF initiative design and options for creating a control group or robust comparison group ideally canvassed in the proposal for funding through EIIF. Various methodological designs that use a control or comparison group counterfactual are explained below, and real-world examples of these designs in practice are presented in Table 1. Common ways for constructing counterfactuals are described in the sections below. In brief, these include:

[3] Rogers, P. (2014). Overview: Strategies for Causal Attribution. *Methodological Briefs: Impact Evaluation 6*, UNICEF Office of Research, Florence.

- Randomised controlled trials, where participants are assigned randomly to receive an initiative or usual care
- Quasi-experimental designs where a comparison group is constructed administratively
- Statistically created counterfactuals, and
- Logically constructed counterfactuals.

Every effort should be made to identify and create a control or comparison group, although we recognise it may not always be possible to establish counterfactual designs, because, for example:

- Alternative services do not exist to address the problem the initiative is trying to address, meaning that a randomised counterfactual condition (where people in need were denied a service) would be unethical
- A suitable comparison group is not available, through for example, a historical baseline, and
- Data is not available to establish counterfactual conditions using statistical matching or other approaches.

In one example provided in Table 1, a counterfactual was unable to be established because of ethical concerns in randomisation and limitations with the database (i.e., data on the specified age group had not been routinely, or accurately collected, previously). In this case, the attribution of impact on children's risk of homelessness to the program could be hypothesised only, through the planned and systematic analysis of cohorts and integration of qualitative data.

1.2.4. Context and complexity

Context matters when it comes to applying and generating evidence. Programs that have demonstrated effectiveness with one cohort, or in one country, may not be able to translate this impact elsewhere. For example, the Nurse-Family Partnership model, where trained nurses regularly visit first-time parents in the home antenatally and postnatally, had been shown to be effective in improving children's outcomes in the US, but failed when first implemented in the UK health system, which has high standards of universal care.[4] In essence, UK parents were already receiving the 'intervention'. Best practice application of evidence involves adaptation to context. Any EIIF proposals seeking funding for implementation of 'evidence-based programs', such as modularised parenting program, or similar should include a description of how the program will be adapted to meet the needs, values, and preferences of, for example, Aboriginal children and families. Capturing experiences of program accessibility, acceptability and cultural safety is particularly important in this context.

That evidence is specific to context has other implications. All EIIF funded programs are implemented in complex and adaptative service systems. These properties include:[5]

[4] Robling, M. et al. (2015). Effectiveness of a nurse-led intensive home-visitation programme for first-time teenage mothers (Building Blocks): a pragmatic randomised controlled trial. *Lancet,* 387: 146-155

[5] Skivington, K. et al. (2021). A new framework for developing and evaluating complex interventions: update of Medical Research Council guidance. *BMJ*, 374: n2061

- Emergence – complex systems have emergent, often unanticipated properties (e.g., at-risk young people attending a group program develop social relationships which reinforce risky behaviour)
- Feedback – one change reinforces, promotes, balances or diminishes another (e.g., a smoking ban in public places reduces visibility of smoking, which leads to fewer people taking up smoking and further reduced visibility)
- Adaptation – change in system behaviour in response to a program (e.g., retailers adapt to alcohol policy by marketing products differently), and
- Self-organisation – order arising from spontaneous local interaction (e.g., a group of people who are unemployed establish a work collective).

Understanding the properties of complex and adaptive systems is critical to inferring causal attribution and interpreting the outcomes of EIIF initiatives. This may have as much meaning when a program, service or innovation is effective (i.e., because it helps to understand why it worked, and what is necessary for it to continue working) as it is when an EIIF initiative does not meet targets (i.e., what is the context in which the initiative was implemented and what could be improved or changed?). Complex systems also means there are likely multiple initiatives targeted to the same service user or cohort at the same time (e.g., separate housing, income support, dental and mental health services, across sectors and federal-state jurisdictions, for an individual experiencing homelessness). This further points to the value of a well-constructed counterfactual condition in demonstrating the impact of EIIF initiatives.

1.3. Methodological approaches to measuring the effectiveness of EIIF initiatives

1.3.1. Establishing baselines to measure effectiveness

Paper 2 canvassed different approaches to establishing baselines for early interventions, including using historical baselines (i.e., outcomes and targets based on an earlier administration of the initiative) or, where necessary, using baselines from similar initiatives (i.e., same or similar intervention, target cohort, outcomes and implementation context). This section focuses exclusively on methodological approaches that require departments, in consultation with sector and service delivery partners, to establish their own EIIF initiative baselines using primary (i.e., data collected from scratch to measure initiative outcomes) or secondary data (i.e., data routinely collected for another purpose that can be used to measure initiative out comes).

Baselines are a collection of data describing the current situation or status quo, which can be used as a fixed reference point against which to monitor and assess changes in service users' outcomes observed since the initiative began. In the ideal situation (and depending on data availability), baselines will be established for up to six outcome measures nominated in the EIIF budget proposals. Not all baselines will however need to be established in the same way. For example, an initiative to reduce offending among at-risk young people may include six outcomes that require measurement using different data from different sources. The primary or long-term outcome of interest is reduced offending, measured using administrative crime data held by the Department of Justice and Community Safety and Victoria Police. The initiative involves implementation of an intensive family support intervention program provided within the home and classroom. This means that EIIF initiative outcomes may also measure, for example, parenting competence using a self-report questionnaire and school attendance using data sourced at the school level. This is not just outcome data gathered from different sources, but also

potentially using different methodological designs, which have different levels of causal attribution.

It may be difficult at EIIF budget proposal stage to establish a baseline, particularly if the initiative is new and it is not possible to rely on administrative data. In this case, it may be necessary to identify a baseline at project start, which could be for example some six to 12 months following a funding decision. It is important that baselines are set as soon as feasibly possible, but equally important not to push ahead too early with baselining if the initiative is not yet appropriately implemented. For example, a baseline survey of parents whose children were eligible to receive a new early childhood service was undertaken while negotiations were still occurring between social service agencies and early childhood centres about how the initiative would be implemented. This meant a delay of more than six months (a significant time developmentally), which necessitated a second baseline be undertaken. Delays between baseline and the start of the initiative raise a risk that factors other than the EIIF initiative, such as developmental maturation, may influence service user outcomes.

Baselines established using historical data for the purposes of the EIIF budget proposal may need to be revisited after funding is received to ensure these baselines are accurate in the current context. This might occur if there has been a significant change to policy affecting the characteristics of people accessing the EIIF funded initiative, such as a change in the definition of a 'risk of harm' report which results in changes to the eligibility of parents to receive a service. Further, other factors such as implementation maturity may also influence when a baseline should be established. Young people who are a 'later intake' to a diversionary service may benefit from a more streamlined service with confident and wizened practitioners compared with 'earlier intake' participants. Depending on the initiative, it may be sensible to delay baselining for up to three months or more.

Data maturity and capability across the Victorian public service is growing. Many departments have highly proficient analysis and research units that can provide advice on constructing baselines and assist with access to linked data and other measurement options. Service providers and DTF can provide further assistance through leveraging learning over the past three years of the EIIF and sharing effective practices. Yet some gaps remain. Even initiatives within the same sector may differ greatly in terms of data system maturity and capability, often because of varied levels of investment over time. Building reliable data systems that measure outcomes and can inform practice and policy, takes time, focused investment, and adequate resourcing, not just in infrastructure, but in departmental capability. Levers and enablers for the EIIF, central government and departments in building capacity for impact measurement are discussed in earlier papers in this series.

1.3.2. Common methodological designs

The objective of the EIIF is to link investment to quantifiable impacts, measured in terms of service user outcomes against expected outcomes. At a fundamental level, this means departments must be able to measure and set baselines, prior to the initiative being implemented and/or as soon as practical following implementation, and report annually as the initiative is being delivered. Baseline and outcomes reporting will be dependent on the type of outcome measure selected. For example, lead outcomes (i.e., outcomes that are present in the short-term, such as parenting competence or emergency department presentations) should be measured and reported immediately. Lag outcomes (i.e., outcomes that are longer-term, such as child development and wellbeing) can be reported later, as per the theory of change, as long as baselines have been set. Administrative data

and linked data can have delays of six months (or sometimes more), and this should be factored into baseline setting and outcome reporting.

Annual EIIF reporting will necessarily differ according to the nature of the early intervention initiative. For example, initiatives may vary according to the size and scope of the service (e.g., an intensive intervention for a small number of families with multiple risk factors and service interactions across sectors versus a population-level alcohol use reduction campaign) and the length of service engagement (e.g., three-months versus 12-months). This means departments will need to define populations and cohorts very clearly in measurement and reporting, including any contextual factors that may influence reporting (e.g., small initial intake sample in year one during set-up means statistical significance from baseline to follow-up is limited). In an example of a therapeutic drug court initiative where participants graduate to a less intensive type of service when their outcomes improve, a department could establish a separate outcome to measure this cohort (e.g., improved wellbeing among participants offered the 'step-down' service compared with a matched cohort with similar characteristics).

Whether it has been expressly considered during an earlier planning phase or not, reporting on an initiative always entails a methodological design choice of some kind. For example, reporting change in service user outcomes before a new service model is implemented (i.e., at baseline) and after the service has been in operation for one-year or more is a 'pre-post' (non-experimental) design. Methodological designs differ according to the degree in which causality can be inferred (i.e., is it possible or highly likely that the chronic disease navigator initiative produced the observed reduction in hospital admissions for people with type 2 diabetes?). The confidence with which departments and DTF will be able to attribute any observed change in outcomes over time to the operation of the EIIF initiative is dependent on the type of methodological design used.

Methodological designs commonly applied to impact measurement of health and human services across Australian governments are described below. These designs can be used to measure EIIF initiative service user outcomes, which vary in terms of rigor (or ability to confidently attribute cause and effect), cost and the technical capacity required to conduct them.[6]

Experimental designs

Experimental designs have randomisation (i.e., the random assignment of individuals or service sites to intervention or control groups) as their defining characteristic. Randomised Controlled Trials (RCT) are the best-known examples of experimental designs. These designs involve randomly assigning, for example, at-risk young people eligible for homelessness prevention services to either a pilot mentoring program, undertaken while receiving standard care case management support (i.e., the intervention) or standard care case management support (i.e., the control group). Random assignment results in intervention and control groups that do not systematically differ (i.e., at-risk young people with different 'characteristics' such as age, gender, cultural background and sexual orientation are spread evenly across the two groups). This means any observed difference in outcomes (such as reductions in homelessness) following the pilot program can be causally attributed to the program itself.

This is a methodologically strong design, but departments should consider:

[6] Lynn, J., Stachowiak, S. & Coffman, J. (2021). Lost causal: Debunking myths about causal analysis in philanthropy. *The Foundation Review, 13,* https://doi.org/10.9707/1944-5660.1576

- Random assignment is not ethical if there is no 'standard care' service or condition that can act as a control
- RCTs can be expensive to set-up, particularly if the measurement of outcomes is solely reliant on the collection of primary data or surveys, although this can be minimised by the use of secondary data or administrative data already collected
- RCTs require adherence to international protocols in the design, conduct and reporting of studies which means they are reliant on expert guidance for implementation.

Randomised controlled trials do not have to be costly, overly-technical or invasive if cleverly designed. This is particularly the case for 'light-touch' interventions which might be expected to demonstrate a small but critical impact across a population. For example, electronic social housing tenant management administrative systems have been used to randomly allocate invitations to a fortnightly savings scheme to prevent 'negative exits' from social housing to homelessness resulting from rental arrears. This application of an RCT was cheap to administer because it included only minor amendments to an existing administrative system, allowed clear unbiased randomisation, and drew on expertise in data and content within the department.

Another experimental design that holds promise for measuring impact of EIIF initiatives is a stepped-wedge design.[7] The benefit of a stepped-wedge design is that it addresses the constraints within which government and program owners operate and mimics how social services are implemented across a state. For example, a new program to safely restore children in the child welfare system to carers may be rolled out by the department site by site over a 12-month period. Using a stepped-wedge design, baseline data is collected for all children across all sites from the start date, and monthly (or another determined interval of time) until the entire program is in place. Each child across sites, experiences the 'control condition' (i.e., standard care) and the 'treatment condition' (i.e., restoration intervention), essentially enabling children and sites to 'act as their own controls'. Requirements for a stepped-wedge design include that roll-out occur in an orderly, spaced or 'stepped' manner (e.g., a site begins implementation every month) and data collection occur over time at regular intervals. This can be challenging given the intensity of data collection, even if using administrative data, the relative lack of specialist experimental trial capability (different from quantitative or statistical skills) within government, and the need to manage roll-out in an environment where the priority is to act quickly. This is likely why this design has not been used frequently to measure impact within government. In the right circumstances however, a stepped wedge design is a good match to social services and could be usefully employed in the measurement of impact for an EIIF funded initiative.

Randomisation can also be undertaken at the site or cluster level (i.e., cluster randomised controlled trials or cluster stepped-wedge designs) because it may make more sense, administratively and ethically, to offer the same innovation to all children at one site rather than, for example, half the children at one site. These designs require large sample sizes with sufficient power to detect differences between the intervention and control group[8] and analysts with specialist statistical skills.

Quasi-experimental designs

Quasi-experimental designs, often used in evaluation of government services because of the availability of administrative data, also examine causation but they do not involve

[7] Hooper, R. (2021). Key concepts in clinical epidemiology: Stepped wedge trials. *Journal of Clinical Epidemiology, 137*: 159-162.

[8] White, H., & S. Sabarwal (2014). Quasi-experimental Design and Methods, *Methodological Briefs: Impact Evaluation 8*, UNICEF Office of Research, Florence.

randomisation. These designs identify and use a comparison group that is as similar as possible to the intervention group (i.e., participants in the new service, program, or innovation) at baseline, allowing comparison between service users who receive the EIIF initiative and those who receive business as usual services. For example, an initiative measuring the impact of an intensive wrap-around support program for children at-risk of contact with the criminal justice system, in using a quasi-experimental design, will seek to identify a comparison group of children at baseline who have the same characteristics as those receiving the program (i.e., same age, gender identification, cultural background, household income etc). The comparison group functions as a counterfactual or what would have happened to children's outcomes had the program not been implemented. This means the program can be said to have caused any observed differences in outcomes.

Quasi-experimental designs are well-matched to measuring the impact of EIIF initiatives because they are pragmatic and feasible, and make good use of existing resources, such as routinely collected administrative data. There are two commonly used methods for constructing comparison groups using administrative data, both of which attempt to minimise bias in measuring impact:

- Propensity score matching, and
- Regression discontinuity design.

Propensity score matching is the statistical process of matching individuals in the intervention group (e.g., students in schools with an active nurse-led mental health program) with the comparison group (e.g., students in schools without this program) based on an analysis of factors that influence their propensity to participate in the program. The process should be undertaken at baseline to allow an exhaustive examination of characteristics that affect participation, and ensure the matching characteristics are different to what the program purports to impact (e.g., school retention, attendance, wellbeing). Propensity score matching requires a large sample size, particularly in the comparison group, because many observations will be discarded until a statistical match is made. Data for intervention and comparison groups can come from different sources as long as the data are defined the same way and are collected at the same time. Or data can come from the same source but at different time-periods, such as in the case of a matched historical comparison group where a contemporary comparison group cannot be identified. This can occur when an EIIF initiative is measuring the impact of a whole of system reform on service user outcomes (i.e., all users receive the new service) and the only available counterfactual is what outcomes were achieved prior to implementation of the new service.

Regression discontinuity designs are used when there is a criterion or threshold for inclusion in a program, such as when children who score at a threshold level on a mental health screener are eligible for the school nurse program. The approach uses statistics to determine the margin around the threshold (i.e., children who score 15 or 16 on a screener where the threshold for entry to the program is 17) and compare the outcomes of children just above and below the cut-off point. Regression discontinuity designs require a large sample size and data collected on the 'threshold measure' and outcome for the entire sample (i.e., intervention and comparison groups).

A common approach to the analysis of data collected through a quasi-experimental design to measure impact is the Difference-in-Differences (DID) method. In essence, the mean difference in days absent from school, for example, between children who participated in the school nurse program and those who did not is calculated and compared. This approach assumes that the outcomes of interest follow the same trajectory, and that

children in the compassion schools, for example, are not subject to a new local policy enforcing school attendance that changes the trajectory of outcomes.

Quasi-experimental designs minimise the chances that the intervention and comparison groups systematically differ from each other, but they are not perfect, and differences may still exist at baseline because of an inability to achieve a good match (although differences can be accounted for, to some extent, in statistical models). Another quasi-experimental design that may prove useful for departments in measuring the impact of EIIF initiatives on service user outcomes is an interrupted time series.[9] In this design, data are collected at multiple and equally spaced time points (e.g., weekly, monthly, or yearly) before and after implementation of an intervention. The main objective of an interrupted time series design is to examine whether the data pattern observed post-intervention is different to that observed pre-intervention.

To make use of these designs in EIIF, departments should have access to administrative data systems that are of high quality (i.e., data is valid and reliable, demonstrated through limited missing data, well defined data dictionaries) and staff with strong statistical skills and training in impact measurement that understand and have competence in using the dataset.

Non-experimental designs

Non-experimental designs represent are commonly used by departments to measure and report on program outcomes. These designs include pre-post designs (i.e., measured at baseline and once at the end of the intervention), repeated measures (i.e., measured at baseline and several times over the course of the intervention and sometimes following), and longitudinal studies (i.e., measured at baseline and across several years, most often used in child development). Non-experimental designs are less desirable than experimental designs for measuring EIIF impact because attribution (i.e., whether the change in outcomes can be attributed to the EIIF project) is limited. They are however the default when an initiative is offered to all who are eligible at the same time (and therefore no randomisation can occur) and high-quality data is not readily available to establish a counterfactual.

Like experimental and quasi-experimental designs, non-experimental designs are interested in inferring cause and effect relationships, but they do not have counterfactuals or comparison groups. While this makes them less rigorous than experimental and quasi-experimental designs, and overall confidence in attributing service user outcomes to the intervention is reduced, assessment of causal attribution should still occur. This should take the form of assessing the consistency of evidence with a causal relationship through, for example:[2,10]

- Aligning the pattern of results, across outcome measures, with the theory of change, both in terms of what is impacted (i.e., whether all participants achieved the intermediate outcome) and when it is impacted (e.g., a long-term outcome occurs with the specified time-frame)
- Identifying 'dose-response patterns', if appropriate (e.g., does parenting competence increase with the number of group sessions attended?)

[9] Ewusie, J.E. et al. (2020). Methods, Applications and Challenges in the Analysis of Interrupted Time Series Data: A Scoping Review. *Journal of Multidisciplinary Healthcare, 13*:411-423

[10] Dickerman, B. A. & Hernan, M.A. (2020). Counterfactual prediction is not only for causal inference. *European Journal of Epidemiology, 35*, 615-617 doi:10.1007/s10654-020-00659-8.

- Checking the consistency of results with empirical literature (i.e., studies that have used robust methodological designs), including identifying and reporting results that are inconsistent, and/or
- Undertaking interviews with service users and key informants that focuses on their own understandings of causal processes and explanations for outcomes.

Hybrid designs

Hybrid designs, which mesh effectiveness and implementation[11], are becoming more popular as the role of context is better understood in impact measurement. Hybrid designs were developed to bridge the research to practice gap by measuring, and learning about, initiative implementation and participant outcomes at the same time – meaning that programs benefit from not only learning 'what works' but what needs to happen to achieve 'what works'. This type of design provides substantial benefits over traditional, independent process and outcome evaluations by generating actionable insights and facilitating practical adoption of effective implementation strategies. Other methodological designs described above can incorporate a focus on rapid evaluation and improvement, particularly if the approach includes a mixed-methods approach; it is the integration of effectiveness and implementation in a hybrid design that allows users to generate insights in the context of delivery, pin-point and tailor implementation while ascertaining whether outcomes are attributable to an initiative.

There are three types of hybrid design:[12]

- Type 1 – which has a primary aim of determining the effectiveness of an intervention and a secondary aim of better understanding the context for implementation
- Type 2 - which has a co-primary aim of determining the effectiveness of an intervention and a co-primary aim of determining feasibility and/or potential impact of an implementation strategy, and
- Type 3 - which has a primary aim of determining the impact of an implementation strategy and a secondary aim of assessing outcomes associated with the intervention.

Type 1 and Type 2 designs are most appropriate in the context of measuring impact through EIIF because an assessment of the effectiveness of the initiative on service user outcomes is primary. Hybrid designs are agnostic in terms of what kinds of designs are used to measure effectiveness (i.e., they may be randomised, quasi-randomised or non-randomised), and the approach used to measure implementation (e.g., different implementation strategies of varying costs randomised across sites, or a single strategy adapted to site context). While measuring implementation is not explicitly required for EIIF funded initiatives, considering both outcomes and implementation through a hybrid-evaluation can provide a holistic view of a program's effectiveness with a line of sight to what is working and why. As an example, a pilot program for families with children at risk of harm may be funded by EIIF for implementation across two regions. The responsible department decides to measure impact using a Hybrid Type 1 design, where the effectiveness of the program in preserving children with families is tested using a matched quasi-experimental design with administrative data (and outcomes are reported to EIIF annually). At the same time as measuring effectiveness, the department measures implementation factors important to understanding scalability (e.g., costs, fidelity and

[11] Curran, G.M. et al (2012). Effectiveness-implementation hybrid designs. *Medical Care, 50*:217-226.
[12] Landes, S.J., McBain, S.A. & Curran, G.M. (2019). An introduction to Effectiveness-Implementation Hybrid Designs. *Psychiatry Research, 280*:112513

adaptation, reach and acceptability, delivery setting and workforce, and implementation infrastructure).[13]

1.3.3. Examples of methodological designs

In Table 1 we outline four methodological designs discussed above that have been applied in measuring the impact of health, education and human services initiatives funded by governments in Australia and internationally. We present an overview of the initiative, the data and measures available, the challenges experienced in creating a counterfactual (the ideal design) and the final methodological design employed to measure impact and why it was selected. These examples have been specifically selected to be relevant to the context of departments implementing, and measuring, the impact of EIIF funded initiatives on service user outcomes.

Table 1: Examples of methodological designs to measure impact

Initiative example	What data and measures were available?	What challenges were experienced in identifying a counterfactual?	What design was used to measure impact and why?
Mentoring program to reduce risk of homelessness among 12- to 15-year-olds in contact with the child protection system	• Administrative data – linked child protection and homelessness services system interactions • Caseworker-reported wellbeing outcomes tool	• Highly vulnerable group meant all eligible children received the program (in the absence of suitable 'standard care') • Inability to create a valid historical comparison group because of database limitations related to the routine collection of data for that age group • Wellbeing outcomes tool newly rolled-out and only administered in this program	Non-experimental (pre-post) design, supplemented with qualitative data, and analysis of cohorts to assist with future program improvement and targeting (i.e., did children with limited child protection interactions fare differently from those who had been 'in the system' for some time?) This design was selected because it addressed practical exploratory questions about the feasibility of the service, service targeting and children's outcomes
Postnatal home visiting program to improve child developmental outcomes among families on a low-income	• No administrative data was available for use in impact measurement • Health and development survey comprising standardised instruments was developed and administered over 3 timepoints	• Need for accurate, up to date birth lists for recruitment and random allocation • Staff resourcing for collection of primary data across both intervention and control groups • Ensuring any incentives/supports for participation administered by the program throughout COVID-19 were the same across intervention and control groups (so impact reflected the home visiting program only)	Experimental design, randomised controlled trial with blinded allocation to intervention and control group This design was selected in conjunction with government stakeholders in the context of an existing usual care condition, so that families randomized to the control group received a service that addressed need

[13] Milat, A. et al (2020). Intervention Scalability Assessment Tool: A decision support tool for health policy makers and implementers. *Health Research Policy and Systems, 18*:1.

Initiative example	What data and measures were available?	What challenges were experienced in identifying a counterfactual?	What design was used to measure impact and why?
Whole of child protection system reform to improve children's permanency	• Administrative data – linked child protection, education and health data	• Whole of system reform, and nature of the service (i.e., statutory child protection) meant no opportunity for randomisation • Need to measure the impact of the program across different cohorts (i.e., family preservation, entry/re-entry to care, and ongoing-care) • Inability to use contemporaneous comparison groups across all cohorts • Variable quality of data across cohorts and data fields resulting from poor completion	Quasi-experimental design, across three cohorts with matched comparison groups: • Children eligible for family preservation package but did not receive one • Children who entered a new episode of foster or kinship care in a historical time-period • Children who were in foster or kinship care in a historical time-period This design was selected because it made ready use of available and reliable administrative and linked data across an entire cohort, allowing creation of multiple cohorts and counterfactuals to measure different service components and outcomes
Early childhood educator's program to improve children's social, emotional and cognitive development	• No administrative data on these outcomes was available for use in impact measurement • A mix of quantitative measures (self-report and observational) were used to measure outcomes at baseline, 6-months and 12-months • A mix of quantitative data (e.g. use of online teaching resources) and qualitative (e.g. focus groups) were used to measure implementation at 12-months	• Difficulties recruiting Early Childhood Centre control sites, some of whom withdrew from the program following randomisation to the control group • Need to minimise data collection burden with time-poor educators • Need for timely data collection across 3 time-points and 12 centres (intervention and control) requiring dedicated resourcing and project management	Type 2 hybrid implementation-effectiveness design using a cluster randomized controlled trial This design was selected because it enabled simultaneous evaluation of outcomes and the effectiveness of program implementation, including differences in outcomes related to implementation fidelity