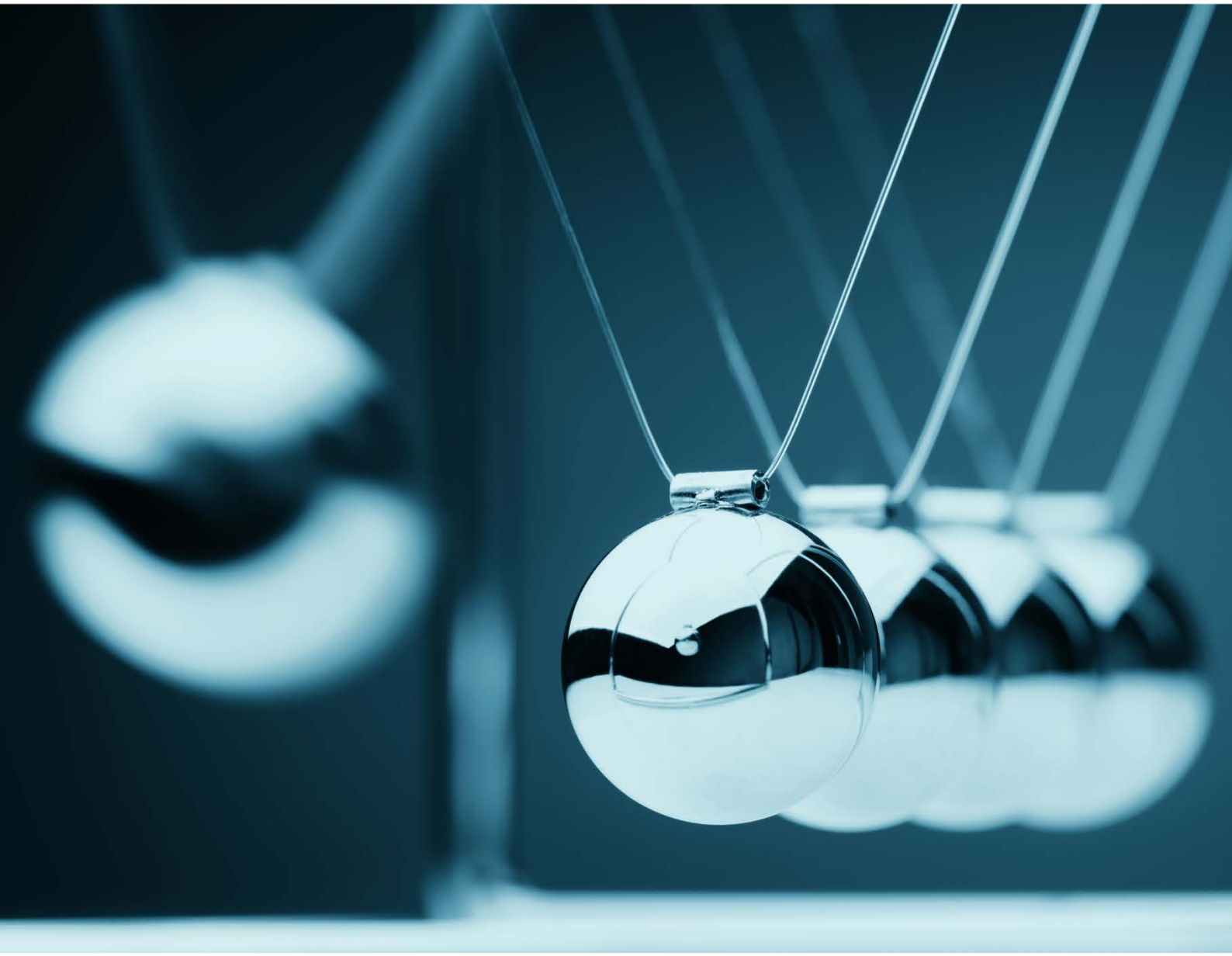
Australian Government
Department of Industry, Innovation and Science

Office of the Chief Economist



Choosing appropriate designs and methods for impact evaluation

2015

WWW.INDUSTRY.GOV.AU/OCE

ARTD CONSULTANTS
strategy & evaluation

RMIT UNIVERSITY

Australian Government
Department of Industry, Innovation and Science

Office of the Chief Economist

Choosing appropriate designs and methods for impact evaluation

Patricia Rogers, Andrew Hawkins,
Bron McDonald, Alice Macfarlan & Chris Milne

November 2015

ARTD CONSULTANTS
strategy & evaluation

RMIT UNIVERSITY

Acknowledgments

This work was completed with the assistance of Andrew Lalor, Louise Talbot, Mark Johnson, Luke Hendrickson, Bilal Rafi, Katherine Barnes, Claire Bramwell and Martine Rodgers of the Department of Industry, Innovation and Science.

We would also like to thank the many interviewees from across the Department for their time and insights, and trust that their views are adequately represented in this report.

The report draws extensively on material on impact evaluation methods from the BetterEvaluation site www.betterevaluation.org.

ARTD/ RMIT Consultancy team

Professor Patricia Rogers, Andrew Hawkins, Bron McDonald, Alice Macfarlan & Chris Milne

ARTD Consultants www.artd.com.au

RMIT University www.rmit.edu.au / BetterEvaluation www.betterevaluation.org

Contents

Tables

Figures

Executive summary

Increasingly, people working in government and public policy are debating the best methods for evaluating government policies and programmes. How can we get the evidence we need to assess the impact of government investment in a way that is both transparent and defensible? How can we determine what works, in what circumstances and why, to the benefit of current and future policies and programmes?

Impact evaluation seeks to determine the longer term results that are generated by policy decisions, often through interventions, projects or programmes. Impacts may be positive or negative, intended or unintended, direct or indirect.

The choice of methods and designs for impact evaluation of policies and programmes in industry, innovation and science is not straightforward, and comes with a unique set of challenges. Policies and programmes may depend on contributions from other agencies and other actors, or take many years to emerge. Measuring direct cause and effect can be difficult.

The Department of Industry, Innovation and Science has commissioned this report to explore the challenges and document a range of possible approaches for the impact evaluations that the department conducts. Research for the project comprised interviews with key internal stakeholders to understand their needs, and a review of the literature on impact evaluation, especially in the industry, innovation and science context. That research led directly to the development of this guide. This research project is the first stage of a larger project to develop materials as the basis for building departmental capability in impact evaluation.

There is not one right way to conduct an impact evaluation. What is needed is a combination of methods and designs that suit the particular situation. When choosing these methods and designs, three issues need to be taken into account: the available resources and constraints; the nature of what is being evaluated; and the intended use of the evaluation.

In terms of the first issue, resources and constraints, the range of possible methods and designs is dependent on the availability of existing data, internal knowledge and expertise, and funding to engage external evaluators and undertake additional data collection and analysis. Other important constraints include the time available before findings will be needed to inform decisions.

The nature of what is being evaluated is the second issue to consider. The way impact evaluation should be done depends in part on: whether the way the programme works is well understood (from previous research and evaluation) or still being developed; whether impacts can be easily observed within a short time frame or only many years later; whether the programme activities are standardised and pre-specified or adaptive; whether the programme is the sole factor producing impacts or works in conjunction with other programmes and other factors. It is useful to develop and use a programme logic of the intervention which specifies how its activities contribute to a chain of intermediate outcomes that produce the intended impacts. In addition to helping to choose impact evaluation methods, this helps to identify gaps in logic or evidence that the evaluation should focus on, and provides the structure for a narrative about the value and impact of a programme.

The third issue to consider is the intended use of the evaluation. This influences the types of questions that are asked, the timing for findings, what will be considered credible evidence – and even whether an impact evaluation is appropriate. In some cases other forms of monitoring or evaluation might be more useful or cost-effective than impact evaluation for informing particular decisions.

Impact evaluation is commonly undertaken for one of four main purposes:

- **Advocacy—**demonstrating the value of investment in a particular programme or portfolio
- **Allocation**—informing how funding will be allocated across potential programmes, including *ex ante* impact evaluation (done before an intervention is funded to estimate likely impacts) and *ex post* impact evaluation (done after implementation to inform decisions about whether or not to continue or scale up)
- **Analysis**—learning what is working to inform continuous improvement including providing information about how to effectively continue or scale up
- **Accountability**—effective risk management.

An impact evaluation involves three different types of questions—descriptive (the way things are or were), causal (how the programme has caused these things to change) and evaluative (overall value judgement of the merit or worth of the changes brought about). In any impact evaluation, a combination of different methods is needed to answer these different types of questions. Like any evaluation, impact evaluation will generally be most reliable and valid when it uses a mixed methods approach where results from one method can be used to test or extend those of another.

Descriptive questions ask about how things are and what has happened. These can include the initial situation and how it has changed, the activities of the intervention and other related programmes or policies, and the environment for implementation. Methods and designs for answering descriptive questions need to address how to sample units of analysis, use appropriate measures and indicators, ensure adequate response rates to questionnaires and interviews, and gather data about hard-to-measure changes.

Causal questions ask whether or not, and to what extent, the intervention being evaluated brought about the observed changes. Increasingly, causal analysis is about understanding how an intervention *contributes to* impacts, along with other factors and other programmes.

Evaluative questions ask about the overall value of a programme or policies, taking into account intended and unintended impacts, the criteria and standards that have been established upfront, and how the different criteria should be weighted and synthesised. A programme that is effective in terms of meeting its objectives might not be judged a success if it also produced large negative impacts or if the impacts were concentrated on sectors that were not the priority focus. On the other hand, in a context of worsening economic circumstances, a programme might be judged successful in terms of reducing a decline in employment even if it has not met its original targets. To answer evaluative questions, methods can include reviewing formal statements of values and articulating unspoken values (especially among diverse partners), negotiating

different values, and synthesising information into an overall judgement of success.

Economic evaluation (such as cost-benefit and cost-effectiveness analyses) adds an extra dimension to evaluative questions by answering questions about the overall value of a programme or policy, taking into account its cost. It combines evidence from an impact evaluation and data about costs. To generate results that accurately compare the costs and benefits of different programmes, consistency in assumptions and measures is essential.

In impact evaluations of DIIS programmes, answering causal questions presents the biggest technical challenges. This stems from the nature of the programmes themselves and the complex systems in which they intervene. They may be undergoing development or be implemented differently in different and constantly changing contexts, with multiple stakeholders involved and long-time lags before intended impacts may be observed.

Some guidance on impact evaluation argues that the most rigorous method for answering causal questions is a randomised controlled trial (RCT) , where individuals, organisations or sites are randomly assigned to either receive a 'treatment' (participate in a programme) or not (in some cases receiving nothing, and in others receiving the current programme), and changes are compared. However RCTs are not always possible or appropriate. To conduct an RCT requires that the evaluators can define the intervention in such a way that what was tested could be reproduced; that they can undertake and maintain random allocation into treatment and control groups; and that the sample size is sufficient to detect differences between treatment and control groups (given the expected strength of the intervention and the length of the casual chains). Meeting these conditions may limit the application of RCTs to components of a programme rather than to an entire programme.

This report therefore also discusses the use of quasi-experimental methods, such as propensity score matching and regression discontinuity, and non-experimental methods, such as contribution analysis and process tracing, to answer causal questions.

The fact that there is the debate around the best methods for impact evaluation points to **social and political considerations** in planning, conducting and using impact evaluation. Impact evaluation is a negotiated process between stakeholders with varied views, interests and power, within specific organisational settings and political environments.

These social, ethical and political considerations for impact evaluation are not generally tied to a particular method, but to overall approaches to commissioning, conducting and concluding an evaluation and communicating findings.

Perhaps the most serious issue for Commonwealth-funded industry programmes is how the choice of evaluation method can influence programme selection and design. The danger is twofold. The first danger is that only relatively simple programmes are developed. There is a risk that programmes become less ambitious when they place a greater emphasis on measuring outcomes. The second danger is that promising programmes are not valued simply because their results cannot be measured, while relatively ineffective programmes are valued because aspects of them can be more easily measured.

Economists will often be asked to conduct a cost-benefit analysis to determine which programme delivers the best value for money. Programmes whose outcomes are not readily measurable will suffer. Understanding the context in which the results were achieved (the why) is also important.

Impact evaluations that provide simple answers are easy to communicate, but can oversimplify the situation. In this case the results may not have external validity—outcomes achieved in the past may not occur in the future because how and when the programme worked was not adequately understood. This underlines the importance of impact evaluations that acknowledge complexities and variations across contexts.

The choice of appropriate designs and methods for impact evaluation will necessarily involve social, ethical and political considerations. The proposed designs of impact evaluations need to be scrutinised from this perspective and their consequences anticipated, with suitable social and political processes and ethical safeguards put in place.

1. Introduction

1.1 This research project

In recent years there has been increasing discussion about the need for better impact evaluation of government programmes and policies. Impact evaluation can inform policy and programme design and implementation as well as resource allocation. It can demonstrate the value of government investments and identify how to improve their value.

The Department of Industry, Innovation and Science (DIIS) facilitates globally competitive industries in Australia and supports the development of critical requirements for productivity, economic growth and scientific capability. DIIS implements a wide range of policies and programmes to provide support, promote growth, facilitate competitive marketplaces, provide regulatory frameworks, and reduce business costs, working closely with business and the scientific community.

DIIS has recognised a need to build capacity to evaluate the impacts of this work. The departmental Evaluation Unit has commissioned this research project to determine the most suitable methods of impact evaluation, as the first stage of a larger project to develop materials and build departmental capability in impact evaluation.

The research has been undertaken in a context where appropriate methods and designs for impact evaluation are keenly debated. While some organisations advocate a narrow range of methods and designs, and organise these in terms of a hierarchy of evidence, there is increasing recognition of the need for a wider range of methods and designs for comprehensive, credible and useful impact evaluation of policies and programmes. This report examines this wider array of approaches and their actual and potential use in contemporary impact evaluation in industry, innovation and science policy.

Evaluating the impact of science and industry policies and programmes comes with a unique set of problems and difficulties. A one-size-fits-all approach is not appropriate to evaluate the diverse activities and interventions implemented by the Department. Many of the policies and programmes depend on contributions from other agencies and other actors, or take many years to emerge, so it can be difficult and sometimes meaningless to talk about the direct cause and effect impact of the Department's work.

The aim of this research project is to review existing literature on impact evaluation, and to analyse the available methodologies to suit the Department. The key research question is:

> Which available impact evaluation methodologies are most suitable to assess the impact of policies and programmes in the Department? How can methods best be matched to particular kinds of policy and programmes?

This report provides a framework for choosing the most appropriate methods for impact evaluation in industry, innovation and science policy with particular application to DIIS policies and programmes. It has been based on a review of relevant theory and practice in impact evaluation in industry and science programmes and more widely, and interviews with key stakeholders in the

Department in May and June 2015. The sources for the literature review are outlined in Appendix 6.

1.2 What we mean by 'impact'

Impacts are the longer-term results produced by a programme, project or policy, usually in conjunction with other factors and activities by other agencies. They include intended and unintended results, positive and negative, direct and indirect impacts.

The Commonwealth of Australia's Resource Management Guide No. 131 (2015, p. 49) defines impact as:

> The ultimate difference made by fulfilling a purpose defined in an entity's corporate plan. Compared to the combined outcome of activities contributing to a purpose, impacts are measured over the longer term and in a broader societal context.

Impacts are broader than stated goals. For the purposes of accountability, learning, value-for-money and ethical conduct, it is important that the term 'impact' also includes unintended impacts, positive or negative.

Unintended impacts in an industry context might take the form of an *externality*, a case where a third party that is not the direct user or adopter receives a direct impact, which is often unintended (CSIRO 2014, p.46).

Reugg and Jordan (2007, p.104) describe potential unintended effects of a firm engaging in R&D in this way:

> The firm uses its new knowledge from research to produce better and/or lower cost products. The innovating firm profits and its consumers benefit by receiving more for their money or by paying less. Knowledge gets into the hands of other firms—through its intended release in papers and patents, but also in unintended ways such as by reverse engineering and worker mobility. Some of these other firms use the knowledge gained from Firm 1's research without compensation to improve their own products competing with those of Firm 1, thereby capturing some of the profit from Firm 1's innovation and driving the price down further for consumers. Some use the knowledge gained to innovate in other product markets, realizing profit from Firm 1's research and benefiting their own customer base. Spillovers result from direct commercialization by the innovator, from knowledge captured by others, and, in this example, from a combination of both. Social benefits are the sum of the gains to all producers and consumers, which […] is much larger—due to market and knowledge spillovers—than the gains realized by the firm who performed the research. For this reason, spillovers are often discussed in terms of social versus private returns.

Unintended impacts can be negative—for example, an Industry Canada evaluation of a mandatory bankruptcy counselling programme describes the following negative unintended impact:

> The audit found that rather than helping steer debtors away from declaring bankruptcy, pre-bankruptcy counselling made their situation worse because it delayed the filing for bankruptcy. By the time debtors attended pre-bankruptcy counselling, their financial standing was almost always beyond rehabilitation and bankruptcy was the only real alternative. Thus, the pre-

> bankruptcy counselling session represented an additional administrative burden rather than an opportunity for debtors to explore options and pursue alternatives (Industry Canada 2013, p. 22).

Impacts do not only refer to what has happened—in some cases, the impact is in terms of **preventing negative changes**: 'Impact also includes the reduction, avoidance or prevention of harm, risk, cost or other negative effects' (Warwick & Nolan 2014, p.9). In the industry and science context, negative changes which programmes and policies seek to prevent include job losses in a regional economy and spread of diseases. Negative changes that have been prevented are more difficult to measure since, by definition, they have not occurred.

1.3 Types of impacts

The type of impacts relevant to industry and science programs will depend upon the nature of the intervention. Impact will have different meanings depending on the ultimate objective of an intervention or programme. The intended impacts of the department's support for science and commercialisation include the development, uptake and commercialisation of innovation and technology, with the long-term objective of an improvement in Australia's productivity, competitiveness and economic growth.

It is helpful to think about impact as including, but not limited to

> an effect on, change or benefit to the activity, attitude, awareness , behaviour, capacity, opportunity, performance policy, practice, process or understanding of an audience, beneficiary, community, constituency, organisation or individuals in any geographic location whether locally, regionally, nationally or internationally (CSIRO 2014, p.vi).

Standardised lists can be used as a guide to potential impacts. However these should not be taken as comprehensive and researchers should consult other sources such as internal and external stakeholders and related research.

CSIRO has established a list of potential economic, environmental and social impacts as a starting point for considering the wider range of possible programme impacts (Appendix 1).

1.4 What we mean by 'impact evaluation'

An impact evaluation is a form of programme evaluation and needs to be planned using the classic steps of an evaluation:

- Identify primary intended users of the evaluation and their primary intended uses and involve them in the planning of the evaluation as much as possible.
- Identify relevant impacts and how they might be produced. Drawing on previous research and evaluation, key informants, and programme documentation, develop a programme theory of the intervention showing how its activities (or planned activities) are likely to generate the intended impacts, and a negative programme theory showing how it could generate negative impacts.
- Develop a short list of key evaluation questions.

- Answer these key evaluation questions, using an appropriate combination of methods and designs.
- Report findings to primary intended users and support them to use the findings.

An impact evaluation provides evidence about the impacts that have been produced (or the impacts that are expected to be produced). It has to not only provide credible evidence that changes have occurred but also undertake credible causal inference that these changes have been at least partly due to a project, programme or policy.

While terminology varies, we suggest using the following terms to clarify an important distinction:

Causal attribution is an appropriate term for this causal inference when it is possible to estimate confidently what proportion of impacts has been caused by a particular programme or policy.

Causal contribution is an appropriate term when it is only possible to be confident that a programme has been one of the contributing factors producing impacts.

Different types of impact evaluation are used before and after as well as during programme implementation

- ***Ex post* impact evaluation** gathers evidence about actual impacts.
- ***Ex ante* impact evaluation** forecasts likely impacts.
- **During implementation** gathers evidence about whether the program is on track to deliver intended impacts.

Impact evaluations differ in their overall intended use.

Formative impact evaluation is used to inform improvements to a programme or policy, particularly when there is an ongoing policy commitment.

Summative impact evaluation is done to help make decisions about beginning, continuing or expanding a programme or policy.

A summative evaluation of a closed programme may be used formatively for a new programme.

Distinctions between *ex-ante* and *ex-post* impact evaluations on the one hand, and formative and summative impact evaluations on the other, are independent of each other. For example, *ex ante* impact evaluation is usually summative but could be used formatively, to estimate likely impacts and inform a redesign of the proposed programme. *Ex post* impact evaluation can be used formatively to identify areas for improvement and elements that need to be retained, but is often used summatively to inform decisions to expand, contract or terminate a programme.

Economic evaluations combine evidence from an impact evaluation and the analysis of data about costs, primarily

- **cost-benefit** analysis which transforms all the benefits (positive impacts) and costs (resources consumed and negative impacts) into monetary terms,

taking into account discount factors over time, and produces a single figure of the ratio of benefits to costs

- **cost-effectiveness** analysis which calculates a ratio between the costs and a standardised unit of positive impacts (for example new patents, or new jobs).

Impact evaluation can include or be complemented by economic analysis, and impact evaluation can provide data on impacts for economic evaluation.

1.5 International discussions about methods for impact evaluation

This report draws on extensive international discussion about methods for impact evaluation in recent years. Some of these discussions have advocated for a narrow range of methods and approaches; this report is consistent with those who have instead advocated for situationally appropriate selection of methods and approaches from a wide repertoire.

Some of the initial debates were led by advocates for randomised controlled trials (RCTs). This is an impact evaluation design where units of analysis (individuals, organisations or communities) are randomly assigned to either receive a 'treatment' (participate in a programme) or not (in some cases receiving nothing, and in others receiving the current programme), and changes are compared. For example, the Coalition for Evidence-Based Policy in the USA (Coalition for Evidence-Based Policy 2015) only includes studies involving a well-conducted RCT when identifying 'social programs that work'.

Some approaches to either undertaking or using impact evaluations have taken a slightly broader view. For example, in the UK the 'What Works Centre for Local Economic Growth' (WWG) (set up in October 2013 as part of the 'What Works Network' to analyse which policies are most effective in supporting and increasing local economic growth) only includes studies which score 3, 4 or 5 on the Maryland Scientific Methods Scale[1]. All of these designs are based on a counterfactual approach to investigating cause and effect—developing an estimate of what would have happened in the absence of a programme or policy and comparing this to what actually happened (the factual).

The report *'Dare to measure', Evaluation designs for industrial policy in the Netherlands* (Impact Evaluation Working Group 2012) recommended specific counterfactual approaches for particular types of interventions.

A third approach to impact evaluation takes an even broader view, arguing that an even wider range of methods and designs can be credible and appropriate in particular circumstances, including using alternatives to counterfactual reasoning. These alternative approaches are particularly relevant when it is not possible to create a credible counterfactual—for example when a programme is universal, or when it is aimed at changing a system rather than individual people or organisations. This approach to impact evaluation has been promoted

[1] The Maryland Scientific Methods Scale rates the strength of evidence according to the research design (Sherman et al. 1998) from Level 1 (correlation between a program and an impact at one point in time) to Level 5 (Random assignment and analysis of comparable units to programme and comparison groups).

particularly in international development to address challenges in impact evaluation of programmes and policies that explicitly work at the system level or in conjunction with other programmes and policies (Ravallion 2009; Deaton 2010).

As an example, the International Initiative for Impact Evaluation (3ie) produced a report *Addressing attribution of cause and effect in small n impact evaluations* (White & Phillips 2012), which discussed possible application of several non-experimental designs and methods. The UK Department for International Development commissioned a report on 'Broadening the range of designs and methods for impact evaluation' (Stern et al. 2012). This in turn informed the recent report 'Impact Evaluation: A Guide for Commissioners and Managers' (Stern 2015), which set out five different bases for causal inference, in addition to using a counterfactual.

This report draws on the broader range of options for impact evaluation outlined in the international literature. The authors have made use of their previous contributions to the field. These include a presentation on 'Learning from the evidence about evidence-based policy' to the Productivity Commission Roundtable on Evidence-Based Policy Making in the Australian Federation (Rogers 2009a), and a paper on matching impact evaluation designs to the nature of the intervention and the purpose of the evaluation (Rogers 2009b), as well as guidance on choosing appropriate methods available on the BetterEvaluation website (Better Evaluation n.d.).

Box 1.1: Report Structure

The remaining chapters of this report consider the following issues:

Chapter 2: Why should we do impact evaluation? What value does it provide?

Chapter 3: Choosing designs and methods for impact evaluation

Chapter 4: How can we describe, measure and evaluate impacts? What methods suit different situations?

Chapter 5: Social, ethical and political considerations for impact evaluation.

2. Why should we do impact evaluation? What value does it provide?

This section describes four reasons for doing impact evaluation as outlined by RAND Europe (Jones et al. 2013):

- **advocacy** (demonstrating the value, or otherwise, of particular programs and of science and industry investments in general)
- **allocation** of investment (funding, staff and other resources)
- **analysis** to inform continuous improvement (including future programme design)
- **accountability** (as required under legislation and better practice performance management).

Any given impact evaluation is likely to have a combination of these reasons, although each may require different evidence and different methods of collecting it. In addition, it must be recognised that, while evidence can inform decisions and actions, other considerations, in particular policy imperatives, are also influential.

2.1 Impact evaluation for advocacy

In practice, many government departments use impact evaluation primarily for advocacy and accountability.

> The main purpose of the industrial policy evaluation system should be to provide evidence on what works to help inform future policy design and strategic economic policymaking. In some countries, however, evaluation systems tend to put more emphasis on transparency and accountability in fund allocation and expenditure rather than on lesson-learning for strategic economic policymaking and development of the national industrial and innovation system… While such systems may have value in demonstrating to external stakeholders that money is well spent, they may be less effective in ensuring that policy learning takes place (Warwick & Nolan 2014).

Impact evaluation can be used to advocate for particular programs and policies which are supported by evidence (and to reduce or stop funding for those found to not be effective) and to advocate more generally for industry and science programs and policies. It might be used to demonstrate how research and business support benefits society. It can also be used to advocate for an organization, by demonstrating its performance in terms of being a good service provider, or innovative and progressive. In an environment where access to resources is contested, impact evaluation can be an important means for advocating for an effective programme or arguing to discontinue an ineffective one.

A portfolio of impact evaluation also can be used to advocate for a policy area or an organisation. For this broader type of advocacy, it would be an advantage to show:

- the way impact evaluation has impacted on the next round of policy decisions
- collectively, how programmes have contributed to policy objectives

- the agency's capability to deliver return on investment and/or to manage and adapt high risk programmes.

Sometimes clear, statistical messages can be most effective; at other times clear stories are more useful.

2.2 Impact evaluation for allocation

Impact evaluation designed to support **allocation** helps to prioritise which projects, people and institutions are given funding.

An impact evaluation to inform future allocation of funds might cover some or all of the following:

- the impact on planned beneficiaries and those who miss out
- cost-benefit or cost-effectiveness and efficiency issues
- risks.

An impact evaluation also will need to consider:

- strategic fit of the proposed intervention with government policy commitments
- the role of government and co-investment and key stakeholders
- capability and capacity to deliver
- innovation
- whether further extension of the programme is required (and evidence for this).

An investment decision-making checklist might be useful to analyse whether the implementing agency has a reputation for delivering what it promises and managing efficiency.

2.3 Impact evaluation for analysis

Impact evaluation designed to support **analysis** is intended to inform improvements. It produces information about where and how funding has been more or less effective in producing the chain of outputs, outcomes and impacts. This is often further defined in terms appropriateness, effectiveness and efficiency. This can be used for formative purposes to make changes to programme or policy design and/or implementation.

Of all the purposes for evaluation this was one that was most frequently mentioned in interviews with DIIS staff. Managers were seeking impact evaluation to inform their decision-making processes and an opportunity to improve the impacts of their programmes. The following questions need to be asked periodically throughout the life of a programme:

- What have we achieved to date and what impacts are emerging?
- What has worked or not worked and why?
- Is there anything we should do now to improve the impact of the programme?
- If we had to start the same programme today, what would we do differently?

Box 2.1: Example of impact evaluation for analysis

An example of impact evaluation for the purpose of analysis is found in the 2010 report *Inspiring Australia: A national strategy for engagement with the sciences* (Commonwealth of Australia) which noted that previous evaluations of DIISR's terminating Science Connections Programme (SCOPE) were used to inform recommendations such as the following:

That DIISR's terminating Science Connections Programme (SCOPE) be replaced with a broader national initiative designed to increase the level of public engagement in the sciences. Such an initiative would provide ongoing support for existing, successful activities while developing innovative approaches to effectively engage a wider audience (Commonwealth of Australia 2010, p. Vii).

Analysis throughout a programme lifecycle might inform programme improvements, but it can be particularly important in designing new programmes. Programme managers might be most interested in what worked and what didn't in a similar programme or industry. Yet this information might not be readily available given political interests which can obscure evaluation findings. Past evaluations should add to the knowledge base about how programmes are effective, which may not occur if evaluation is mainly done for accountability or advocacy.

2.4 Impact evaluation for accountability

Over a period of time, impact evaluations can show a consistent pattern of competent management that reassure stakeholders that public money is being managed wisely. The Australian Government Public Management Reform Agenda, *Public Governance, Performance and Accountability Act 2013* (PGPA Act) and enhanced Commonwealth performance framework emphasise the value of 'comprehensive evaluations to provide a better understanding of the overall impact of an activity' (Department of Finance 2015).

Impact evaluation aimed at accountability needs to be clear about who is being held accountable for what (Rogers 2009a). In particular, when the causal chain from activities to impacts is not closed, and other factors affect the achievement of intended impacts, it is not reasonable to hold implementers accountable for achieving impacts. Neither is it reasonable to only hold them accountable for producing outputs. Instead it has been suggested (Perrin 2012) that managers should be held accountable for managing for impact—gathering information about actual and likely impacts, and visibly using this to improve their decisions about implementation and resource allocation.

Accountability can be a broad concept. For example, the Victorian Department of Finance and Treasury (2014) spelled out requirements for impact evaluation across the investment lifecycle and provided a range of options for providing evidence of impact, both quantitative and qualitative. Considerations included:

- whether the expected benefits of the investment have been realised
- what lessons can be learned from the project for both current and future projects, such as

- successful elements to reinforce in future processes
- aspects of the current project requiring remedy
- ways of improving the management of future projects.

These lessons were expected to be shared for effective organisational learning about investment development and project planning, procurement, implementation, and ongoing management.

By spelling these out during the *ex-ante* phase, project teams were able to build these requirements into their monitoring, research and process evaluation plans from the beginning, and also start to inform the other purposes for which impact evaluation is conducted.

For some agencies, accountability can be their main reason for evaluation. For example, in an impact evaluation of natural resource management research programs for the Australian Centre for International Agricultural Research (ACIAR), Mayne and Stern (2013) noted that

> In … ACIAR, impact assessments are undertaken primarily from an accountability perspective. Accountability driven evaluations focus on the results achieved and associated causal processes; assessing whether programs 'produced' impacts and their magnitude.

Many evaluations conducted for accountability ask investors whether their funds are being well spent and tend to focus on estimating the net economic benefits.

In complex environments, when impacts are not under the control of programmes, it is often more appropriate for evaluation that is done for the purposes of accountability to focus on whether management processes are adequate, including responding to changing circumstances and understanding what the impact has been, rather than simply checking that intended impacts have been achieved.

3. Choosing designs and methods for impact evaluation

This section presents a framework for matching impact evaluation methods and designs to programmes and policies. It addresses the overarching research question of this report:

Which available impact evaluation methodologies are most suitable to assess the impact of policies and programmes in the Department of Industry, Innovation and Science?

How can methods best be matched to particular kinds of policy and programmes?

3.1 A framework for designing impact evaluations

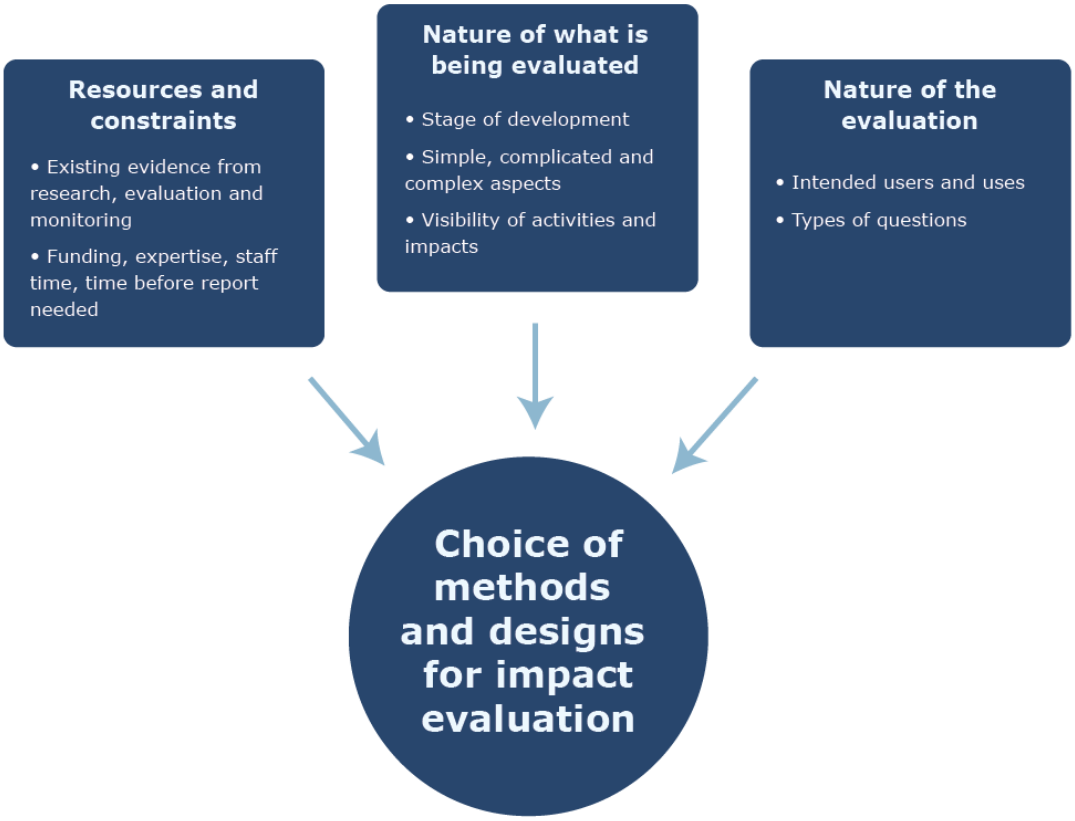
The framework we present is a variation of the 'design triangle' developed by Stern et al. (2012) for designing an evaluation. This has three elements (see Figure 1), which are discussed further in the remainder of this chapter:

1. **The available resources and constraints**—including time, timing, expertise and existing data, as well as organisational standards and definitions for evaluation.
2. **The nature of what is being evaluated**—important features of the project, programme or organisation being evaluated, often described by a programme logic or theory of change, also considering the stage of the policy or programme in its lifecycle, and whether it has aspects that are simple, complicated or complex.
3. **The nature of the evaluation**—in particular its purpose, the key evaluation questions it is intended to answer and the requirements of key stakeholders.

All three elements of the framework should be considered concurrently and need to be brought together when choosing methods that are credible, useful and cost effective for producing necessary information.

While this report is about choosing methods for impact evaluation it should not be assumed that every evaluation should focus on evaluating impacts. Many types of monitoring and evaluation exist that will be more or less useful in different circumstances (section 3.5). For example, an impact evaluation that is conducted before a programme is being implemented as intended might conclude the impacts do not warrant continuation of the programme, when it is the implementation, rather that the programme itself that is not performing.

Figure 3.1: Framework for choosing appropriate methods and designs



3.2 Resources and constraints

The appropriate evaluation method depends on available resources and other constraints. This includes whether there is relevant and credible evidence from previous research, evaluation and monitoring and capacity to undertake additional evaluation in terms of:

- funding to cover the cost of an external evaluator or evaluation team, including procurement and contract management costs
- direct or in-kind funding to cover the cost of internal staff time, including an internal project manager and steering committee
- expertise (knowledge and skills in evaluation and in the content area) for both internal and external contributors
- policy and implementation lag
- time available before a report is needed to inform a particular decision
- political sensitivities affecting resourcing, for example in relation to terminating programs.

As well as resource limitations, there can be organisational constraints. These might include norms, policies and processes about credible evidence, a credible evaluation team including an element of independence, ethical evaluation

practice, definitions and standards for evaluation including a voice for intended beneficiaries, or equity issues.

3.3 Nature of what is being evaluated

Programme logic or theory of change

A programme logic, or theory of change, describes how intervention activities are understood to contribute to a chain of intermediate outcomes that produce intended and potential unintended impacts. Programme logic also identifies the assumptions and external factors that will influence the extent to which outputs lead to intended outcomes.

Programme logic can help to identify gaps in logic or evidence which should then be a priority for the evaluation. It identifies intermediate outcomes which could be evident in a shorter time frame and provide early indications of whether or not the programme is on track to achieve the intended impacts and whether the risks of negative impacts are being effectively managed. Programme logic also provides a structure for constructing an evidence-based narrative about the value and impact of a programme.

Making programme logic a part of standard policy and programme development helps to improve the quality of design and implementation as well as the quality of evaluation by identifying logical gaps, key risks and providing a shared understanding among stakeholders about what is involved. It may help to manage expectations of stakeholders who are unfamiliar with other ways of describing or understanding the impacts of an intervention other than simply by reference to ultimate long term impacts which may be difficult to measure.

Stage of development

The maturity of a programme or policy has important implications for the range of impact evaluation designs. For a new program, it can be possible to stage roll-out during the pilot phase in ways that create a control group. These designs are more difficult and often impossible to implement for a programme which has already been widely implemented.

Nature of activities and impacts

In addition to gathering evidence of impacts, it is important to gather evidence of implementation activities. This is needed to be able to distinguish between implementation failure (where the programme failed because it was not properly implemented) and theory failure (where the programme failed despite adequate implementation).

For some programmes, both activities and impacts are visible, so it can be relatively easy to develop valid measures. If activities and/or impacts are not readily observable, a combination of indicators will be needed.

For programmes with standardised processes (like assessing applications), a standard can be developed to measure the quality of implementation in a consistent way. For programmes that involve customised implementation (for example, specific advice to small businesses), expert judgement will be needed of the quality of the activities provided.

It is easier to conduct impact evaluations of programmes with impacts that will be visible in the short-term. For programmes with a long lead-time before impacts will be evident, evaluations can usually only gather evidence of shorter-term outcomes that previous research has demonstrated (or logic suggests) are associated with the long-term impacts of interest. Where possible these evaluations should be complemented by some longitudinal studies to check for long-term impacts.

The impact evaluation of programmes which have transformational impacts that are not readily reversed (such as the transmission of new knowledge) is easier, as measurement can be done at any time and confidently projected into the future. Programmes with fragile impacts which are easily undone present additional challenges for the timing of evaluations.

Simple, complicated and complex aspects of programmes and policies

It can be helpful to consider an intervention in terms of a three-part typology—simple, complicated or complex (Stacey 1992; Glouberman 2001; Glouberman & Zimmerman 2002; Kurtz and Snowden 2003). This has been shown to be useful for planning and analysing evaluations (Guijt 2008; Patton 2008; Rogers 2008b; Rogers 2011). The typology is particularly useful for classifying aspects of interventions rather than the whole intervention.

'Simple' aspects of interventions can be tightly specified and are standardised—for example, a specific product, technique or process. '**Complicated'** aspects of interventions have multiple components, are part of a larger multi-component intervention, or work differently as part of a larger causal package. They might work, for example, in particular implementation environments, for particular types of participants, or in conjunction with another intervention. '**Complex'** refers to appropriately dynamic and emergent aspects of interventions, which are adaptive and responsive to emerging needs and opportunities.

In programmes with simple focus and governance, there is a single organisation involved in the programme and a single set of intended impacts. In programmes with a complicated focus and governance, there may be multiple organisations involved and/or multiple different perspectives about the impacts that are valued. In programmes with a complicated focus and governance, impact evaluations need to identify and gather evidence about multiple possible impacts and synthesise these in ways that reflect the values of the different organisations and agreed trade-offs between different impacts. Attention may also need to be paid to negotiating access to the different data held by partner organisations. Programmes with a complex focus and governance may have not only multiple stakeholders and values but also **emergent** stakeholders and values. As these new players and issues arise, a nimble impact evaluation is needed that can gather evidence about impacts for an emergent list of stakeholders and intended users.

In programmes with simple cause-and-effect relationships, a well-designed programme will be both necessary and sufficient to produce the intended impacts. Counterfactual approaches, which compare what happened to an estimate of what would have happened in the absence of the programme or policy, are appropriate in these circumstances. Where the programme is not sufficient to produce the intended impact, and a complicated causal package of

other programmes or favourable context is also needed, simple comparisons with and without the programme will incorrectly estimate its effectiveness. Where the programme is not logically necessary to produce the intended impacts (that is, there are other ways of achieving them), impact evaluations will need to investigate what services have been accessed by the control group to avoid understating the impact of the programme.

Further implications of complicated and complex aspects of programmes and policies can be found in Appendix 3.

3.4 Nature of the impact evaluation

Intended uses and users

Different intended uses of an impact evaluation, as outlined in Section 2, lead to different types of evaluation questions. Different intended users might have different intended uses and also different views about what constitutes credible evidence.

Table 1 elaborates on the types of impact evaluation questions identified by Stern et al. (2012) and begins to suggest some appropriate methods.

Table 3.1: Different types of impact evaluation questions and relevant methods

Intended use	*Typical evaluation question*	*Conditions*	*Relevant methods and designs*
Attribution	Did the intervention cause the impact(s)?	Requires a single cause and a small number of effects. Needs either a homogenous effect (it works the same for everyone) or knowledge about the contextual factors that influence impacts	RCTs, regression discontinuity, propensity scores
Apportioning	To what extent can a specific impact be attributed to the intervention?	Requires a single effect, large data sets on relevant contributing factors.	Regression, econometrics, structural equation modelling
Contribution	Did the intervention make a difference?	Requires an understanding of the different configurations that could produce the results (which can include contextual factors, variations of the programme and other programmes).	Contribution analysis, comparative case studies, process tracing, Bradford Hill criteria
Explanation	How has the intervention made a difference?	Requires the development of a clear programme theory which sets out a change theory (how change is understood to come about) and an action theory (what activities will be undertaken to trigger this). This can be informed by exploring how actors in the intervention attribute cause and investigate these for plausibility, as well as drawing on research literature and theoretical frameworks.	Actor attribution, theory-based evaluation, realist evaluation, process tracing.
		Where it is possible to identify potential 'active ingredients' in the programme and develop different combinations of what is delivered and test their relative effectiveness. Requires homogeneity of effects as it only provides information about average effects.	Multi-arm RCTS with 2-way or 3-way interactions designed to identify the 'active ingredient'
Generalisability or transportability	Is the intervention likely to work elsewhere? What is needed to make it work elsewhere?	Need an understanding of contextual factors that have affected the implementation and results. Need to identify alternative action theories which might be more suitable in different contexts, or even alternative change theories.	Realist evaluation

Note: See the glossary in Appendix 7 for definitions of methods and designs

It is not always appropriate to ask attribution or apportionment questions in an impact evaluation, as the PGPA Act Resource Management Guide makes clear:

> Many government activities are delivered in complex environments that are constantly changing. Limited control over external factors can make it difficult to link the results of a particular activity with changes observed in the broader environment. In such cases it may necessary to just measure the changes observed, and provide evidence that supports a theory that links those changes to the results of a specific government activity *(Department of Finance 2015*).

Timing of the evaluation

Impact evaluations are usually concerned with the actual effects of an intervention—these are referred to as *ex-post* impact evaluations.

Ex ante impact evaluation occurs before a programme is implemented. It is used to inform decisions about whether or not a programme should be funded based on its likely impacts. This kind of analysis will often draw on existing data about similar programmes. A common form of *ex-ante* evaluation is a prospective cost benefit analysis (See Section 4.4) that may be cited in a new policy proposal to justify expenditure on a programme.

A third option in terms of the timing of an impact evaluation is *during implementation*, when the intended use is to check within a shorter term policy cycle whether a project is on track to deliver longer term impacts (Table 2).

Table 3.2: Impact evaluation before, during and after implementation

Type and intended use	*Typical evaluation question*	*Implication*
Ex post—done after implementation (although started well before this) to inform funding of subsequent programmes or continuation of existing ones	What have been the actual impacts of this programme and policy?	Need a feedback loop to ex ante and during implementation evaluations to iteratively improve estimates
Ex ante—done before implementation to inform funding of potential programmes	What are the likely impacts of this programme or policy if it is undertaken?	Need credible assumptions about likely impacts based on previous research and evaluation.
During implementation—done to provide evidence about likely impacts given current progress	What are the likely impacts of this programme or policy given the current situation?	Need credible assumptions about likely impacts based on evidence about intermediate outcomes and previous research and evaluation.

Monitoring data that has been collected systematically during programme implementation can be used to estimate the programme's contribution, or to confirm that it's on track, alert programme managers to issues or raise questions to be explored. It is not always practical or possible to wait until long-term impacts can be measured in an impact evaluation to generate evidence to inform policy, investment and implementation decisions.

In some cases, assurance that a programme is on track to deliver outcomes is all that can be expected from monitoring and evaluation. This is especially so when a programme is being developed at the same time as it is being implemented and questions are being asked about its impact early in the policy cycle. In complex open systems, characterised by diverse, interdependent entities that adapt to changing conditions (e.g. people, firms) this might be all that can be achieved.

Types of impact evaluation questions

The articulation of key evaluation questions crystallises many of aspects of the design triangle. The purpose, resources and constraints for the evaluation, the

stage of the policy or programme itself and whose information needs it will serve will all shape the evaluation questions.

An impact evaluation involves answering at least three different types of questions—descriptive, causal and evaluative.

- **Descriptive questions** ask about how things are and what has happened. They describe the initial situation and how it has changed, the intervention and other related programmes or policies, the context (participant characteristics) and the implementation environment.
- **Causal questions** ask whether or not, and to what extent, observed changes are due to the intervention in question. Increasingly, causal analysis is about understanding how an intervention contributes to impacts along with other factors and other programs.
- **Evaluative questions** ask about the overall value of a programme or policies, taking into account intended and unintended impacts, criteria and standards and how performance across different domains should be weighted and synthesised.

In addition, impact evaluations that include recommendations need to answer action questions by identifying and assessing possible options for responding to findings.

These types of questions are often combined into specific Key Evaluation Questions, each of which requires a different bundle of methods to answer:

- Did the intervention make a difference?
- How much of a difference did the intervention make?
- For whom, in what situations, and in what ways did the intervention make a difference?
- To what extent can a specific impact be attributed to the intervention?
- How did the intervention make a difference?
- Will the intervention work elsewhere?
- What is needed for the intervention to work elsewhere?

An evaluation is likely to have elements of each of these evaluation questions—and a package of methods will be needed to answer them. For example, an evaluation informing whether to continue funding for a programme will need to include descriptive questions (about what has happened), causal questions (about the role of the programme in producing observed changes) and evaluative questions (about whether the programme has been a success overall, and whether the costs and benefits justify continuation).

3.5 Impact evaluation and other types of evaluation

Impact evaluation is one type of evaluation. Five types of evaluation are commonly recognised as being useful depending on the questions being asked (Owen & Rogers 1999) summarised in Table 3 below.

Ideally types of evaluation are cumulative—impact evaluation uses data from the needs analysis, intervention design, monitoring and process evaluation and economic evaluation requires data from impact evaluation. For example, an

impact evaluation that is conducted prior to a process evaluation will not be able to determine whether it was the programme itself or poor implementation that led to a lack of observed impacts—unless it also incorporates a process evaluation.

Table 3.3: Other types of evaluation

Type of evaluation	*Types of questions asked*
Values clarification	What is needed? What would success look like?
Monitoring	How is it going? (Regular reporting of metrics)
Process evaluation	How is it going (periodic investigations)? Is it being implemented according to plan or according to data-informed revisions to that plan? What has been done in an innovative program?
Economic evaluation	What has been the value of the intervention? Has the intervention been cost-effective (compared to alternatives)? What has been the ratio of costs to benefits?

A **values clarification or needs analysis** will identify the criteria for success, including identified and prioritised needs, and policy positions about the appropriate role for government. This is needed in an impact evaluation to be able to make an overall judgement about the success of the programme.

Monitoring can be useful for accountability by facilitating construction of performance measures, data aggregation and comparisons or trends over time. It can be used in advocacy by providing performance measures that enable the programme administrator to 'tell the story' and in allocation by ensuring standardised data is collected, allowing for comparisons of relative merit. It can be used in analysis by providing a relatively simple means of tracking inputs, outputs and to an extent, outcomes, over time.

Monitoring and process evaluation will provide information about the quality of implementation and changes or trends in the problem being addressed by a programme. This can allow an impact evaluation to distinguish between

- implementation failure—where the programme did not produce the intended impacts but was not adequately implemented
- theory failure—where the programme was adequately implemented but did not produce the intended impacts (meaning that the theory of how the programme should work is incorrect).

One particular approach to monitoring which has been used in government is Results Based Accountability (RBA) developed by Mark Friedman. Like many forms of monitoring RBA pays attention to the results an intervention or programme is trying to achieve, i.e. changes in problem conditions, rather than measuring attributable impacts of particular interventions or programmes. The particular feature of this approach is distinguishing between results for clients of a programme (the responsibility of programme managers) and results for the

wider population (the responsibility of a partnership of programmes and organisations). The approach is outlined in Appendix 4.

Monitoring relies on the ongoing collection of good quality data. Successful monitoring strategies use data already collected through administrative systems and only collect new data that is actually needed. They also provide those who submit data with reports that explain how their information has been used, motivating them to provide good quality data.

4. How can we describe, measure and evaluate impacts? What methods suit different situations?

The previous section described how impact evaluation involves answering at least three different types of questions—descriptive, causal and evaluative. In this section we look at the different methods that can be used to answer these questions and when they can be used effectively. We then turn to economic analysis and consider how to answer questions about economic impact.

4.1 Descriptive questions

Descriptive questions ask how things are and what has happened, including describing the initial situation and how it has changed, the intervention and other related programmes or policies, the context in terms of participant characteristics, and the implementation environment. This might include pre and post-implementation review of the changes in a situation that a programme set out to address.

A range of data collection and analysis methods can be used to gather evidence showing that changes have (or have not) occurred by:

- sampling (random and purposeful sampling)
- using measures, indicators and metrics from existing data sets and sources
- collecting and retrieving data
- managing data, including organisation, storage and quality checking of data
- combining qualitative and quantitative data
- analysing data, particularly options for identifying patterns in the data
- visualising data.

Random sampling or a census will be needed to produce a firm quantitative measure of what happened before, during and after the programme. Purposeful sampling is appropriate when seeking to learn from the most or least successful cases (extreme case sampling) or for testing the programme theory by understanding cases which don't fit the overall patterns (theoretical sampling).

Measures, indicators and metrics can be built into programme monitoring systems and draw on existing administrative data as well as sources that indicate the condition (e.g. ABS data on population or economic activity).

Internal validity—It is important to be confident about the quality of data collection tools, especially where there are reasons why responses might not represent the situation comprehensively or accurately. Programme managers should ask to what extent the descriptions truly reflect what is going on. Triangulation, using multiple data sources, can improve confidence in the internal validity of findings.

External validity—Programme managers should also ask to what extent and in what ways the findings can be confidently generalised to other sites and times. For example, an evaluation of changes in one industry or science programme may not be relevant for another industry or programme.

Likely impacts can be estimated by developing a programme logic of the intervention, identifying intermediate outcomes and important contextual factors, reviewing previous research and evaluation and using evaluation data to provide evidence of shorter-term outcomes and previous data to provide evidence of likely future impacts.

Methods and designs for answering descriptive questions

Table 4 below gives an overview of some of the different tasks involved in answering descriptive questions in an evaluation and relevant methods and processes. Definitions of these techniques can be found in the glossary at Appendix 7.

Table 4.1: Methods and processes for answering descriptive questions

Task	*Options*
Sampling	
Probability samples	Multi-stage; simple random sample; stratified random sample
Purposeful samples	Confirming and disconfirming; criterion sample; critical case; homogenous; intensity; maximum variation; outlier; snowball; theory-based; typical case, extreme case
Convenience samples	Convenience sample; volunteer sample
Use measures and indicators (develop or use existing)	Targets; indexes; standards
Collect and/or retrieve data	
From individuals	Interviews; opinion polls; questionnaires and surveys; assessment scales or rubrics; goal attainment scales; logs and diaries; mobile phone logging; expert reviews; polling booth; postcards; projective techniques; seasonal calendars; mapping; stories and anecdotes
From groups	After action review; brainstorming; concept mapping; Delphi study; dotmocracy; fishbowl technique; focus groups; future search conference; hierarchical card sorting; keypad technology; mural; ORID technique; Q-methodology; SWOT analysis; world cafe; writeshop
Observation	Field trips; participant observation; non-participant observation; photography and video; transect walks
Physical measurements	Biophysical; geographical
Existing data	Big data; official statistics; previous evaluations and research; project records; reputational monitoring dashboard
Combine qualitative and quantitative data	
In terms of when qualitative and quantitative data are gathered	Parallel data gathering; sequential data gathering
In terms of when qualitative and quantitative data are combined	Component design; integrated design
In terms of the purpose of combining data	Enriching: using qualitative work to identify issues or obtain information on variables not obtained by quantitative surveys. Examining: generating hypotheses from qualitative work to be tested through the quantitative approach. Explaining: using qualitative data to understand unanticipated results from quantitative data. Triangulation (confirming/reinforcing; rejecting): verifying or rejecting results from quantitative data using qualitative data (or vice versa)

Analyse data	
Numeric analysis	Correlation; cross-tabulations; data mining; exploratory techniques; frequency tables; measures of central tendency; measures of dispersion; multivariate descriptive; non-parametric inferential statistics; parametric inferential statistics; summary statistics; time series analysis
Textual analysis	Content analysis; thematic coding; framework matrices; timelines and time-ordered matrices
Visualise data	
See relationships among data points	Scatterplot; matrix chart; network diagram
Compare a set of values	Bar chart; block histogram; bubble chart
Track rises and falls over time	Line graph; stacked graph
See the parts of a whole	Pie chart; treemap; icon array
Analyse a text	Word tree; phrase net; word cloud
See the world	Demographic mapping; geotagging; GIS Mapping; interactive mapping; social mapping

4.2 Causal questions

Causal questions ask about the cause and effect relationship between the intervention and the changes that have been observed. This includes attribution (where the intervention can reasonably be said to have caused the changes) and contribution (where the intervention is one of several factors together producing the changes).

These are the three main causal inference strategies, with a list of possible methods for each one in the following table (Table 5):

- Counterfactual—Constructing an estimated or hypothetical case of what would have happened without the programme and comparing this to what actually happened. Includes experimental designs, which construct a control group through random assignment (randomised controlled trials) and quasi-experimental designs which create a similar comparison group (matched comparisons, regression discontinuity, propensity score matching).
- Consistency—Checking evidence is consistent with the programme theory in terms of the timing and patterns in the data, including actively searching for outliers and data that don't match and seeking to explain them.
- Alternative explanations—Ruling out alternative explanations by identifying other possible explanations for the observed changes and investigating whether they can be ruled out.

Table 4.2: Possible methods and designs for different causal inference strategies

Causal inference strategy	*Possible methods 2*
Compare results to the counterfactual	
Experimental research designs	Control group; Randomised controlled trial (RCT)
Quasi-experimental research designs	Difference-in-difference (or double difference); instrumental variables; judgemental matching; matched comparisons; propensity scores; sequential allocation; statistically created counterfactual; regression discontinuity
Non-experimental options	Key informant interviews (hypothetical counterfactual); Logically constructed counterfactual
Check results support causal attribution	
Gathering additional data	Actor attribution; modus operandi; process tracing
Analysis	Bradford-Hill criteria (dose-response patterns; intermediate outcomes check timing of outcomes); compare to expert predictions; comparative case studies; qualitative comparative analysis (QCA); realist analysis of testable hypothesis.
	Contribution analysis; collaborative outcomes reporting; multiple lines and levels of evidence (MLLE); rapid outcomes assessment.
Investigate possible alternative explanations	Force field analysis; general elimination methodology; key informant interviews; process tracing; ruling out technical explanations; searching for disconfirming evidence / following up exceptions; statistically controlling for extraneous variables

Broadly, there are three groups of possible methods and designs for answering causal questions:

- **Experimental designs** which construct a control group through random assignment such as randomised controlled trials.
- **Quasi-experimental designs** which construct a comparison group through matching for example, regression discontinuity, and propensity scores.
- **Non-experimental designs** which construct hypothetical counterfactuals or use non counterfactual strategies to test causal inference. They look systematically at whether the evidence is consistent with what would be expected if the intervention were producing the impacts, and also whether other factors could provide an alternative explanation, for example, contribution analysis; econometric modelling; qualitative comparative analysis; process tracing; comparative case studies; multiple lines and levels of evidence.

Appendix 5 provides examples of causal inference methods.

[2] More information on these methods can be found in the glossary in Appendix 7.

Experimental designs

Many frameworks such as the Maryland scale for evaluating causal evidence have a hierarchy of evidence that prioritises **randomised controlled trials** (RCTs) as the most credible and robust evidence of the effectiveness of an intervention. Even more compelling is a systemic review or meta-analysis, combining the results of numerous RCTs.

By randomly allocating a sufficient number of potential participants to either the control or treatment(s) groups, RCTs are intended to create groups equivalent on initial conditions. Any differences between groups observed at a later time can be considered the result of the intervention. In RCTs the process of randomisation controls the effects of both observable and non-observable or unknown factors affecting outcomes. This provides a more accurate measure of the independent and attributable impact of an intervention or programme than approaches that rely only on what is known or estimated.

RCTs are appropriate *when the types of questions they can answer are being asked*, and *when these pre-conditions for their use can be met*:

- The intervention can itself be defined in a meaningful way such that what was tested could be reproduced. This would mean that the intervention is mature rather than still under development and evolving; and has been or is being implemented as intended.
- Random allocation into treatment and control groups can be undertaken and maintained.
- Sample size is sufficient to accurately detect differences between treatment and control groups given the expected strength of the intervention and the length of the causal chains.

RCTs measuring long term outcomes of DIIS programmes are likely to require very large sample sizes. As an intervention interacts with other factors also affecting long term outcomes, its influence over time will dissipate. So even if the intervention is still significant, it will require increasingly sensitive experiments or in experimental evaluation, very large sample sizes, and raises the practical difficulty of constructing very large control groups.

It is important when doing an RCT to consider that the ‘average’ effect hides both winners and losers. Doing more work to understand who benefits and who doesn’t from a programme and then targeting those that stand to benefit the most will lead to more efficient programmes and evaluations with a larger effect size. It is for this reason that RCTs are appropriate at the end of a long period of programme development, not as the first method for evaluating a new programme or intervention. Most other scientific approaches to theory development, experimentation and evaluation use this approach and this is why RCTs are part of Stage IV in clinical trials following 10-15 years of small-scale testing. Applying RCTs to whole programmes that are not sufficiently developed will generate a very low, and possibly negative return on investment (ROI) from evaluation, and likely also on any ROI calculated for the intervention itself.

If the question is *‘what is it the makes the programme work, when does it work, and how should we shape it to maximize the benefits?’* then RCTs are unlikely to be fine-grained enough to reduce uncertainty or inform decisions about how a programme could be delivered more efficiently, or effectively, for whom and

under what circumstances. Formative or developmental impact evaluation might be more appropriate. Even if sub-group analysis of the data from an RCT can give some hints about programme improvement, this approach will be less cost effective than methods focused on understanding how and when the intervention generated impacts.

At the heart of the RCT is the desire to control for 'confounding factors'. This works in agricultural research, but in complex systems, these contextual factors are often exactly what determines when and where an intervention is effective. One approach to this problem is to move away from the idea of using RCTs of whole programmes to endorse specific interventions (Bonell et al. 2012) to testing components or causal theories *within* a program. These approaches suggest that an intervention may be effective in certain contexts by firing a causal mechanism—and that these relations between mechanism and context are more appropriate units of analysis than entire programs (Pawson 2013).

Quasi-experimental designs

Quasi-experimental designs seek to mimic what can be achieved in terms of causal inference by a well-constructed RCT. These types of impact evaluations are useful when random allocation to treatment or control groups is not possible. Instead of using random allocation to construct treatment and control groups they rely on other methods for constructing a counterfactual or comparison group.

The key issue for these methods is reducing selection bias (which is avoided by random allocation in an RCT) such that any differences between the treatment and comparison groups can reasonably be attributed to the impact of the intervention (Shadish et al. 2002, p. 14).

Regression discontinuity is useful where an intervention is made available above or below a cut-off point, and data is available about participants and non-participants. This method is particularly useful with large data sets where the large sample sizes make it is easier to detect small differences.

Propensity score matching is useful when participation is voluntary. It creates comparison groups by matching people on the factors which influence their propensity to participate.

Hunting for **'natural experiments'** where through some chance event some people experienced an intervention and other did not, may be useful and low cost.

Difference in Difference can be useful for measuring changes in the amount of difference between treatment and comparison groups over time. This compares differences that occurred during the intervention for the treatment group and the comparison group. This method is best used with either propensity score matching to take into account any *known* differences between the two groups on initial conditions, or regression discontinuity to take into account different *trends* in the treatment and comparison group.

Non-experimental designs

Experimental and quasi-experimental approaches to causal impact evaluation are not always possible. In these contexts other methods are required.

Hypothetical and logical counterfactuals It is sometimes possible to construct a credible counterfactual (an estimate of what would have happened in the absence of a programme or policy) without constructing a comparison or control group. This involves consulting with key informants to identify either a hypothetical counterfactual (what they think would have happened in the absence of the programme or policy) or a logical counterfactual (what would logically have happened in its absence).

For example Industry Canada's 2015 evaluation of its contribution to Canada's Advanced Research and Innovation Network (CANARIE) which supports research, discovery and innovation in Canada by providing Canadian research and education communities with a high-speed network for data transfer) found that

> In all, the absence of CANARIE would have a profoundly negative effect on research, education and innovation in Canada. Third-party evaluation interview and survey respondents outlined some of the following potential impacts:
>
> - Without CANARIE, interprovincial and international linkages for Canadian researchers and educators would be seriously jeopardized and the Canadian position on the international arena, fragmented
> - Without CANARIE and its community of resources, research would become more isolated and happen in silos. It would become much less creative and less productive.
> - It would put Canada at a significant disadvantage, nationally and internationally.

The evaluation concluded that discontinuing CANARIE would be 'catastrophic' and some research and educational activities would stop. It found the impact would be greater in smaller provinces that were more dependent on CANARIE for connectivity, with few alternative commercial options available (Industry Canada 2015a, p.8).

Using a hypothetical counterfactual is only appropriate when the situation is reasonably predictable and the key informants have extensive knowledge about usual patterns of outcomes. It is most appropriate when informants have no incentive to present a particular view, or where their reported hypothetical counterfactual can be justified and tested by reference to other information and other informants.

Qualitative comparative analysis is particularly useful where there are a number of different ways of achieving positive impacts, and where data can be iteratively gathered about a number of cases to identify and test patterns of success.

Contribution analysis

Contribution analysis is a way of combining evidence for systematic causal inference. It can draw on evidence from experimental, quasi-experimental and/or non-experimental studies. It derives from theory-driven approaches to evaluation that emphasise the programme theory or programme logic of an intervention, but it addresses the need to take into account external factors also expected to

affect outcomes. Contribution analysis is designed to reduce uncertainty about the contribution of an intervention to observed results by providing an increased understanding of why the results have occurred (or not) and the roles played by the intervention and external factors.

A contribution analysis produces a credible contribution story about how the programme or intervention contributed to desired outcomes. There are two main features: developing a theory of how the programme works and the external factors expected (or known) to shape outcomes; and gathering evidence about the extent to which the programme worked as intended and the external factors which actually affected outputs and outcomes. In short, it is a claim supported by evidence that pays attention to the internal logic of the programme and knowledge of external factors also affecting outcomes.

The production of a contribution analysis performance story can be achieved in six steps:

- **Set out the attribution/ contribution problem** to be addressed. This means acknowledging the problem and setting out the extent to which specific cause and effect relationships will be addressed and the likely external factors that will also affect the production of outcomes.
- **Develop a theory of change** or programme logic and provide the evidence on which it is based. This answers why you would think this programme could or would be effective in producing the intended outputs and outcomes.
- **Gather the existing evidence** on the theory of change and external factors. This will help determine whether each step in the programme logic seems to have been achieved. If some steps have not been fulfilled there may be a failure of theory or failure of implementation that should be addressed. This is often the last step in programme logic-based evaluation. But contribution analysis gathers evidence about whether external factors have affected (supported or worked against) the extent to which outputs were translated into outcomes.
- **Assemble and assess the contribution story**, or performance story, as it will answer questions about whether the programme appears to have been implemented as intended. It will also consider the extent to which expected higher order outcomes can be linked to programme activities. It might lead to a re-evaluation of the programme's success (if external factors appear to have driven outcomes or if external factors worked against an otherwise promising programme).
- **Seek out additional evidence** to strengthen the credibility of the performance story. The theory of change may itself need to be changed as a result of evidence collected. Stakeholders or subject matter experts may offer suggestions about other factors that supported or worked against the realisation of expected outcomes that can be further investigated.
- **Revise and strengthen the contribution story**, where the additional evidence permits, through successive iterations that progressively address weaknesses in the analysis or new insights and alternative explanations.

As its name suggests, contribution analysis is focused more on contribution than attribution. Typically data will be assembled that supports (or refutes) a claim about the extent to which a programme was responsible for delivering an

intended outcome. Its main weakness for causal impact evaluation is that it does not readily allow quantification, i.e. measurement of the attributable impact of an intervention.

Advice on the choice of causal evaluation methods

When doing impact evaluation it is important to follow the framework described in Chapter 2, to ensure evaluation methods are designed to answer the questions being asked rather than have questions based on the methods to be used.

In this section we describe some common situations that DIIS programme managers may find themselves in when thinking about causal impact evaluation methods. We start our discussion with situations in which DIIS programme managers may be thinking about causal impact evaluation methods with simple interventions and short causal chains. We move to more complicated programs that include multiple interventions and longer causal chains. We conclude with more complex situations where the programme itself is changing in response to emerging results while the intended beneficiaries are adapting both to what the programme has to offer and to their ever-changing environment. In this situation a programme may have only very small impact on behaviour relative to other influences.

Common situations for DIIS programme managers and causal impact evaluation methods that may be useful include the following.

- **When there is a very simple intervention and a very short causal chain**, where short term outcomes are likely to be good predictors of long term outcomes and/or when you don't need to understand how an outcome was achieved, A/B testing may be used. This form of RCT is often used in IT where people are randomly served one version of a website and some behavioural outcome of interest is measured, such as clicking a link. You might test whether to put a login box on the left of right of a website screen. Here the causal chain to the behaviour is extremely short (logging in or not logging in); it is not important to understand why people logged in more to one box than the other; short term behaviour is not being used to predict use of a website more broadly (i.e. long term impact); and there is opportunity for trial and error. In short it can be useful to 'test, learn and adapt' (Haynes et al. 2012) because the stakes are sufficiently low that false positive or negatives don't matter too much and it is not necessary to understand how or in what contexts an outcome was achieved.
- **A combination of different interventions** is **deployed into very different contexts** (such as access to business advice and/or resources etc.). In these situations RCTs are not very useful (especially where programmes are partially formed, implementation is inconsistent, and it is highly likely that interventions will have different impacts in different short and long term average effects). A theory-driven approach complemented with RCTs might be more useful. This means developing programmes based on current understanding (or theory) of what is likely to be effective where and for whom. In these cases, you might use factorial or multi-arm RCTs to test components of well-understood interventions, implemented as intended, where they are expected to have an impact. The bigger the expected impact, the shorter the causal chain and the larger the sample size for a

specific intervention or combination of interventions, the more likely the results are to be accurate and replicable.

- **When the programme is stable, mature and composed of powerful elements being implemented as intended, with significant outcomes expected across specific contexts,** RCTs may be used to test whole programmes. An RCT applied to a whole programme can provide a statistic about the size of the effect attributable to a programme. The dangers of applying RCTs in situations other than these include that a promising programme may not be sufficiently valued and is discontinued, or a trend develops towards programmes that are not very sophisticated or wide-ranging in their benefits, but easy to measure.
- **When managing complicated or complex programmes and their component interventions** where it is not feasible to randomly allocate participants to treatment and control groups, impact evaluation may involve testing theories about what works for whom in different circumstances. It may involve quasi-experimental methods such as propensity score matching or regression discontinuity which can rely on routine administrative data where counterfactual groups are generated from within the dataset. These methods may be used when the administrative data includes information about elements thought to affect results (i.e. contextual factors used in propensity score matching). Alternatively, they may be used where eligibility for the programme turns on an arbitrary data item e.g. a tax incentive provided to business with 19 employees but not 20 employees (i.e. where results from 'discontinuity' in regression for those with 19 and 20 employees can provide evidence for impact).
- **When a programme is new, not well understood or not being implemented as intended, formative or developmental evaluation** (with monitoring of outputs and problem conditions) and descriptive impact evaluation may be required. Non-experimental methods based on programme logic and knowledge of external factors designed for monitoring in complex situations, or dealing with unmeasurable factors, might be more useful in decision-making than an RCT. Where a programme has diverse elements that are deployed in a variety of ways, or where outcomes may only emerge over long periods of time, an RCT is likely to return a very low return on investment/ neither showing positive effects for advocacy, nor providing useful information for analysis, unless there is good reason to expect that large measurable outcomes will generally result from activities.

Examples of answering causal questions

We have developed several examples to illustrate issues in answering causal questions in the context of industry and science policy. The first example (Project Gate) was identified in the literature search and provides an instructive example of a well-executed randomised controlled trial that still led to uncertain results.

The second example considers what could be learnt about impacts of the Entrepreneurs' Infrastructure Programme and Manufacturing Transition Programme.

Project GATE—high quality RCT but uncertain results

Project GATE provides an interesting example of an RCT that could specify the intervention, was able to randomise and did have sufficient sample sizes and follow-up periods—but was still generally considered not useful for policy-making by the National Bureau of Economic Research (NBER) in the US.

Project GATE was a demonstration programme designed to offer an array of self-employment training services through the US workforce development system to individuals interested in self-employment. The program, implemented from 2003 through 2005 in Maine, Minnesota, and Pennsylvania, included an outreach campaign for recruiting applicants, with designated One-Stop Career Centres serving as central points of recruitment. At the end of the recruitment period, 4198 individuals applied to participate and were randomly assigned to the treatment group or to the control group. Participants and control group members were followed up at six months, 18 months, and 60 months.

At least two reports were published by IMPAQ (2008 and 2009) and the data was reanalysed by the National Bureau of Economic Research (2012). The 2008 IMPAQ report found evidence of short-term gains for business ownership. The 2009 evaluation moderated the extent to which short term gains held over time, and now found these only held for participants that commenced the programme when they were unemployed. It supported the continued funding of the programme for unemployed participants. The evaluation reported:

> Our results show that Project GATE was effective in assisting unemployed participants start their own business, leading to significant gains in self-employment and overall employment in the first months following programme entry. The program was also effective in assisting unemployed participants to remain self-employed even five years after program entry. However, we find no evidence that the program was effective in assisting non-unemployed participants improve their labour market outcomes. Based on these results, we conclude that U.S. state workforce agencies should consider adopting self-employment training programs targeting the unemployed as part of their workforce development agenda.

The 2012 re-analysis of the same RCT data set by the NBER confirmed patterns in the data but concluded that the benefits even to the unemployed did not hold over time.

> Using data from the largest randomized control trial ever conducted on entrepreneurship training, we examine the validity of such motivations and find that training does not have strong effects (in either relative or absolute terms) on those most likely to face credit or human capital constraints, or labour market discrimination. On the other hand, training does have a relatively strong short-run effect on business ownership for those unemployed at baseline, but not at other horizons or for other outcomes. On average, training increases short-run business ownership and employment, but there is no evidence of broader or longer-run effects on business ownership, business performance or broader outcomes.

The implications from this analysis were not to continue with the programme in any form—at least based on the available evidence which lacked nuance about which mechanisms of the intervention were effective and the different contexts in which this was the case. The final paragraphs of the NBER report conclusion of this large-scale long-term study are in line with the advice in this paper about testing the components of programmes.

> In all, the absence of positive treatment effects across numerous measures of business ownership, business performance and broader outcomes, and the estimated $1,321 per-recipient cost of providing GATE training, suggests that entrepreneurship training may not be a cost effective method of addressing credit, human capital, discrimination, or social insurance constraints. This conclusion contrasts with the positive benefit/cost conclusion reached in the final evaluation report submitted to Department of Labor, and with similarly positive arguments proffered by advocates of state-level programs. ***Our results also speak to the importance of understanding which components of training are more and less helpful, and for which populations [emphasis added].*** Should subsidies for entrepreneurship training be re-allocated to job training? Should content from entrepreneurship training be grafted onto job training? Are there groups thus far not identified for whom entrepreneurship training may be beneficial in the longer run? Understanding more about the effects and mechanisms of entrepreneurship training is particularly important given the continued growth and popularity of these programs around the world.

There are three main lessons from this case. The first is that even if the overall impacts of an entire programme could be measured by an RCT, the approach did not generate useful information about the value of the components and it was this information that the National Bureau of Economic Research needed to make decisions about the future of the program. The second lesson was how the choice of different outcome measures and definitions by various evaluation teams led to different conclusions about the impact of a program, and whether funding should be maintained, extended or cancelled. The third and associated lesson was about the danger of making conclusions about the value of a programme using an RCT for either measuring short-term or long-term gains—short-term outcomes may not hold, and long-term outcomes may not be evident until many years into the future, or may tail off due the influence of other factors on both the treatment and control groups.

Entrepreneurs Infrastructure Programme (now known as the Entrepreneurs' Programme)

The Entrepreneurs Infrastructure Programme (EIP) has three distinct elements (business management, research connections and accelerating commercialisation). As Paul Jensen (2015, p. 25) points out, the external validity of an RCT applied to this programme would be compromised by the inability to specify the intervention.

Problem: The first problem with specifying the intervention is, how would we know how much of each one of the three elements is necessary, and in what order or whether it is just a dose of EIP that works? The second problem is, as Jensen points out, that it appears the success of the programme is closely related to the skill of the adviser.

An RCT might tell us what happened, but it will not deliver information about what is likely to happen in a future situation with different advisers or what to advise new programme managers about the best course of action with future businesses. As people with knowledge of the programme within DIIS have said

> With the Entrepreneurs Infrastructure Programme (EIP) we are up against the limit in terms of complexity. It is a facilitation first model, a tailored service, so everything is different every time!

Possible solution: An alternative could be used to test theories about which elements within the intervention work in different circumstances, perhaps with some control of the level of expertise of the advisors. Administrative data and a prospective propensity score matching approach or regression discontinuity could be used (for example, comparing outcomes for relevant firms just under and just over the $20 million turnover threshold for eligibility for the accelerating commercialisation component). Other alternatives might be more descriptive or evaluative impact methods, such as comparative case studies, multi-criterion qualitative causal analysis or contribution analysis. While these methods will never have the same internal validity as an RCT, they may have greater external validity and usefulness, making them more cost effective in the information they can generate to inform decision-making. They will not 'prove' the programme is effective (but neither will an RCT). Neither will they deliver a neat outcome measure for a cost benefit analysis, but what they lack in precision about the average effect they would make up in providing better information about the range of impacts produced in different contexts.

Manufacturing Transition Programme

The Manufacturing Transition Programme (MTP) requires the government to pick promising manufacturing projects for funding to make them more competitive and sustainable. The intervention is in the form of a grant between $1-10 million from a budget of $50 million. In this case, inadequate sample size and long causal chains and inadequate theory about contexts in which RCT will be effective mean RCT is likely to be invalid.[3]

Jensen (2015, p. 26) identifies a problem of self-selection bias (that is, only those motivated to apply for the grant will apply) and whether randomisation is possible. This is a real world issue with which all grant programmes must deal.

Problem: There is a problem with the size of the treatment group. As Jensen points out (2015, p. 27), the maximum size is 50 (if the $50 million budget is allocated in $1 million grants) but could be as small as five (if the $50 million budget is allocated in $10 million grants). This means that the sample sizes will have insufficient power for statistical analysis if the suggested long term outcomes of 'new product launches 'or 'new job creations 'and 'new export markets' are to be measured for treatment and control groups.

If an RCT shows no effect, there will not be adequate statistical power to determine whether there truly was no effect, or whether the sample size was insufficient to detect the effect. This could lead to a programme being labelled ineffective just because there was a lack of evidence that it was effective rather than because there was evidence that it was actually ineffective. If the programme has a very strong effect it may be possible, even with this sample size, to identify differences and conclude it was effective. However, this might be unlikely given the long causal chains and the potential for companies to apply for a grant to fund things they were going to do anyway.

[3] There are two issues here: the first is whether there is sufficient reason to expect the programme will provide a net benefit rather than a deadweight loss 'on average' (i.e. companies applying for the grant to do what they planned to do anyway) and the second is whether there are reasons to expect it will work better in some situations rather than others—which should then be the situations in which it is provided—but this is primarily an issue for programme design rather than evaluation.

A second problem relates to the lack of theory about when the programme will be effective. If the programme could be considered to be effective in all situations, there is a danger that the average effect hides evidence of situations where there was a large effect and situations where there was no apparent effect, or possibly harm as a result of the grant (e.g. job losses).

Possible solution: A useful approach in this situation would be to develop an understanding of when cash grants lead to competitive and sustainable business. Realist qualitative analysis could be used to identify contexts where additional finance leads to a manufacturing business being reinvented or reorientated. These theories could be based on behavioural economics and/or other research. Comparative case studies or qualitative causal analysis with recipients could identify the situations where it is likely to be most effective. Contribution analysis could identify the likely causal impact—see below).

Once developed these theories may be tested with more fine grained RCTs, testing theories within the MTP and using shorter causal chains (such as changes to decision-making) to deal with the problem of small sample size before applying an RCT to the whole programme. This would help to develop a programme with more evidence about its component effectiveness before it is put to the test as a stable and mature programme. In the long run, this would result in a programme with bigger impacts and it would avoid the risk of discarding a promising intervention by not understanding the conditions under which it works and/or having an inadequate sample size in which to detect the average effects.

A **contribution analysis** for the MTP would follow the steps below:

Step 1 might start with identifying what part of the causal question is to be answered, perhaps the extent to which grants lead firms to be more competitive and sustainable by building skills and/or moving into higher value activities or markets. External factors would need to be identified that previous research suggests also affect these outcomes, for example the currency exchange rate, the supply of skilled labour, the demand for higher value products etc.

Step 2 begins with the theory of change or programme logic that would need to address *how* the grants are expected to affect business behaviour. A useful programme theory in this situation might attempt to address the ability of the programme to handle situations where the grant represents a dead-weight loss (that is, the business would have done the transition anyway). If this is not the case there may be a problem with the internal logic of the programme such that it would probably not be a good use of resources to do an impact evaluation. Instead a formative, developmental or process evaluation might be more useful.

Step 3 involves gathering evidence. This might include surveys and case studies with grantees, interviews with the grant committee, and/or analysis of key administrative data sets with evidence about important outcomes for grantees. Evidence also needs to be gathered about external factors that may have affected the ability to produce intended outputs, or the extent to which these were translated into ultimate outcomes. This might include interviews with selected firms or subject-matter experts as well as access to statistical data sets on identified factors such as the exchange rate and other factors thought to affect outcomes.

Step 4 would require taking all this data and evidence that supports or challenges the internal logic of the programme and the extent to which each step in the programme theory was or was not affected by external factors to develop a contribution performance story. This is essentially an argument about the programme effectiveness. A claim about observed changes (not exactly attributable outcomes) is supported with evidence about how these changes can be associated with the programme. This evidence will be the extent to which the programme's outputs or intermediate outcomes occurred, and the extent to which other factors either supported or worked against the transition and sustainability of the firms involved.

Step 5 might include dealing with feedback from internal or external stakeholders who challenge the claims and the evidence offered and offer alternative explanations, by talking to other experts or identifying journal articles that describe other external factors not originally identified and data collected.

Step 6 would be to produce a new contribution analysis story with greater credibility, taking into account additional factors or interpretations of the data.

4.3 Evaluative questions

Evaluative questions ask about the overall value of a programme or policies, and whether a programme or policy can be considered a success, an improvement or the best option. They take into account intended and unintended impacts, criteria and standards and how performance across different domains should be weighted and synthesised.

A programme that is effective in terms of meeting its objectives might not be judged a success if it also produced large negative impacts or if the impacts were concentrated on sectors that were not the priority focus. On the other hand, in a context of worsening economic circumstances, a programme might be judged successful in terms of reducing a decline in employment even if it has not met its original targets.

In any impact evaluation, it is important to define first what is meant by 'success' (quality, value). One way of doing so is to use a specific rubric that defines different levels of performance (or standards) for each evaluative criterion, deciding what evidence will be gathered and how it will be synthesized to reach defensible conclusions about the worth of the intervention. At the very least, it should be clear what trade-offs would be appropriate in balancing multiple impacts or distributional effects.

Criteria relate to the aspects of a programme that define success. For example, one criterion may be 'satisfied customers'. Standards provide an objective context for interpreting criteria or outcomes. They may specify a minimum standard of customer satisfaction or that an 'adequate' programme will achieve a certain score on a criterion, while an 'excellent' one achieves some higher score.

After criteria and standards have been agreed, appropriate data collection methods are chosen. For example evidence of 'customer satisfaction' may come from a satisfaction survey using questions that have been shown to accurately measure customer satisfaction, or from direct observations of customer behaviour, or by inspecting reported complaints.

Methods and designs for answering evaluative questions

The selection of methods, designs and approaches should include making appropriate choices in terms of articulating transparent and defensible:

- evaluative criteria—the domains of performance
- evaluative standards—the levels of performance
- evaluative synthesis—how these should be combined when judging success and whether there are minimum essential standards for some or all criteria.

These forms of impact evaluation seek to understand the value of a programme in a future context or application. They will emphasise the alignment of an intervention with the needs of a particular situation. The goal is still that impacts may be maximised, but this approach does not assume that the best way to generate future impacts is to measure past impacts which may no longer be achieved when re-applied to different actors or in different contexts. What worked as subsidies for IT firms in 2008 may not work as subsidies for manufacturing firms in 2016.

Options for identifying the criteria and standards for judging success include:

- reviewing formal statements of values—including stated objectives, policy statements, international obligations
- articulating tacit values, especially among diverse partners—including values clarification interviews, public opinion polling, concept mapping, Most Significant Change
- negotiating different values—including consultations, Delphi study.

Options for synthesising evidence about impacts in terms of performance across different criteria include: Numeric Weighting; Multi-Criteria Weighting; Rubrics.

Options for synthesising evidence from impact evaluations with consideration of resources used include: Cost Benefit Analysis; Cost-Effectiveness Analysis; Cost Utility Analysis; Social Return on Investment; Value for Money.

We note that most previous impact evaluations of industry and science programmes have largely been based on stated objectives. In an increasingly complex environment there is considerable value in going beyond objectives and expanding the range of methods used to clarify values and synthesise evidence in terms of those values.

Explicit trade-offs need to be made between positive and negative impacts (for example, positive economic results but negative social results), between success in some aspects and not in others, and for some market segments and not for others. To do this, it is necessary to combine or summarise the data across one or many evaluations.

Synthesise data from a single evaluation—techniques include consensus conference; expert panel; lessons learnt; multi-criteria analysis; numeric weighting; qualitative weight and sum; rubrics. The emerging method of social return on investment (SROI) explicitly addresses both economic and social values, reflecting the interests of both funders/investors, and direct programme beneficiaries.

Synthesise data across evaluations—techniques include best evidence synthesis; lessons learnt; meta-analysis; meta-ethnography; rapid evidence assessment; realist synthesis; systematic review; textual narrative synthesis; vote counting.

For **action questions**, techniques include generalisation of findings (statistical generalisation, analytical generalisation), contribution analysis and positive deviance (a participatory evaluation practice).

4.4 Economic analysis

Questions about economic impact are particularly important for allocation decisions. Economic analysis methods answer questions about the value of a programme or policy—in particular 'Was it worth it?' or 'Is it likely to be a good investment?' For impact evaluation these are questions about the value of the impacts produced (or likely to be produced) relative to the costs of producing them.

In a public policy environment, costs often include one-off, fixed or sunk costs for the design, administration and evaluation of programmes, such as staff time and other overheads. There are also often variable costs associated with delivery, such as the amount of grants, or subsidies provided to those implementing a programme. Costs also include negative impacts from a programme or policy.

Economic analysis can be conducted prior to a programme being delivered (*ex-ante)* or after a programme has run (*ex-post*). In the former it is used to weigh up the **likely** costs and benefits, in the latter to weigh up **actual** costs and benefits.

Economic analysis is used primarily for the purpose of making decisions about the allocation of funds for a programme. It is often conducted to justify new or continued expenditure on a programme, but may be most reliable when used to make decisions about the relative merits of different programmes and where the same assumptions and proxy measures are used in different analyses.

Economic analysis can also be important for advocacy purposes, demonstrating the value of investments.

Choosing which economic analysis method to use

The two most common forms of economic analysis in programme evaluation are cost benefit and cost effectiveness. The key similarity between cost benefit and cost effectiveness is in the collection of data on costs and the use of common outcomes metrics.

Cost Benefit Analysis (CBA)—allows for a conclusion about whether the programme costs were greater than the benefits provided and requires all benefits to be expressed in monetary terms. In principle at least, there should be no problem comparing the costs and benefits of a programme designed, for example, to provide free coaching sessions with one designed to provide subsidies for exports, as all benefits are expressed in the same dollar terms.

Cost-Effectiveness Analysis—similar to CBA but while costs are still expressed in monetary terms, benefits are expressed in a non-monetary terms using a common outcome metric. This metric will be something that all the

programmes being compared aim to achieve, such as patents filed, or jobs created. As a result, cost effectiveness requires comparisons among families of programmes in order to determine which option is the most cost effective.

Cost Utility Analysis—a type of cost-effectiveness analysis that expresses benefits in terms of a standard unit such as Quality Adjusted Life Years (QALYs).

Value for Money—a term used in different ways, including as a synonym for cost-effectiveness, and as systematic approach to considering these issues throughout planning and implementation, not only in evaluation.

The implications for DIIS are to:

- Use **cost effectiveness** to select the most cost-effective intervention or programme among a family of interventions or programmes aiming at a single common key outcome.
- Use **cost benefit analysis** for single programmes (to assess whether the benefits outweigh the costs or provide an adequate return on investment) or to compare different programmes with different intended outcomes by using a common set of assumptions and measures. This may require a DIIS manual on cost benefit analysis that specifies such things as acceptable assumptions about whose benefits to include, discount rates, proxy measures and common monetisation rates for different outcomes.
    - Where very different programmes have very different intended outcomes for different stakeholders and monetisation rates have a degree of uncertainty, it may be more reliable to limit allocative decision-making based on cost benefit analysis to families of interventions.

Comparing the cost effectiveness of programmes aiming for a common outcome

In cost effectiveness, different programmes are compared on the same outcome measure. This limits comparisons to programs with a similar aim. In health this might be 'quality adjusted life years' or QALYs—a common objective of many health programs being to increase the number of years of healthy life. This common outcome metric allows conclusions about which of a number of programmes delivers the most QALYs relative to their cost, and therefore which are most cost effective. The incremental cost effectiveness ratio (ICER) is commonly used to determine the added benefit of an intervention against the base case—in health economics it is not possible to determine if a programme is cost effective on its own, rather its relative effectiveness in achieving an outcome is determined in comparison to other programmes.[4]

Outside of health it is not easy to identify a key outcome to which very different programmes might aspire. One programme may be strong on stimulating innovation in science and another on commercialisation of science—which key outcome would be chosen for a cost effectiveness analysis? It would be a useful outcome if such a measure like QALYs for industry innovation and science

[4] Value for money analysis using rubrics and qualitative information may allow for evaluative determinations of the worth of single program to be achieved when a cost benefit analysis is not possible.

policy existed, but as this is unlikely in the immediate future, cost effectiveness will be limited to choosing among alternative programmes or interventions with a common aim and single outcome that can be specified.

Determining which benefits to measure in cost benefit analysis

Benefits beyond a single outcome are complicated to measure because decisions are required about what benefits (and potential negative outcomes) will be included in an analysis. This can be a major source of variation in economic analysis. Benefits will often include increased government revenue (from taxes or fines) but should usually extend to include intended beneficiaries and the broader economy—for example if a new wheat strain provides benefits to exporters there should also be consideration of the effect on those who will be disrupted by the new strain. Deciding on who is being counted and who is not, and whether it is only immediate effects or effects that ripple out as the result of former effects (i.e. multiplier effects) are not easy and will vary from analysis to analysis. Similarly, as different benefits accrue over smaller or longer timeframes, future benefits must be attributed to a programme, and this requires that they are 'discounted' at some rate due to the preference for outcomes that occur sooner rather than later.

Cost benefit ratios need common assumptions, monetisation rates and proxy measures

In principle there are no limits on comparing different programmes aiming for different outcomes when calculating cost benefit analysis—all costs and benefits are expressed in monetary terms. However, in practice almost every cost benefit analysis makes assumptions and uses estimates or proxy measures when hard data do not exist. Decisions about assumptions and proxy measures can result in very different cost benefit analyses and lead to different decisions about whether or not to fund a programme even when identical outcomes data is being used—as for example the decision to fund or not fund Project GATE.

In cost benefit analysis every benefit is monetised, that is, an estimated dollar value is placed on each unit of outcome, so for example every job created is assigned a dollar value. The source of this dollar value may vary from study to study. One study may calculate it based on a survey of employers, another on a revealed preferences experimental study, and a third on an academic literature review for a particular industry or time period. It is easy to see that if one cost benefit analysis placed a high value on a job being created while another analysis placed a low value then different results may occur and imply different decisions about the future allocation of funds. These decisions can lead to very different cost benefit ratios for the same programme that have little to do with the inherent merit, value or worth of the programme.

While there will never be one correct method, it is important for cost benefit analyses designed to inform decisions about funding or not funding programmes to use similar monetisation rates and assumptions about benefits and proxy measures to allow for comparison of cost benefit ratios.

5. Social, ethical and political considerations for impact evaluation

All evaluations function in a social and political context, and can raise ethical issues. The focus of this report on choosing appropriate designs and methods points to social and political processes and decisions. While some issues are similar in evaluation more broadly, in this section we briefly highlight issues for impact evaluation particularly in the industry and science policy, and how these may be addressed.

5.1 Evaluation is inherently social and political

The earlier sections of this report have been about approaches, designs and methods for impact evaluation. While the focus has been on the technical dimensions of impact evaluation, it has also involved decisions between alternatives by people with varying degrees of knowledge, interest and influence. In this way contemporary evaluation is a negotiated process which has inherent social and political dimensions. People involved in an evaluation bring their personal views as well as organisational perspectives, and have different degrees of influence and power. More broadly, policy in science and industry is positioned within institutional and political frameworks, and future directions may be strongly contested.

Any evaluation is about a particular programme at a particular time with its own specific stakeholders, functioning in a particular organisational environment and an economic and political context— in part this is what distinguishes evaluation from research. It is a negotiated process with its own micro-politics that involves relations between commissioners, evaluators, programme managers, possible partners, beneficiaries and other stakeholders, functioning in a wider political environment. Evaluations themselves are discrete projects that follow the classic steps of planning, identifying primary intended users, involving stakeholders to some degree, agreeing on relevant impacts and evaluation questions, and reporting findings. Each of these steps involves negotiation and decisions.

Culture is an important social influence on evaluation. It is most apparent with international differences, or in Australia between Indigenous and mainstream interventions. But cultural differences also arise between organisations (whether government agencies, research organisations, universities, or private sector), between industry sectors, and between the varied disciplines involved in industry and science. For impact evaluation, views about the credibility of evidence and the suitability of evaluation designs are shaped by these different social and cultural perspectives.

The core of ethical considerations in evaluation is preventing harm to individuals or groups, and aiming for equity and fairness. Different social, cultural or political perspectives can create ethical dilemmas (for example choosing between two “least harmful” approaches, or dealing with conflicts between credible evidence and political pressure).

5.2 The limits of single methods

Social, ethical and political considerations for impact evaluation are not generally tied to a particular method, but to overall approaches to the

commissioning, conducting, concluding and communicating of evaluation findings. All methods have strengths and weaknesses, can be appropriate or inappropriate for the questions being asked, may be implemented well or poorly and lead to strong or weak, useful or less useful evaluative conclusions. In addition to the appropriate selection and implementation of methods, the key to mitigating risks associated with any particular method is to avoid evaluation that relies on a single method.

Impact evaluation, like any evaluation, will generally be most reliable and valid when it uses a mixed methods approach where the results of one method can be used to validate or extend those of another method. For example, a quantitative causal impact evaluation method may generate evidence of the average effect size associated with an intervention, but it may not be immediately apparent how and when positive effects are most likely to be generated, or identify negative unintended consequences of an intervention. These may be more likely to come from a rigorous analysis of qualitative data, or descriptive impact evaluation methods. In a mixed methods approach the focus is using all available evidence to form an evaluative judgment based on the weight of evidence. This involves the making of claims, provision of evidence, and justifications for making the claim based on the evidence. It is akin to 'the balance of probabilities' or in some cases 'beyond all reasonable doubt' concepts in civil and criminal legal argument rather than the proofs associated with formal logic and mathematics (Schwandt 2015).

5.3 Risks to impact evaluation

Impact evaluations are more useful if there is a measure of independence from key stakeholders and those delivering the government's political agenda. While these groups need to be able to have input into the design, efforts need to be made to ensure they do not unduly influence the results or promote a positive bias.

While no method is perfect, some methods have particular limitations which need to be considered and managed. For example, qualitative 'stories' may be self-serving and highlight small successes even when the overall programme has not been successful. Another risk is being too ambitious, for example being unable construct a control group where there is contamination across groups or moving straight to RCTs to measure impact when other information is required for understanding an intervention, modifying it and targeting it to maximise impact (even if these are not yet measured).

Another potential social, political and ethical feature of all impact evaluation is how the report is used—whether it informs decisions and is able to contribute to the broader knowledge base within the department and externally, or collects dust. Impact evaluations that have agreed strategies for communicating and disseminating evaluations and their results will help apply the lessons more widely across the department and avoid reinventing the wheel or repeating mistakes.

Possible social, political and ethical issues for impact evaluation are listed in table 6 together with a range of strategies that can be drawn upon to reduce these risks.

Table 5.1: Social, political and ethical considerations

Issues and problems	*Possible mitigation strategies*
Social	
There will likely be both winners and losers from decisions following an evaluation	Communicate with stakeholders from the beginning and ensure results are transparent and credible
Bias towards positive findings that result in continuing funding for poor programmes	Avoid incentives for positive reports by limiting high stakes evaluation and scope for gaming the system.
Burden of participation in an evaluation (e.g. grantees providing detailed data)	Build in value for participants, and communicate this to them, such as reporting back aggregate data for benchmarking
Evaluations may not be useful to key stakeholders	Evaluation brief explicitly makes the case for its purpose, benefit and cost. Organisation has evaluation agenda, criteria and priorities.
Equity and distributional effects when only average effect size is reported	Analyse and report on variations in success, especially in terms of equity and targeting issues
Unintended impacts are not being addressed, either positive (and under-valued) or negative	Ensure unintended impacts are included in the scope of the evaluation questions, methods and reports. Engage with diverse informants to identify potential and actual unintended impacts are included in data collection
Political	
Politicians look for simple statements or sound bites that may not reflect evaluation findings; or want simple answers or “facts” for more complex findings	Rise to the challenge of reporting findings in ways that are clear, nuanced and contextualised. Present persuasive evaluative arguments using multiple sources of evidence
Focus of what is evaluated, why and how is driven by unacknowledged and self-serving agendas	Form a reference group and engage with key stakeholders to focus the evaluation on important areas and make transparent decisions about the evaluation focus, design and reporting Ensure technical review of methods chosen and the report (meta-evaluation)
Findings manipulated by vested interests	Use multiple methods that cross-validate or triangulate findings, with balanced evaluative argument. Report notes data manipulation. Ensure technical review of methods chosen and report (meta-evaluation
Benefit to government vs. benefit to community especially in economic analysis	Form a reference group including key community Transparency through publication of reports
Evaluation report withheld from public release without adequate justification	Develop guidance regarding what is appropriate to withhold from public reporting and how this should be managed
Ethical	
Participants in evaluation face harm or loss of privacy or confidentiality	Promote and pro-actively use professional and sector guidelines and standards, such as AES Guidelines for the Ethical Conduct of Evaluation. Be explicit with participants about privacy or confidentiality conditions for people or organisations
How to distribute knowledge, share IP	Develop and negotiate arrangements on a case-by-case basis during evaluation planning

Issues and problems	*Possible mitigation strategies*
	Transparency through publication of reports Stakeholder and peer input and review
Withholding a programme in order to test it	Ensure that 'control groups' are not deprived of an existing intervention unless there is existing evidence casting doubt on its effectiveness.
Evaluation that identifies early implementation difficulties can prematurely curtail promising programme	Develop an evaluation strategy that fits the timeframe for expected impacts, and manage expectations of early evaluations

5.4 Wider pressures on impact evaluation

Different stakeholders will have different ideas about the purposes of the evaluation and the outcomes they wish to measure. They will also have more or less ability to influence the purpose and type of impact evaluation that is undertaken. In addition, there will be an ongoing trade-off between simplicity and comprehensiveness in any impact evaluation.

Perhaps the most serious social, ethical and political consideration for impact evaluation in the context of Commonwealth-funded industry programs is how the choice of evaluation method can influence programme selection and design.

The danger is twofold. The first danger is that only relatively simple problems are developed, for example programs to improve the number of staff with R&D in their position description. There is no evidence to suggest this is a problem for DIIS, quite the contrary, but there is a risk that programmes become less ambitious when they place a greater emphasis on measuring outcomes.

The second danger is that promising programmes are not valued simply because their results cannot be measured while relatively ineffective programmes are valued because aspects of them can be more easily measured. In an attempt to determine the relative value of programmes (i.e. impact evaluation for 'allocation') economists will often be asked to conduct a cost benefit analysis to determine which programme delivers the best value for money. Programmes whose outcomes are not readily 'measureable' will suffer.

Decisions about what methods are most appropriate and how to interpret the results can be contentious. Calls for more use of RCTs in industry evaluation from groups such as NESTA and the Campbell collaboration seek to reduce the complexity of evaluation to the results of RCTs. This is understandable—everyone wants simple answers to questions, but as described above RCTs are not always possible or useful. Different methods will be more appropriate and cost effective in different situations.

Impact evaluations that provide simple answers are readily communicable, but can oversimplify the situation. In this case the results may not have external validity—outcomes achieved in the past may not occur in the future because how and when the programme worked was not adequately understood. They may also lack internal validity if the average impact is very different to the impact on particular sub-groups. Differentiated impact evaluations that acknowledge variations across contexts can have greater validity, but can be difficult to

communicate in political terms. Providing tangible examples of why context matters can be critically important in effective communications.

In conclusion, it is clear that the choice of appropriate designs and methods for impact evaluation will necessarily involve social, ethical and political considerations. The proposed designs of impact evaluations need to be scrutinised from this perspective and their consequences anticipated, with suitable social and political processes and ethical safeguards put in place.

Appendix 1 Types of impacts

CSIRO has established a list of potential economic, environmental and social impacts that evaluators of its programmes can consult as a starting point for considering the wider range of possible impacts.

Table A1: Economic Impacts

Economic Impact	*Definition*
The macro economy	The capability to influence or change at the macroeconomic level i.e. economy-wide impact such as changes in unemployment, national income, rate of growth, gross domestic product, inflation and price levels.
The micro economy	The capability to influence or change the section of the economy that analyses market behaviour of individual consumers and firms in an attempt to understand the decision-making process of firms and households. It is concerned with the interaction between individual buyers and sellers and the factors that influence the choices made by buyers and sellers. In particular, the micro economy focuses on patterns of supply and demand and the determination of price and output in individual markets (e.g. dairy industry).
International trade	The capability to influence or change the international agreements on trade or trade assistance, protection and policy.
Management & productivity	The capability to influence or change the management, management systems or production of products and services. This also includes not only the risk, marketing, profitability and productivity aspects but also sustainability of the production and consumption system.
Measurement standards & calibration services	The capability to influence or change the measurement standards and calibration services for sectors such as agriculture and the environment, defence, manufacturing and the service Industries.
Economic frameworks & policies	The capability to influence or change economic systems and policies, for example, the taxation, government expenditure systems, the Carbon and Emissions Trading Scheme and ecological economic systems.
New products or services	The capability to develop new products and services, through technological and organisational innovations, in the following areas: · Plant · Animal · Minerals · Energy · Manufacturing · Construction · Transport · Information and communications · Commercial services and tourism · Sustainable products and services

Source: CSIRO Impact Evaluation Guide (2014, p. 22-24)

Table A2: Environmental Impacts

Environmental Impact	*Definition*
Air quality	The variety and connections between plant and animal life in the world or in a particular habitat. Focus on plants and animals within an area and how they interact with each other as well as with other elements such as climate, water and soil. Also the ecosystem services provided to protect ecosystems and biodiversity.
Climate & climate change	Focus on atmospheric, land and ocean patterns and the changes in these over time.
Disaster mitigation	Steps taken to contain or reduce the effects of an anticipated or already occurred disastrous events (such as drought, flood, lightning, various levels and types of storms, tornado, storm surge, tsunami, volcanic eruption, earthquake, landslides).
Energy generation and use	The creation of energy using various technologies and processes and its effect on the environment. The effect of the use of created energy and the benefits of efficiency measures.
Land quality and management	Land use and management with effects on soil and the surrounding environment. Actions taken to rehabilitate the land after production processes.
Water quality and management	Water systems, availability, quality, access and management.
Oceans and marine environments	Includes quality and quantity of marine and other ocean resources.
Sustainable industry development	Features of sustainable industry development are: energy efficiency, resource conservation to meet the needs of future generations, safe and skill-enhancing working conditions, low waste production processes, and the use of safe and environmentally compatible materials.

Source: CSIRO Impact Evaluation Guide (2014, p. 22-24)

Table A1: Social Impacts

Social Impact	*Definition*
Life & Health	The capability to be alive and healthy.
Equity and equality	Equity involves trying to understand and give people what they need to enjoy full, healthy lives. Equality, in contrast, aims to ensure that everyone gets the same things in order to enjoy full, healthy lives. Both aim to promote fairness and justice.
Social connectedness	Social connectedness refers to the relationships people have with others and the benefits these relationships can bring to the individual as well as to society.
Standard of living	The degree of wealth and material comfort available.
Safety	Safety means protection from dangerous materials, products or processes.
Security - Civil	Physical and psychological protection against others.
Security - Military	Protection from an actual or perceived threat from an internal or external combatant that will affect the greater society.
Security - Cyber	Information security as applied to computers and networks.
Social consciousness	Social consciousness is an awareness or realisation shared within a society. An individual with an acquired social consciousness derives his or her viewpoint from the mainstream culture.
Social licence to operate and community confidence	Ongoing approval or broad social acceptance and achieving successful integration between industry/government/NGOs and community.
Resilience (community and industry)	The capacity to recover from a disturbance.

Source: CSIRO Impact Evaluation Guide (2014, p. 22-24)

Appendix 2 Examples of indicators for industry programmes

Table B1: Examples of indicators for industry programmes

Programme	*Key performance indicators*
Australia-China Science and Research Fund	Number of collaborative research projects completed that reported strengthened international relationships.
Cooperative Research Centres Program.	Total value of grants and contracts.
Commercialisation Australia Program	Value of total investment in commercialisation of projects receiving Commercialisation Australia grants. Number of customers that report they are working towards successful commercialisation of supported projects. Number of respondents who met projects milestones and who are achieving successful commercialisation outcomes.
Research and Development (R&D) Tax Incentive	Number of entities registering R&D expenditure with AusIndustry in order to claim the tax concession through their annual tax returns. R&D expenditure registered with AusIndustry in order to claim the tax incentive or tax concession through their annual tax returns.
Green Building Fund	Reductions in greenhouse gas emissions from completed Green Building Fund projects, expressed as kilotonnes of carbon dioxide (equivalent) per annum.
Small Business Advisory Services	Increased facilitation, advice, support and assistance provided to businesses and industry by increasing the number of services utilised, increasing the number of SMEs developing new and sustainable capabilities (knowledge, tools, expertise) and increasing the number of assisted clients implementing advice. Number of businesses assisted through provision of advice, referrals and services to improve capabilities: Number of services provided to small businesses through SBAS.
Community Energy Efficiency Program	Improved energy management practices within councils, organisations and the broader community through the Community Energy Efficiency Program.

Appendix 3 Implications of complication and complexity for impact evaluation

The following table sets out some possible implications of complication and complexity for seven different aspects of programmes and policies.

Table C1: Characteristics and implications of complicated and complex aspects of programmes and policies for impact evaluation

Aspect	*Simple*	*Complicated*	*Complex*
1. Focus	Single set of objectives	Different objectives valued by different stakeholders Multiple competing imperatives Objectives at multiple levels of a system	Emergent objectives and imperatives
Implication	Impacts to be included can be readily identified from the beginning	Need to identify and gather evidence about multiple possible changes Need an agreed way to weight or synthesise results across different domains	Need nimble impact evaluation systems that can gather adequate evidence of emergent intermediate outcomes or impacts
2. Governance	Single organisation	Multiple organisations (which can be identified) with specific, formalized responsibilities	Emergent organizations working together in flexible ways
Implication	Primary intended users and uses easier to identify and address	Likely to need to negotiate access to data and ways to link and co-ordinate data Might need to negotiate parameters of a joint impact evaluation, including negotiating scope and focus	Need nimble impact evaluation systems that can gather evidence about the contributions of emergent actors and respond to the different ways they value intended and unintended impacts
3. Consistency	Standardized – one-size-fits-all service delivery	Adapted – variations of a programme planned in advance and matched to pre-identified customer profiles	Adaptive – evolving and personalised service delivery that responds to specific and changing needs
Implication	Quality of implementation should be investigated in terms of compliance with 'best practice'	Quality of implementation should be investigated in terms of compliance with the practices prescribed for that type of situation	Quality of implementation should be investigated in terms of how responsive and adaptive service delivery was
4. Necessity	Only way to achieve the intended impacts Works the same for everyone	The intervention is one of several ways of achieving the impacts, and these can be identified	Possibly one of several ways of achieving the intended impacts

Table C1: continued

Aspect	*Simple*	*Complicated*	*Complex*
Implication	Counterfactual reasoning appropriate	Counterfactual reasoning not appropriate as it does not accept a causal relationship between the intervention and the impacts unless they would not have occurred in the absence of the intervention	Counterfactual reasoning not appropriate as it does not accept a causal relationship between the intervention and the impacts unless they would not have occurred in the absence of the intervention
5. Sufficiency	Sufficient to produce the intended impacts Works the same for everyone	Works only in specific contexts which can be identified (e.g. implementation environments, participant characteristics, support from other interventions)	Works only in specific contexts which are not understood and/or not stable
Implication	Counterfactual reasoning appropriate Reasonable to ask 'Does it work?'	Impact evaluation question needs to be 'For whom, in what circumstances and how does it work?' Counterfactual reasoning only appropriate if the causal package of supportive context and other activities can be identified and included	Impact evaluation question needs to be 'For whom, in what circumstances and how does it work?' Counterfactual reasoning not appropriate as the causal package of supportive context and other activities is changing and/or poorly understood and cannot be adequately identified
6. Change trajectory (how impact variables will change over time – for example, straight line of increase, or J curve)	Simple relationship that can be readily predicted and understood	Complicated relationship that needs expertise to understand and predict	Complex relationship (including tipping points) that cannot be predicted or understood except in retrospect
Implication	Measurement of change can be done at a convenient time and confidently extrapolated	Timing of the measurement of changes should be undertaken when it will be most meaningful – expert advice will be needed	Changes will need to be measured at multiple times as the change trajectory cannot be predicted
7. Unintended impacts	Readily anticipated and addressed	Likely only in particular situation; need expertise to predict and address	Cannot be anticipated, only identified and addressed when they occur
Implication	Need to draw on previous research and common sense to identify potential unintended impacts and gather data about them	Need advice from experts about potential unintended impacts and how these might be identified	Need to include a wide net of data collection that will catch evidence of unexpected and unanticipated unintended impacts

Source: adapted from Funnell and Rogers (2011, p. 90-91)

Appendix 4 Results Based Accountability

A particular approach to monitoring, which can provide useful data for impact evaluation, is Results Based Accountability (RBA) developed by Mark Freidman. RBA works by monitoring outcomes at the population level and the three programme levels. It looks for patterns in the data to tell a story. Rather than a pre and post-test it includes ongoing data collection. The logic is that comparisons with a programme's own history are more meaningful than trying to make comparisons with other programmes—although this is still possible. This approach requires data on the problem to be addressed as well as data on the programmes being implemented. Data on programmes should be informed by programme logic and be set at three levels. Data items are usually drawn from administrative or survey data. At each level a small number of data items (1-5) reflect key questions around the quantity or quality of the interaction. It is also often useful to provide a free text data item about comments or suggestions wherever possible (especially if the data is being collected using a survey method) as this may be useful for more in-depth exploratory analysis.

There are three levels of performance that are monitored in RBA and these are associated with three questions. This is referred to as performance accountability.

- At the first level the question is, how much did we do? This is usually simple data counting outputs, such as number of firms applying for the MTP. This is something over which DIIS has a relative degree of control, is easy to measure but is ultimately only an output rather than an outcome.
- At the second level the question is, how well did we do it? This is also something over which DIIS has some control, may be moderately easy to assess (such as with a satisfaction questionnaire) but again it is not as important as the effects.
- At the third level the question is; is anyone better off? This is something that is hard to control and difficult to assess but is ultimately very important. Note the question is not whether we made a difference. This is because it is very hard to measure an outcome that is attributable to a program. This may be self-reported data on how workplace practices have changed among firms participating in the MTP.

RBA also monitors the problem/condition which the programme exists to address. This is referred to as population accountability, and is about observed changes in the population for whom the programme exists. It may as wide as the whole population or just a subset, so for example it may be all the manufacturing firms or just those in a particular location or industry. The point is that the population includes both those participating in the programme and those not participating—obviously the more effective the programme and the more people participating the better the population level outcomes will be.

RBA seeks to link performance and population accountability and provide programme administrators with a means of telling the story of a program. It provides an aspirational target and reminder of why the programme exists in the

form of data on the problem being addressed. In the best case scenarios it allows a performance story to talk about ‘turning the curve’ or how the programme affected population factors. The curve relates to long term trends. Long term trend provides important context for understanding the value of a program. An effective programme may reduce the rate at which a problem is increasing as much as it may increase the rate at which it is improving. As with contribution analysis it is useful to monitor other key factors known to affect population level outcomes, so for example in the case of the MTP tracking the currency exchange rate may be a very important external factor that can be used to describe why a programme may be effective, but not changing the problem being addressed (e.g. manufacturing going out of business). The focus is on bringing together data from different levels of the programme and problem being addressed to make an evidence-based judgment about the value of the program, while at the same time allowing for real time monitoring to flag areas for further exploration, and adoption of the programme in light of evidence of its effectiveness without waiting for a full-scale evaluation.

RBA is designed for individual programs but can be modified to provide an agency wide monitoring framework. This can be used to enable ongoing monitoring of performance, setting of DIIS benchmarks and the comparison of various DIIS interventions or services with each other. This type of work is more often associated with agency accountability and initiatives such as Agency Corporate Reports under the PGPA Act. An example based on this approach is the Data Exchange Framework and SCORE method developed by ARTD in 2014 for the Commonwealth Department of Social Services.

Appendix 5 Examples of causal inference methods

The following case studies illustrate how different forms of impact evaluation were used for different purposes and contexts, why they were chosen and how they were interpreted.

Example

Cross-case analysis for explaining impact

This Industry Canada Evaluation of the Community Futures Program demonstrates how cross-case analysis was used to provide context around whether/how outcomes occurred and how activities and outcomes contributed to the intended outcomes. These cases were selected according to a range of criteria.

> The case studies were undertaken to answer evaluation questions pertaining to the achievement of immediate, intermediate and ultimate outcomes. In addition, they also provided information regarding the operational environment of the Community Futures Development Corporation (CFDC) sites and the process by which projects are undertaken; illustrative examples that support the programme theory (i.e., not only whether outcomes have occurred, but how activities and outputs contribute to the intended outcomes); challenges and lessons learned.
>
> To obtain representative results from the cross case analysis, and to have a range of illustrative examples to support other lines of evidence, five CFDCs were selected for case studies. The following criteria were applied for case study selection: geographical representation (representing both regions of Northern Ontario), community economic development activities, First Nation population, level of unemployment, population density, strategic community planning, support to community-based projects and special initiatives, materiality (i.e., the dollar value of projects), and nature and extent of partnerships with other community organizations.

Industry Canada 2015b, Evaluation of the Community Futures Program, https://www.ic.gc.ca/eic/site/ae-ve.nsf/eng/h_00351.html, p.10–11).

Qualitative Comparative Analysis for analysing combinations of attributes

In this paper, Greckhamer et al. (2008) illustrate how Qualitative Comparative Analysis (QCA) 'may be used to study the sufficiency of combinations of industry, corporate, and business-unit attributes for the occurrence of superior or inferior business-unit performance. Rather than trying to understand the relative independent contribution of each of the various industry, corporate, and business-unit level effects to performance, this research approach instead examines what combinations of industry, corporate, and business-unit attributes are necessary and/or sufficient for superior or inferior performance. For example, instead of ''How much does corporate strategy matter?'' the pertinent question becomes ''How do corporate factors combine with industry and business-unit factors to matter?'' for business-unit performance. Therefore, QCA allows for the investigation of the complex interdependencies among industry, corporate, and business-unit attributes that potentially underlie business-unit performance.'

Using QCA, they are able to demonstrate 'substantial interdependence among industry, corporate, and business-unit attributes in determining business-unit performance. Moreover, they illustrate the causal complexity that underlies the determination of performance. The results suggest that two or more different combinations of attributes can be sufficient for attaining the same outcome and that any particular attribute may have different and even opposite effects depending on the presence or absence of other attributes.'

Greckhamer, T, Misangyi, V, Elms, H & Lacey, R 2008, 'Using Qualitative Comparative Analysis in Strategic Management Research', Organizational Research Methods, vol. 11, no. 4, p. p. 720.

Realist analysis — understanding relations between context and mechanisms for achieving impact

A study of public involvement in research (Evans et al, 2014) used a realist evaluation.

> The aim was to identify the contextual factors and mechanisms that are regularly associated with effective public involvement in research. The objectives included identifying a sample of eight research projects and their desired outcomes of public involvement, tracking the impact of public involvement in these case studies, and comparing the associated contextual factors and mechanisms.

Example

The research design was based on the application of realist theory of evaluation, which argues that social programmes are driven by an underlying vision of change – a 'programme theory' of how the intervention is supposed to work. The role of the evaluator is to compare theory and practice. Impact can be understood by identifying regularities of context, mechanism and outcome. Thus the key question for the evaluator is 'What works for whom in what circumstances . . . and why?' (Pawson 2013). We therefore planned a realist evaluation based on qualitative case studies of public involvement in research.

Conclusions: A revised theory of public involvement in research was developed and tested, which identifies key regularities of context, mechanism and outcome in how public involvement in research works. Implications for future research include the need to further explore how leadership on public involvement might be facilitated, methodological work on assessing impact and the development of economic analysis of involvement

(Evans et al. 2014, p. v-vi).

Comment

When is a realist analysis appropriate? A realist evaluation design is well suited to assess how interventions in complex situations work because it allows the evaluator to deconstruct the causal web of conditions underlying such interventions.

A realist evaluation yields information that indicates how the intervention works (i.e., how it leverages generative mechanisms) and the conditions that are needed for a particular mechanism to work (i.e., specification of contexts). In many cases programmes work by the way activities or interventions introduce resources into a situation that leads to new decisions being made by programme participants and, thus, it is likely to be more useful to policymakers than other types of evaluation which assume that programs somehow 'work' on their own irrespective of the people involved.

As with any evaluation, the scope of the realist evaluation needs to be set within the boundaries of available time and resources. Using a realist approach to evaluation is not necessarily more resource or time-intensive than other theory-based evaluations, but it can be more expensive than a simple pre-post evaluation design (BetterEvaluation) and requires a depth of thought about what actually causes change that is not always apparent in more simplistic approaches that lead to an average effect size or cost benefit ratio.

Evans, D, Coad, J, Cottrell, K, Dalrymple, J, Davies, R, Donald, C, Laterza, V, Long, A, Longley, A, Moule, P, Pollard, K, Powell, J, Puddicombe, A, Rice, C & Sayers, R 2014, 'Public involvement in research: assessing impact through a realist evaluation', Health Services and Delivery Research, vol. 2, no. 36, pp. 1–128.

Quasi-Experimental Research Designs

Difference-in-difference

Falck et al. (2010) evaluate a cluster programme introduced in Germany in 1999, the Bavarian High Technology cluster initiative. This programme aimed to increase innovation and competitiveness in the region of Bavaria by stimulating cooperation between science, business, and finance in five target industries. The main activity was to improve the supply of joint research facilities. The authors used a difference-in-difference-in-differences estimator (comparing the innovation performance of target and control firms in Bavaria as well as firms in other German States, both before and after the policy). The authors found that the initiative increased the probability that firms in a target industry would innovate by 4.5 per cent to 5.7 per cent (depending on the indicator of innovation used). Interestingly, at the same time, research and development (R&D) spending fell by 19.4 per cent on average for firms in the target industries, although the use of external know-how, cooperation with public scientific institutes, and the availability of suitably qualified R&D personnel increased. The authors suggest that the increase in innovation in the presence of falling R&D can be explained by improvements in R&D efficiency consequent on greater access to complementary inputs such as qualified R&D personnel.

In Warwick, K. and A. Nolan (2014), 'Evaluation of Industrial Policy: Methodological Issues and Policy Lessons', OECD Science, Technology and Industry Policy Papers, No. 16, OECD Publishing, http://dx.doi.org/10.1787/5jz181jh0j5k-en, p.37.

Example

Matched comparison group

In matched comparison groups, groups that are expected to be comparable are created based on observable characteristics of participants that are thought to effect outcomes.

Example

The 2014 Australian Industry Report describes the use of propensity score matching (which it refers to in the text as 'matching' in an analysis of the pre- and post-programme financial performance of participants and similar non-participants:

> In 2013, the department commissioned the ABS to analyse the financial performance of businesses that received a business review and/or a tailored advisory service grant from the former Enterprise Connect programme. The study assessed the post-programme (up to five years) performance of firms that received a business review during the 2007–08 to 2010–11 period.
>
> Data and methodology
>
> The financial indicators analysed were revenue, total wages and proxies for value-added and gross operating profit. The statistical analysis was twofold. The pre-programme performance of participants was firstly analysed against their own post-programme performance. Secondly, participants' performance was compared with that of a set of non-participants with similar characteristics (control group).
>
> A comparison of the pre- and post-programme performance of the participants alone could generate misleading conclusions due to factors such as overall macroeconomic conditions affecting both participants and non-participants and selection bias of the participants. For example, it is more likely in a period of high economic growth that revenues would increase irrespective of programme participation status. Additionally, a well-managed business may be more likely to seek opportunities for improvement, and hence be more likely to apply for a business review. The technique that was used to abstract from these factors is known as matching.
>
> Matching is required when some businesses are exposed (treated) to a policy and others are not (untreated). The aim is to construct a sample of treated and untreated businesses that are as similar as possible in terms of observable characteristics prior to the policy intervention. The impact of a programme can then be analysed as the mean difference in growth of the treated and untreated firms (the difference in differences). That is, we observe how participants performed, and we estimate how they would have performed (the counterfactual) by using the performance profile of non-participants that were similar to participants in other respects prior to the programme.
>
> Variables used for matching were wages, export status, R&D status, degree of foreign ownership, Australian and New Zealand Standard Industrial Classification (ANZSIC) Subdivision, main state of activity and type of legal organisation. The selection of these variables was guided by economic theory, previous empirical results and programme entry criteria.
>
> Upon completion of the review, businesses could apply for a small grant to implement recommendations from the review and a number of businesses received this grant. The ABS analysis made a distinction between review-only and review-and-grant businesses since review-and-grant was considered a more intensive 'treatment' than review-only. Two industries (Manufacturing and Professional, Scientific & Technical Services) and two levels of participation (review-only and review-and-grant) generated four groups of cohorts matched separately.
>
> The data used in this analysis were constructed by merging the programme administrative data, ATO data and ABS data. ATO and ABS data were accorded with the programme's

Example

> administrative data using Australian Business Numbers (ABNs) as unique identifiers. Businesses were excluded from the analysis if ABNs were missing in the programme database, if the business was part of a GST grouping, or if the business was part of a complex business that operates across multiple industries, or had multiple ABNs ().

Department of Industry 2014, Australian Industry Report, http://www.industry.gov.au/Office-of-the-Chief-Economist/Publications/Documents/Australian-Industry-Report.pdf, p. 176–177.

Comments

There are two main approaches, the first is matching each treatment individual with a control individual based on recorded observable data (such as may be kept in an administrative data set, such as size of business, location, profit, history, patents filed, survey responses about innovation etc.). The second main approach is using regression statistics to weight aggregate treatment and control group data. There are benefits and drawbacks of both approaches. The latter allows for the full variation in factors affecting outcomes to be analysed in regression models, while the former allows for the combined impact of contexts and mechanism to be included in the measurement of impact and is therefore more likely to be accurate.

Comparison matching requires statistical computations and is best conducted using statistical programs such as Stata or SPSS. It may be useful to involve an experienced statistician, depending on levels of staff knowledge.

Comparison matching demands a deep understanding of the observable covariates that drive participation in an intervention and requires that there is substantial overlap between the propensity scores of those subjects or units which have benefited from the programme and those who have not; this is called the 'common support'. If either of these two factors are lacking, it is not a suitable methodology for estimating causal effects of an intervention.

It also requires a large sample size in order to gain statistically reliable results. This is true for many methodologies but is particularly true for PSM due to the tendency to discard many cases which cannot be matched.

Matched comparisons do not represent a panacea in avoiding selection bias (where participants and non-participants are systematically different in ways that affect impacts) — because this method matches only on observed information, it may not eliminate bias from unobserved differences between treatment and comparison groups.

It is important to understand the trade-offs between reducing bias and reducing standard errors that arise when choosing the specifications of the matching algorithms. For example, when choosing the calliper size for the radius matching, if the calliper size is too large there is a risk that very dissimilar individuals will be matched, while if the calliper size is too small the sample size may become too small to obtain statistically convincing results. Similarly, for neighbour matching, choosing multiple neighbours decreases bias, relative to single neighbour matching, but increases standard errors due to the smaller sample size caused by a more stringent specification. Such trade-offs exist in each matching algorithm.

Instrumental Variables

> Oosterbeek et al. (2010) evaluate the impact of the student mini company scheme (SMC) on students' entrepreneurial competencies and intentions using an instrumental variable (IV) approach and a difference in difference framework. They drew their data from a vocational college in the Netherlands that offered the scheme, part of the Junior Achievement Young Enterprise programme, at one of two of its locations providing similar Bachelor's programme. The latter provided a natural control group but since students may have self-selected into different school locations, location choice (and thus treatment) may account for changes in outcome variables due to unobserved differences between the students of both locations.
>
> Entrepreneurial competencies were measured using the Entrepreneur Scan Test, a validated self-assessment test based on 114 items which are converted (loaded) into 10 traits and skills identified as important in the entrepreneurship literature. While traits may be invariant to the programme, skills such as market awareness for example, can be learned and improved through participation and changes are more likely to be observed. The test and a survey that included questions on background and likelihood of becoming an entrepreneur (intention) were administered at the start of the programme and again at the end both in the treatment location and the control. The results show that the SMC participation did not impact on entrepreneurial intention nor stimulate the skills of students. The effect on entrepreneurial intention was negative and significant. In other words, entrepreneurial intention in the control group was higher than for those in the SMC programme. In addition the effects on students' self-assessed entrepreneurial skills and traits were negative and not significantly different from zero. Oosterbeek et al. (2010) suggest that that the SMC programme may have had a discouraging or 'sorting' effect as in participating, students were able to form a more realistic assessment of both themselves as well as what it takes to be an entrepreneur.

In Rigby, J and Ramlogan, R 2013, The impact and effectiveness of entrepreneurship policy, Nesta Working Paper Series, No. 13/01
https://www.nesta.org.uk/sites/default/files/the_impact_and_effectiveness_of_entrepreneurship.pdf p.13–14.

Bryson, Dorsett & Purdon note:

> The IV method is possible when a variable can be identified that is related to participation but not outcomes. This variable is known as the 'instrument' and it introduces an element of randomness into the assignment which approximates the effect of an experiment. Where it exists, estimation of the treatment effect can proceed using a standard instrumental variables approach (2002, p.9).
>
> Bryson, Dorsett & Purdon 2002, The use of propensity score matching in active labour market policies, Working Paper No. 4, Department for Work and Pensions.

Matched comparison

'While not a perfect experiment, comparison of programme participants with matched comparison groups approximates equivalency' (Royse 1991; Rossi & Freeman 1993). Most examinations of microenterprise outcomes have been before-versus-after comparisons rather than with versus without (Schreiner 1999a). Without a control group, 'although the analyst can observe users both before and after a MEP, the analyst cannot observe users both with and without a MEP. It [before-versus- after evaluation] ignores that changes in outcomes might have happened even without a MEP' (Schreiner 1999a, p. 20). But this study compares three similar groups of workers to examine whether they diverge significantly in economic outcomes over time. It thus makes it possible to estimate whether microenterprise assistance programs have a programme effect.

Low-income self-employed and wage workers were identified in the Panel Study of Income Dynamics (PSID) and matched as closely as possible to the programme participants on 1991 data. Matching focused on six demographic factors—age, education, race, gender, marital status, and presence of young children. Matching is carried out in the aggregate (Rossi & Freeman 1993). That is, individuals are not matched one to one on every factor, but the overall distributions on each variable are made to correspond between groups.

However, because the number of self-employed Latinos (in PSID) reporting data in 1991 and 1995 was limited, Latinos and African Americans are grouped together as non-whites. Further, because self-employed workers drawn from PSID included a higher frequency of male versus female self-employed people, a greater proportion of men were drawn from PSID. Significant differences across groups on any of the matching factors are used

Instrumental Variables

as a covariate in analyses. While matched groups are very similar, sample selection does present study limitations. Unobserved differences may exist between groups that cannot be controlled. For example, education level is used as a proxy for human capital in the matching process. However, human capital may include relevant business skills and qualities that are not captured in the measure of education level.

Sanders 2002, ‘‘The Impact of Microenterprise Assistance Programs: A comparative study of program participants, nonparticipants, and other low-wage workers’, Social Service Review, vol. 76, no. 2, pp 325–326.

Regression discontinuity

Regression discontinuity (RD) is a powerful and potentially very useful approach to impact evaluation when both of the following conditions are met: administrative data sets exist and there are relatively arbitrary eligibility criteria for programme participation. For example, if business with 20 or more employees is eligible for a subsidy, and the subsidy was introduced on certain date, then changes in outcomes of interest with firms of 20 employees can be compared to firms with 19 employees. If the trends in the outcome of interest diverge over the implementation time frame, but there were no other changes introduced at the time based on the same criteria, then changes can be attributed to the intervention.

RD can be used in situations where:

a continuous eligibility index (e.g. test scores, age etc.) is used to rank the population

a clearly defined cut-off point that determines eligibility for treatment is part of the programme design and

the same cut-off score has not been used in determining eligibility for other treatments.

Be aware that RD provides limited external validity as results are only generalizable around the cut-off - provision of a service might make more or less difference to people who are further away from the cut-off point.

Example

Cambodia: World Bank Girls Secondary School Scholarship Fund. Two levels of scholarship for poorest ($60) and next poorest ($45) girls. The evaluation aimed to assess programme impacts, particularly the effectiveness of providing larger scholarships to poorer girls. Regression discontinuity design comparing groups above and below the eligibility cut-off point for the maximum $60 scholarship. Richer data set also permitted analysis of learning, child labour and inter-household issues.

Bamberger, M., & Kirk, A. (2009), ‘Making Smart Policy: Using Impact Evaluation for Policy Making—Case Studies on Evaluations that Influenced Policy’, Doing Impact Evaluation, 14. Chicago, p.36.

Example

Comparison between Australian firms receiving government assistance for innovation and non-assisted firms.

DIISRTE provided the ABS with ABNs of companies in the programmes under analysis. The ABS matched these ABNs with Australian Taxation Office (ATO) data to obtain financial information. The ABS then created a dataset of similar businesses that did not receive DIISRTE assistance. In order to make valid comparisons with assisted firms, the ABS established ’non-assisted groups’ of firms with similar characteristics by size and industry.

Firms that receive assistance from DIISRTE programs have different characteristics from the general business population, so it was important that the control group of non-assisted firms had a similar profile to allow accurate comparisons. The ABS carefully investigated this issue and undertook appropriate statistical analyses to ensure the comparisons were valid. ABS used the approach empirically tested by Zimmerman (1998) where non-normally distributed assisted and non-assisted samples were adjusted by trimming the top and bottom of the distribution by one per cent to enable a more rigorous statistical comparison.

The analysis was conducted by comparing changes in dollar values for turnover, value added, profits and wages between the two periods. The ABS used the following statistical methods to compare DIISRTE assisted programme firms with non-assisted firms (ABS 2011): significance tests (t-test and paired t-test); ordinary Least Square Regression method; and comparison of difference in difference. The results were broadly consistent across methods giving additional confidence in the findings. The statistical methods were determined by the ABS and externally peer reviewed by Dr Robert Clark, School of Mathematics and Applied Statistics, University

Instrumental Variables

of Wollongong.

The data showed that the overwhelming majority of firms receiving assistance outperformed non-assisted firms. However, not all these results could be proven to be statistically significant at the 95 per cent confidence level. This may be the result of a number of factors including the way in which the sample was selected, the methodology used and the presence of unobserved factors that cannot be controlled during the study. Given these factors, it is important to note that differences which are not statistically significant may still indicate positive results for assisted firms.

Department of Industry, Innovation, Science, Research and Tertiary Education (DIISRTE) 2012, Data Matching Paper - A comparison between Australian firms receiving government assistance for innovation and non-assisted firms (p.5)

Appendix 6 Sources for literature review

Bamberger, M & Kirk, A 2009, 'Making Smart Policy: Using Impact Evaluation for Policy Making—Case Studies on Evaluations that Influenced Policy', *Doing Impact Evaluation*, vol. 14.

Benus, J, McConnell, S, Bellotti, J, Shen, T, Fortson, K & Kahvecioglu, D 2008, *Growing America Through Entrepreneurship: Findings from the evaluation of Project GATE*, U.S. Department of Labor, Employment and Training Administration, Washington, DC.

BetterEvaluation 2015, http://betterevaluation.org/

Bonell, C, Fletcher, A, Morton, M, Lorenc, T & Moore L 2012, 'Realist Randomised Controlled Trials: A New Approach to Evaluating Complex Public Health Interventions', *Social Science & Medicine*, vol. 75, no. 12, pp. 2299–2306.

Bryson, A, Dorsett, R & Purdon, S 2002, *The use of propensity score matching in the evaluation of active labour market policies*, Working Paper No. 4, Department for Work and Pensions.

Coalition for Evidence-Based Policy 2015, *What Works in Social Policy? Findings From Well-Conducted Randomized Controlled Trials*, Coalition for Evidence-Based Policy, http://evidencebasedprograms.org/ .

Commonwealth of Australia 2010, *Inspiring Australia: A national strategy for engagement with the sciences*, Canberra.

CSIRO 2014, *Impact Evaluation Guide* (unpublished).

Deaton, A 2010. Instruments, randomization, and learning about development. Journal of economic literature, 424-455.

Department of Finance 2015, *Resource Management Guide No.131: Commonwealth Resource Management Framework Companion,* Canberra.

Department of Industry 2014, *Australian Industry Report*, http://www.industry.gov.au/Office-of-the-Chief-Economist/Publications/Documents/Australian-Industry-Report.pdf

Department of Innovation, Industry, Science and Research 2010, *Inspiring Australia: A national strategy for engagement with the sciences,* Canberra, http://www.industry.gov.au/science/InspiringAustralia/Documents/InspiringAustraliaReport.pdf

Department of Treasury and Finance 2014, *Investment Lifecycle and High Value/High Risk Guidelines,* Victoria State Government, http://www.dtf.vic.gov.au/Investment-Planning-and-Evaluation/Understanding-investment-planning-and-review/What-are-the-investment-lifecycle-and-high-value-high-risk-guidelines

DIISRTE 2012, Data Matching Paper - A comparison between Australian firms receiving government assistance for innovation and non-assisted firms.

Evans, D, Coad, J, Cottrell, K, Dalrymple, J, Davies, R, Donald, C, Laterza, V, Long, A, Longley, A, Moule, P, Pollard, K, Powell, J, Puddicombe, A, Rice, C & Sayers, R 2014, 'Public involvement in research: assessing impact through a realist evaluation', Health Services and Delivery Research, vol. 2, no. 36, pp. 1–128.

Friedman, M 2009, Trying Hard Is Not Good Enough, BookSurge Publishing, Charleston, SC.

Funnell, S, & Rogers, P 2011, Purposeful Program Theory: Effective Use of Theories of Change and Logic Models, John Wiley & Sons, Qld.

Glouberman, S 2001, Towards a new perspective on health policy, Canadian Policy Research Networks, Ottawa.

Glouberman, S & Zimmerman, B 2002, Complicated and Complex Systems: What Would Successful Reform of Medicare Look Like?, Commission on the Future of Health Care in Canada, http://www.hcsc.gc.ca/english/pdf/romanow/pdfs/8_Glouberman_E.pdf

Greckhamer, T, Misangyi, V, Elms, H & Lacey, R 2008, 'Using Qualitative Comparative Analysis in Strategic Management Research', Organizational Research Methods, vol. 11, no. 4, pp. 695-726.

Guijt, I 2008, Seeking Surprise: Rethinking monitoring for collective learning in rural resource management, published PhD thesis, Wageningen University, Wageningen.

Haynes, L, Service, O, Goldacre, B & Torgerson, D 2012, Test, Learn, Adapt: Developing Public Policy with Randomised Controlled Trials, Cabinet Office Behavioural Insights Team, London, https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/62529/TLA-1906126.pdf

Impact Evaluation Working Group 2012, 'Dare to measure', Evaluation designs for industrial policy in the Netherlands, http://www.cpb.nl/en/article/dare-measure-evaluation-designs-industrial-policy-netherlands.

IMPAQ 2008, Growing America Through Entrepreneurship: Findings from the Evaluation of Project GATE

IMPAQ 2009, Growing America Through Entrepreneurship: Final Evaluation of Project GATE

Industry Canada 2012, Evaluation of the Strategic Activities Program, https://www.ic.gc.ca/eic/site/ae-ve.nsf/eng/03583.html

Industry Canada 2013, Evaluation of Mandatory Counselling, https://www.ic.gc.ca/eic/site/ae-ve.nsf/eng/h_00351.html

Industry Canada 2015a, Evaluation of Industry Canada's contribution to CANARIE, https://www.ic.gc.ca/eic/site/ae-ve.nsf/eng/h_00351.html

Industry Canada 2015b, Evaluation of the Community Futures Program, https://www.ic.gc.ca/eic/site/ae-ve.nsf/eng/h_00351.html

Jensen, P 2015, Randomised controlled trials and industry programme evaluations, Department of Industry and Science, http://www.industry.gov.au/Office-of-the-Chief-Economist/Publications/Pages/Randomised-controlled-trials-and-industry-programme-evaluations.aspx

Jones, M, Castle-Clarke, S, Manville, C, Gunashekar, S & Grant, J 2013, Assessing research impact: An international review of the Excellence in Innovation for Australia Trial, RAND Europe, Cambridge, http://www.rand.org/content/dam/rand/pubs/research_reports/RR200/RR278/RAND_RR278.pdf

Kurtz, G & Snowden, D 2003, 'The new dynamics of strategy: Sense-making in a complex and complicated world', IBM Systems Journal, vol. 42, no. 3, pp. 462-483.

Mayne, J & Stern, E 2013, Impact Evaluation of Natural Resource Management Research Programs: A Broader View, ACIAR Impact Assessment Series Report No. 84, Australian Centre for International Agricultural Research, Canberra.

National Bureau of Economic Research 2012, 'Behind the Gate Experiment: Evidence on Effects of and Rationales for Subsidized Entrepreneurship Training', NBER Working Paper Series, No. 17804

Owen, J & Rogers, P 1999, Program evaluation : forms and approaches, Allen & Unwin, St Leonards, NSW.

Patton, M 2008, Utilization-focused evaluation, 4th edn, Sage Publications, Thousand Oaks, CA.

Pawson, R 2013, The Science of Evaluation: A Realist Manifesto, SAGE Publications Ltd, London.

Perrin, B 2012, 'Linking Monitoring and Evaluation to Impact Evaluation', Impact Evaluation Notes, No. 2.

Ravallion, M 2009, 'Should the randomistas rule?', The Economists' Voice, 6(2).

Rigby, J & Ramlogan, R 2013, The impact and effectiveness of entrepreneurship policy, Nesta Working Paper No. 13/01, Manchester Institute of Innovation Research, http://www.nesta.org.uk/wp13-01

Rogers, PJ 2008a, A Map of Impact Evaluation (slides), Edinburgh Evaluation Summer School, 26-27 May 2008.

Rogers, PJ. 2008b, Using programme theory to evaluate complicated and complex aspects of interventions. Evaluation, 14(1), 29-48.

Rogers, PJ 2009a, 'Learning from the evidence about evidence-based policy', in Productivity Commission (ed.), Strengthening Evidence-Based Policy in the Australian Federation, Productivity Commission, Melbourne.

Rogers, PJ 2009b, Matching impact evaluation design to the nature of the intervention and the purpose of the evaluation. Journal of development effectiveness, 1(3), 217-226.

Rogers, PJ 2011, Implications of complicated and complex characteristics for key tasks in evaluation. Evaluating the Complex: Attribution, Contribution, and Beyond, ed. K. Forss, R. Schwartz, and M. Marra, 18, 33-52.

Ruegg, R & Jordan, G 2007, Overview of evaluation methods for R&D programs: A Directory of Evaluation Methods Relevant to Technology Development Programs, prepared for U.S. Department of Energy, Office of Energy Efficiency and Renewable Energy.

Sanders, C 2002, 'The Impact of Microenterprise Assistance Programs: A comparative study of program participants, nonparticipants, and other low-wage workers', Social Service Review, vol. 76, no. 2, pp. 321–340.

Schwandt, T 2015, Evaluation Foundations Revisited: Cultivating a Life of the Mind for Practice, Stanford University Press, Stanford.

Shadish, W et al. 2002, Experimental and Quasi-Experimental Designs for Generalized Causal Inference, Houghton Mifflin Company, Boston.

Shen, T, Zhang, S, Chan, M & Hansen, B 2009, Growing America Through Entrepreneurship: Final Evaluation of Project GATE, US Department of Labor, Employment and Training Administration, Washington, DC.

Sherman, L, Gottfredson, D, MacKenzie, D, Eck, J, Reuter, P & Bushway, S 1998, Preventing crime: what works, what doesn't, what's promising, Report to the United States Congress, National Institute of Justice, Washington, DC.

Stacey, R 1992, Managing the unknowable: Strategic boundaries between order and chaos in organizations, John Wiley & Sons, London.

Stern, E, Stame, N, Mayne, J, Forss, K, Davies, R & Befani, B 2012, Broadening the range of designs and methods for impact evaluations, Working Paper 38, Department for International Development, London.

Stern, E 2015, 'Impact Evaluation: A Guide for Commissioners and Managers', Report prepared for the Big Lottery Fund, Bond, Comic Relief and the Department for International Development, London: Bond.

Warwick, K & Nolan, A 2014, 'Evaluation of Industrial Policy: Methodological Issues and Policy Lessons', OECD Science, Technology and Industry Policy Papers, No. 16, OECD Publishing, Paris.

White, H. & Phillips, D 2012, Addressing attribution of cause and effect in small n impact evaluations, International Initiative for Impact Evaluation (3ie) http://www.3ieimpact.org/media/filer_public/2012/06/29/working_paper_15.pdf

Appendix 7 Glossary of evaluation options

The following and more can be found on the BetterEvaluation Website, 2015

Actor Attribution: providing evidence that links participation plausibly with observed changes.

After Action Review: bringing together a team to discuss a task, event, activity or project, in an open and honest fashion.

Analytical generalisation: making projections about the likely transferability of findings from an evaluation, based on a theoretical analysis of the factors producing outcomes and the effect of context. Realist evaluation can be particularly important for this.

Archive Data for Future Use: systems to store de-identified data so that they can be accessed for verification purposes or for further analysis and research in the future.

Assessment Scales: providing an overall rating of performance across multiple dimensions (also called a rubric).

Best evidence synthesis: a synthesis that, like a realist synthesis, draws on a wide range of evidence (including single case studies) and explores the impact of context, and also builds in an iterative, participatory approach to building and using a knowledge base.

Big data: data sets that are so voluminous and from such different sources that traditional analysis methods are not feasible or appropriate.

Biophysical: measuring physical changes over a period of time related to a specific indicator by using an accepted measurement procedure.

Block Histogram: presenting a frequency distribution of quantitative data in a graphical way.

Bradford Hill criteria: a group of minimal conditions necessary to provide adequate evidence of a causal relationship between an incidence and a possible consequence. Bradford Hill was an epidemiologist who developed the criteria in the 1960s to guide exploration of cause and effect in situations where experimental designs were not possible. They cover strength, consistency, specificity, temporality (the timing of the cause and the effect), dose-response, plausibility, coherence, (small 'e') experiment and analogy

Brainstorming: focusing on a problem and then allowing participants to come up with as many solutions as possible.

Bubble Chart: providing a way to communicate complicated data sets quickly and easily.

Card Visualization: brainstorming in a group using individual paper cards to express participants thoughts about particular ideas or issues.

Check Dose-Response Patterns: examining the link between dose and response as part of determining whether the programme caused the outcome.

Check Intermediate Outcomes: checking whether all cases that achieved the final impacts achieved the intermediate outcomes.

Check Results Match a Statistical Model: comparing results with a statistical model to determine if the programme caused the outcome.

Check Results Match Expert Predictions: making predictions based on programme theory or an emerging theory of wider contributors to outcomes and then following up these predictions over time.

Check Timing of Outcomes: checking predicated timing of events with the dates of actual changes and outcomes.

Collaborative Outcomes Reporting: mapping existing data against the theory of change, and then using a combination of expert review and community consultation to check for the credibility of the evidence.

Comparative Case Studies: using a comparative case study to check variation in programme implementation.

Component design: collecting data independently and then combining at the end for interpretation and conclusions.

Concept Mapping: showing how different ideas relate to each other - sometimes this is called a mind map or a cluster map.

Confirming and Disconfirming: providing deeper insights into preliminary findings and highlighting the boundaries of the findings.

Consensus Conference: a process where a selected group of lay people (non-experts) representing the community are briefed, consider the evidence and prepare a joint finding and recommendation

Consistent Data Collection and Recording: processes to ensure data are collected consistently across different sites and different data collectors.

Content analysis: reducing large amounts of unstructured textual content into manageable data relevant to the (evaluation) research questions.

Contribution Analysis: assessing whether the programme is based on a plausible theory of change, whether it was implemented as intended, whether the anticipated chain of results occurred and the extent to which other factors influenced the program's achievements.

Control Group: comparing an untreated research sample against all other groups or samples in the research.

Convenience sample: based on the ease or 'convenience' of gaining access to a sample simply in which data is gathered from people who are readily available.

Convergent Interviewing: asking probing questions to interviewees and then using reflective prompts and active listening to ensure the conversation continues.

Correlation: a statistical measure ranging from +1.0 to -1.0 that indicates how strongly two or more variables are related. A positive correlation (+1.0 to 0) indicates that two variables will either increase or decrease together, while a

negative correlation (0 to -1.0) indicates that as one variable increases, the other will decrease.

Cost Benefit Analysis: compares costs to benefits, both expressed in monetary units, taking into account discount factors over time, and produces a single figure of the ratio of benefits to costs.

Cost Utility Analysis: a particular type of cost-effectiveness analysis that expresses benefits in terms of a standard unit such as Quality Adjusted Life Years.

Cost-Effectiveness Analysis: calculates a ratio between the costs and a standardised unit of positive impacts (for example new patents, or new jobs).

Criterion: involving the identification of particular criterion of importance, the articulation of these criterion, and the systematic review and study of cases that meet the criterion.

Critical Case: identifying cases that have the potential to impact other cases.

Cross tabulations: using contingency tables of two or more dimensions to indicate the relationship between nominal (categorical) variables. In a simple cross tabulation, one variable occupies the horizontal axis and another the vertical. The frequencies of each are added in the intersecting squares and displayed as percentages of the whole, illustrating relationships in the data.

Data Backup: onsite and offsite, automatic and manual processes to guard against the risk of data being lost or corrupted.

Data Cleaning: detecting and removing (or correcting) errors and inconsistencies in a data set or database due to the corruption or inaccurate entry of the data.

Data mining: computer-driven automated techniques that run through large amounts of text or data to find new patterns and information.

Deliberative Opinion Polls: providing information about the issue to respondents to ensure their opinions are better informed.

Delphi Study: soliciting opinions from groups in an iterative process of answering questions in order to gain a consensus.

Demographic Mapping: using GIS (global information system) mapping technology to show data on population characteristics by region or geographic area.

Difference in Difference (or Double Difference): the before-and-after difference for the group receiving the intervention (where they have not been randomly assigned) is compared to the before-after difference for those who did not.

Dotmocracy: collecting and recognizing levels of agreement on written statements among a large number of people.

Effective Data Transfer: processes to move data between systems, including between software packages, to avoid the need to rekey data.

Email Questionnaires: distributing questionnaires online via email.

Enriching: using qualitative work to identify issues or obtain information on variables not obtained by quantitative surveys.

Examining: generating hypotheses from qualitative work to be tested through the quantitative approach.

Expert Panel: a process where a selected group of experts consider the evidence and prepare a joint finding

Explaining: using qualitative data to understand unanticipated results from quantitative data.

Exploratory Techniques: taking a 'first look' at a dataset by summarising its main characteristics, often by using visual methods.

Face Questionnaires: administering questionnaires in real time by a researcher reading the questions (either face to face or by telephone). Global

Field Trips: organizing trips where participants visit physical sites.

Fishbowl Technique: managing group discussion by using a small group of participants to discuss an issue while the rest of the participants observe without interrupting.

Focus Groups: discovering the issues that are of most concern for a community or group when little or no information is available.

Force Field Analysis: providing a detailed overview of the variety of forces that may be acting on an organizational change issue.

Framework matrices: a method for summarising and analysing qualitative data in a two-by-two matrix table. It allows for sorting data across case and by theme.

Frequency tables: a visual way of summarizing nominal and ordinal data by displaying the count of observations (times a value of a variable occurred) in a table.

Future Search Conference: identifying a shared vision of the future by conducting a conference with this as its focus.

General Elimination Methodology: this involves identifying alternative explanations and then systematically investigating them to see if they can be ruled out.

Geographical: capturing geographic information about persons or objects of interest such as the locations of high prevalence of a disease or the location of service delivery points.

Geotagging: adding geographic information about digital content, within 'metadata 'tags - including latitude and longitude coordinates, place names and/or other positional data.

GIS Mapping: creating very precise maps representing geographic coordinates that could include information relating to changes in geographical, social or agricultural indicators.

Goal Attainment Scales: recording actual performance compared to expected performance using a 5 point scale from -2 (much less than expected) to +2 (much more than expected).

Hierarchical Card Sorting: a participatory card sorting option designed to provide insight into how people categorize and rank different phenomena.

Homogenous: selecting similar cases to further investigate a particular phenomenon or subgroup of interest.

Horizontal Evaluation: An approach that combines self-assessment by local participants and external review by peers

Icon array: a matrix of icons (usually 100 or 1000 icons) typically used as a frequency-based representation of risk, simultaneously displaying both the number of expected events and the number of expected non-events.

Index: a composite variable made up of data from individual items.

Instrumental Variables: a method used to estimate the causal effect of an intervention.

Integrated design: combining different options during the conduct of the evaluation to provide more insightful understandings.

Intensity: selecting cases which exhibit a particular phenomenon intensely.

Interactive mapping: maps that allow users to interact – e.g. zooming in and out, panning around, identifying specific features, querying underlying data such as by topic or a specific indicator (e.g., socioeconomic status), generating reports

Internet Questionnaires: collecting data via a form (with closed or open questions) on the web. Interviews: in-depth, structured, semi-structured, or unstructured.

Judgemental Matching: a comparison group is created by finding a match for each person or site in the treatment group based on researcher judgements about what variables are important.

Key Informant Interviews: interviewing people who have particularly informed perspectives.

Key Informant: asking experts in these types of programmes or in the community to predict what would have happened in the absence of the intervention.

Key Informant: asking experts in these types of programmes or in the community to identify other possible explanations and/or to assess whether these explanations can be ruled out.

Keypad technology: gauging audience response to presentations and ideas in order to gain provide valuable feedback from large group settings.

Lessons learnt: Lessons learnt can develop out of the evaluation process as evaluators reflect on their experiences in undertaking the evaluation.

Line Graph: displaying information as a series of data points connected by straight line segments, on two axes.

Logically constructed counterfactual: using the baseline as an estimate of the counterfactual. Process tracing can support this analysis at each step of the theory of change.

Logs and Diaries: monitoring tools for recording data over a long period of time.

Mapping: creating visual representations ('map') of a geographically based or defined issue.

Matched Comparisons: participants are each matched with a non-participant on variables that are thought to be relevant. It can be difficult to adequately match on all relevant criteria.

Matrix Chart: summarising a multidimensional data set in a grid.

Maximum Variation: contains cases that are purposefully as different from each other as possible.

Measures of central tendency: a summary measure that attempts to describe a whole set of data with a single value that represents the middle or centre of its distribution. The mean (the average value), median (the middle value) and mode (the most frequent value) are all measures of central tendency. Each measure is useful for different conditions.

Measures of dispersion: a summary measure that provides information about how much variation there is in the data, including the range, inter-quartile range and the standard deviation.

Meta-analysis: a statistical method for combining numeric evidence from experimental (and sometimes quasi-experimental studies) to produce a weighted average effect size.

Meta-ethnography: a method for combining data from qualitative evaluation and research, especially ethnographic data, by translating concepts and metaphors across studies.

Mobile Phone Logging: targeted gathering of structured information using devices such as smartphones, PDAs, or tablets.

Modus Operandi: drawing on the previous experience of participants and stakeholders to determine what constellation or pattern of effects is typical for an initiative.

Most Significant Change: a cyclic process of identifying domains of change that are of interest, gathering stories of change, working with individuals in a group to select the most significant change and explain why, reflecting on the stories selected to better understand the values of different stakeholder groups in terms of what is valued

Multi-Criteria Analysis: a systematic process to address multiple criteria and perspectives

Multiple Lines and Levels of Evidence (MLLE): reviewing a wide range of evidence from different sources to identify consistency with the theory of change and to explain any exceptions.

Multi-Stage: cluster sampling in which larger clusters are further subdivided into smaller, more targeted groupings for the purposes of surveying. Sequential: selecting every nth case from a list (e.g. every 10th client).

Multivariate descriptive: providing simple summaries of (large amounts of) information (or data) with two or more related variables.

Mural: collecting data from a group of people about a current situation, their experiences using a service, or their perspectives on the outcomes of a project.

Network Diagram: a depiction of how people or other elements are related to one another.

Non-Parametric inferential statistics: methods for inferring conclusions about a population from a sample's data that are flexible and do not follow a normal distribution (i.e., the distribution does not parallel a bell curve), including the chi-square test, binomial test.

Non-participant Observation: observing participants without actively participating.

Numeric analysis: Analysing numeric data such as cost, frequency, physical characteristics.

Numeric Weighting: developing numeric scales to rate performance against each evaluation criterion and then add them up for a total score.

Official Statistics: obtaining statistics published by government agencies or other public bodies such as international organizations. These include quantitative or qualitative information on all major areas of citizens' lives such as economic and social development, living conditions, health, education, the environment.

ORID: enabling a focused conversation by allowing participants to consider what is known (Objective) and their feelings (Reflective) before considering issues (Interpretive) and decisions (Decisional).

Outlier: analysing cases that are unusual or special in some way, such as outstanding successes or notable failures.

Parallel Data Gathering: gathering qualitative and quantitative data at the same time.

Parametric inferential statistics: methods for inferring conclusions about a population from a sample's data that follows certain parameters: the data will be normal (ie, the distribution parallels the bell curve); numbers can be added, subtracted, multiplied and divided; variances are equal when comparing two or more groups; and the sample should be large and randomly selected.

Participant Observation: identifying the attitudes and operation of a community by living within its environs.

Peer/Expert Reviews: Drawing upon peers or experts with relevant experience and expertise to assist in the evaluation of some aspect or all of a project.

Photo Voice: promoting participatory photography as an empowering option of digital storytelling for vulnerable populations.

Photography/video: discerning changes that have taken place in the environment or activities of a community through the use of images taken over a period of time.

Photolanguage: eliciting rich verbal data where participants choose an existing photograph as a metaphor and then discuss it.

Phrase Net: depicts, in a network diagram, the relationships between different words in a source text using pattern matching (i.e., looks for pairs of words that fit particular patterns). Matching different patterns provides different views of concepts contained in the text.

Polling Booth: a method of allowing informants to provide sensitive information through an anonymous voting process.

Positive Deviance: involves intended evaluation users in identifying 'outliers' – those with exceptionally good outcomes - and understanding how they have achieved these.

Postcards: collecting information quickly in order to provide short reports on evaluation findings (or an update on progress).

Previous Evaluations and Research: using the findings from evaluation and research studies that were previously conducted on the same or closely related areas.

Process Tracing: case-based approach to causal inference which focuses on the use of clues within a case (causal-process observations, CPOs) to adjudicate between alternative possible explanations

Project Records: retrieving relevant information from a range of documents related to the management of a project such as the project description, strategic and work plans, budget and procurement documents, official correspondence, minutes of meetings, description and follow-up of project participants, progress reports.

Projective Techniques (photo-elicitation): participants selecting one or two pictures from a set and using them to illustrate their comments about something.

Propensity Scores: statistically creating comparable groups based on an analysis of the factors that influenced people's propensity to participate in the program.

Q-methodology: investigating the different perspectives of participants on an issue by ranking and sorting a series of statements (also known as Q-sort).

Qualitative Comparative Analysis: comparing the configurations of different cases to identify the components that produce specific outcomes.

Qualitative Weight and Sum: using qualitative ratings (such as symbols) to identify performance in terms of essential, important and unimportant criteria

Randomized Controlled Trial (RCT): creating a control group and comparing this to one or more treatment groups to produce an unbiased estimate of the net effect of the intervention.

Rapid evidence assessment: a process that is faster and less rigorous than a full systematic review but more rigorous than ad hoc searching, it uses a combination of key informant interviews and targeted literature searches to produce a report in a few days or a few weeks.

Rapid Outcomes Assessment: assessing and mapping the contribution of a project's actions on a particular change in policy or the policy environment.

Realist Analysis of Testable Hypotheses: Using a realist programme theory (what works for whom in what circumstances through what causal mechanisms?) to identify specific contexts where results would and would not be expected and checking these.

Realist Evaluation: Analyses the contexts within which causal mechanisms produce particular outcomes, making it easier to predict where results can be generalised.

Realist synthesis: synthesizing all relevant existing research in order to make evidence-based policy recommendations.

Regression Discontinuity: comparing the outcomes of individuals just below the cut-off point with those just above the cut-off point.

Reputational Monitoring Dashboard: monitoring and quickly appraising reputational trends at a glance and from a variety of different sources.

Rubrics: using a descriptive scale for rating performance that incorporates performance across a number of criteria

Ruling Out Technical Explanations: identifying and investigating possible ways that the results might reflect technical limitations rather than actual causal relationships.

Scatterplot: displaying the relationship between two quantitative variables plotted along two axes. A series of dots represent the position of observations from the data set.

Searching for Disconfirming Evidence/Following Up Exceptions: Treating data that don't fit the expected pattern not as outliers but as potential clues to other causal factors and seeking to explain them.

Seasonal Calendars: analysing time-related cyclical changes in data. Sketch

Secure Data Storage: processes to protect electronic and hard copy data in all forms, including questionnaires, interview tapes and electronic files from being accessed without authority or damaged.

Sequential Allocation: a treatment group and a comparison group are created by sequential allocation (e.g. every 3rd person on the list).

Sequential Data Gathering (Sequencing): gathering one type of data first and then using this to inform the collection of the other type of data.

Simple Random: drawing a sample from the population completely at random.

Snowball: asking a number of people where else to seek information creates a snowball effect as the sample gets bigger and bigger and new information rich examples are accumulated

Social mapping: identifying households using pre-determined indicators that are based on socioeconomic factors.

Social Return on Investment: a systematic way of incorporating social, environmental, economic and other values into decision-making processes

Stacked Graph: visualising how a group of quantities changes over time. Items are 'stacked' in this type of graph allowing the user to add up the underlying data points.

Statistical generalisation: statistically calculating the likely parameters of a population using data from a random sample of that population.

Statistically Controlling for Extraneous Variables: where an external factor is likely to affect the final outcome, it needs to be taken into account when looking for congruence.

Statistically Created Counterfactual: developing a statistical model, such as a regression analysis, to estimate what would have happened in the absence of an intervention.

Stories (Anecdote): providing a glimpse into how people experience their lives and the impact of specific projects/programs.

Stratified Random: splitting the population into strata (sections or segments) in order to ensure distinct categories are adequately represented before selecting a random sample from each.

Summary statistics: providing a quick summary of data which is particularly useful for comparing one project to another, before and after.

Survey: collecting data in response to structured questions.

SWOT Analysis: reflecting on and assessing the Strengths, Weaknesses, Opportunities and Threats of a particular strategy.

Systematic review: a synthesis that takes a systematic approach to searching, assessing, extracting and synthesizing evidence from multiple studies. Meta-analysis, meta-ethnography and realist synthesis are different types of systematic review.

Telephone Questionnaires: administering questionnaires by telephone.

Textual analysis: Analysing words, either spoken or written, including questionnaire responses, interviews, and documents.

Textual narrative synthesis: dividing the studies into relatively homogenous groups, reporting study characteristics within each group, and articulating broader similarities and differences among the groups.

Thematic coding: recording or identifying passages of text or images that are linked by a common theme or idea allowing the indexation of text into categories.

Theory-based: selecting cases according to the extent to which they represent a particular theoretical construct.

Time series analysis: observing well-defined data items obtained through repeated measurements over time.

Timelines and time-ordered matrices: aids analysis by allowing for visualisation of key events, sequences and results.

Transect: gathering spatial data on an area by observing people, surroundings and resources while walking around the area or community.

Treemap: makes use of qualitative information in the form of important distinctions or differences that people see in the world around them. They help overcome some of the problems that may be encountered when dealing with qualitative information.

Triangulation (Confirming/reinforcing; Rejecting): verifying or rejecting results from quantitative data using qualitative data (or vice versa)

Typical Case: developing a profile of what is agreed as average, or normal.

Value for Money: a term used in different ways, including as a synonym for cost-effectiveness, and as systematic approach to considering these issues throughout planning and implementation, not only in evaluation.

Volunteer: sampling by simply asking for volunteers

Vote counting: comparing the number of positive studies (studies showing benefit) with the number of negative studies (studies showing harm).

Word Cloud: assists an evaluator identify important words during the process of textual analysis.

Word Tree: a visual display of the words in qualitative dataset, where frequently used words are connected by branches to the other words that appear nearby in the data.

World Cafe: hosting group dialogue in which the power of simple conversation is emphasised in the consideration of relevant questions and themes.

Writeshop: a writing workshop involving a concentrated process of drafting, presenting, reviewing and revising documentations of practice