Evaluating Innovation Programs

Report for the Victorian Department of State Development, Business and Innovation (DSDBI)

Paul Jensen and Elizabeth Webster

Melbourne Institute of Applied Economic and Social Research

University of Melbourne

FINAL REPORT

June 2014

Executive Summary

- Evaluations must have a control group (or counterfactual). Reporting on how funds were spent, the number of people participating in a program and the outcome of program participants is a descriptive report, **not** an evaluation.
- It is difficult to conduct a convincing evaluation unless the evaluation framework is set up before the program began. *Ex post* evaluations can be conducted but the results are characteristically heavily qualified. The level of confidence in these *ex post* results can be too poor to provide a clear voice on what works and what does not.
- The evaluator does not need to use the gold-standard evaluation method (randomised control trials) to get solid evidence. Other ways of selecting a control group (such as using firms located in a different State) and clever estimation methods (such as difference-in-difference estimations; regression discontinuity) can provide robust evidence in less time and at cheaper cost.
- A series of consistent empirical findings are needed before people will accept a finding as a 'stylised fact' (well-acknowledged empirical regularity). This is similar to all areas of empirical research in the social, life and physical sciences. For example, our acceptance of the link between smoking and cancer is based on numerous converging studies, not a single study.
- The Government should commit to spending 1% of its program budget on evaluation. The evaluation team should be brought into the program at the design stage.
- Certain practical difficulties caused by short corporate memory hinder *ex ante* evaluations. The turnover of staff in and out of government departments makes it difficult for data which has been collected at the start of a program to be used several years later to evaluate the program. The establishment of a whole-of-government evaluation unit to oversee or undertake program evaluations could ameliorate this problem.
- The Government should consider hiring external evaluation experts from the university sector to run practical annual workshops to train the government evaluation team. These workshops need to be undertaken annually to replenish the skills within the government sector.
- Most reliable evaluations use unit-record data. The analyst must be able to see the characteristics of the individual firm in order to conduct rigorous analysis and draw strong inferences about the effects of policy. The Government should consider new initiatives to link data and generally make linked data more accessible. Technological advances are making it much cheaper (and safer in terms of privacy) than ever before.

1. Introduction

> "You change your laws so fast…without inquiring after results past or present, that it is all experiment, seesaw, doctrinaire; a shuttlecock between battledores."
>
> Florence Nightingale[1], as quoted by Gary Banks in Productivity Commission (2010, p.3).

The notion that Governments do a poor job at evaluating their laws, policies and programs still resonates today. The ultimate goal of better policy evaluation is to provide an evidence base to develop better public policies. In this Report, we take a look at the state-of-the-art as it relates to policy evaluation in the 21st century. Although the specific emphasis is on 'innovation policy', much of the Report has implications for other policy domains. In doing so, we consider techniques, methods and data requirements, and also consider potential impediments to implementing best practice policy evaluation.

In this report, an innovation program is defined as an intervention to enhance the rate at which firms adopt new or substantially improved products, processes, operations and marketing practise. The Victorian Department of State Development, Business and Innovation (DSDBI) is the primary Victorian Government agency responsible for supporting organisations and Government to boost their productivity through innovation and technology. Specific objectives targeted by the Department's innovation activities include:

- Strengthening coordination and connections in the local innovation system;
- Stimulating additional business innovation for productivity, growth and exports;
- Supporting the development of innovation skills;
- Attracting new business investment into Victoria's R&D strengths to build innovation capabilities and create new high-value jobs.

In order to achieve these objectives, DSDBI delivers a range of programs and activities to address identified areas of deficiency within the Victorian innovation system. But how do we know whether innovation programs are successful in (helping to) achieving these objectives?

It is common for people to confuse monitoring with evaluation. Whereas monitoring fulfils an auditing function – reporting on how the moneys were spent, the number of people participating in a program and the numbers still in business or employment – an evaluation aims to establish what would have happened to the people or the business in the absence of the program. It is common for people to believe that monitoring is all that is required to justify a program. This is far from good policy. Competent monitoring is a necessary condition for probity and efficient policy but it does not tell us whether public funds have been spent in a manner that delivers the maximum benefits to citizens. **The key challenge for a good evaluation is to identify a counterfactual**: what would have happened in the absence of participating in the program?[2] Or what would have happened to the Victorian economy if this program had not been run? This is quite difficult to do but it is an important pillar of evidence-based policymaking. Although it is arduous, there are ways in which a counterfactual can be calculated. And once these challenges are embraced, the process of designing

[1] Florence Nightingale in a letter to Sir Francis Galton 7 February 1891.

[2] There are many examples of 'evaluations' that do not use or attempt to use a counterfactual i.e. NIH (2009). These are properly called descriptive reports, not evaluations.

and implementing better innovation policy can continue apace. Adopting a coherent analytical framework and systematic evaluation of existing programs will help with the introduction of better innovation policy in the long-run. The aim of this project is to identify the key methods and data required to undertake such evaluations.

If evaluation is difficult – and individual Government departments are investing in specific programs – why is it worthwhile improving the quality of the evaluations? The reason is simple: in order for programs to be supported (or even expanded over time) by the central agencies, it is important to provide them with evidence of sufficiently high quality that is believable and trustworthy. In the innovation policy domain, this is perhaps even more important since the rationale for government support of innovation is often questioned.

The cornerstone of the evidence-based policy approach is that good public policy can only come about through rigorous evaluation of existing programs in order to determine what works *and* what doesn't work. With careful evaluation of the evidence, programs that don't work can be identified and dropped, while those programs that do work can continue (or be enhanced). Of course, it is difficult to determine and isolate 'what works' since most programs are instituted in complex settings where there are many factors that influence their success (e.g. even 'good programs' could produce 'bad' results during a downturn in the business cycle). Nevertheless, there are a range of methods and techniques – from randomized controlled trials to difference-in-differences and case studies – that can be used to examine the effects of specific programs.

In this Report, we will consider and critique the available evaluation methods in order to uncover how to establish robust, *causal* evidence of the effects of a specific program. Since establishing causality is very difficult, we will also explore the data requirements which underpin each method. Moreover, the most appropriate methods to use will be context-specific and will also depend on the whether baseline data were collected before the program began; the budget; and the length of lapsed time since the program was undertaken. For more practical guidance on how to choose the best evaluation protocol, see DSDBI (2013).

In the next section of the Report, we will provide some background on the issues underpinning good innovation policy and a framework for understanding the issues. In particular, we will examine what constitutes best-practice innovation policy and any obstacles that must be overcome in order to achieve this standard in Victoria. Following that, we will review Victoria's recent experience with innovation program evaluation in order to assess: (i) the quality of the evidence provided; and (ii) to highlight ways in which the quality of evidence could be improved. A number of specific program evaluations have been identified by DSDBI and these will be critically reviewed. In the course of reviewing these evaluations, we will shed light on the following questions:

- What evaluation methods provide the most rigorous evidence of program efficacy?
- What standard of evidence has been produced by recent innovation program evaluations?
- How can innovation program evaluation move to more rigorous standards of evaluation?
- What are the limitations in reaching higher standards of evaluation, and what developments are or could be enabling new standards of evaluation?

2. Background

All public policymakers face a difficult task in trying to maximise the returns from their programs. If we abstract from the political issues which may shape the existence and/or size of specific programs and simply focus on the economic, social and environmental (i.e. non-political) objectives of the program, it is obvious that all policymakers want the best possible outcomes given the prevailing economic climate (which, of course, changes over time). There are lots of problems in achieving the best possible outcomes for government interventions: for example, there are numerous examples of 'unintended consequences' resulting from specific government programs (see Gans and Leigh 2009 for an analysis of the change in the timing of births via inducements and caesarean sections, as a result of the Federal Government's baby bonus). If similar policies have been implemented elsewhere we can learn from their experiences. In other instances, it is more difficult to predict the effects of a specific policy, particularly if it is a new (i.e. new-to-the-world) policy.

2.1 Importance of Evaluation to Inform Policy

Typically, we can only learn from policy experiences if the objectives and outcomes are rigorously documented. The quality of the analysis undertaken on a program directly influences the quality of the lessons learned. Since it is extremely difficult to establish causal effects of policies, such analysis must be designed and performed very carefully.[3] High quality evaluation usually requires *ex ante* design of programs, to ensure the right data is collected and an appropriate methodology is selected up front. In doing so it may become apparent that the program design will not be conducive to evaluation at a high level.

A major issue is to separate the effect of the program from other co-incidental occurrences such as a change in the macro-economy or the trajectory the firm was on regardless of its participation in the program. We take the position that the target audience of program evaluation are ultimately the central agencies who control the purse strings of the government departments: our intention is that high-quality evidence should convince them of the merits of the program or the need to redirect money to other uses. Of course, high-quality evidence on its own is not sufficient to guarantee that better policies will be implemented: there are obvious political issues that will need to be addressed as well. But, abstracting from good luck, better evidence is a necessary but not sufficient condition for better public policies.

2.2 A Poor Innovation Policy Evidence Base

With regard to innovation policy, the lessons we are able to learn from other programs (i.e. the 'evidence base') is rather limited. As pointed out recently by the US Government's Chairman of the Federal Reserve, Ben Bernanke:

> "If the government decides to foster R&D, what policy instruments should it use? A number of potential tools exist, including direct funding of government research facilities, grants to university or private-sector researchers, contracts for specific projects, and tax incentives. ... **Unfortunately, economists know less about how best to channel public support for research and development than we would like**; it is good news, therefore, that considerable new work is being done on this topic, including recent initiatives on science policy by the National Science Foundation (see Lane 2009)." May 16th, 2011, see http://www.federalreserve.gov/newsevents/speech/bernanke20110516a.htm

[3] In Section 3 of this Report, we provide a detailed account of what we mean by the 'quality of evidence'.

This is a major problem facing innovation policymakers: although just about everyone believes that innovation has a role to play in generating long-term economic prosperity, there is scant evidence that governments need to play a role in fostering innovation. To make this case convincingly requires serious analysis. As Bernanke points out, this is a serious problem in all countries, not just in Australia. But it is slowly being addressed (in the US at least) with the recent initiatives being launched by the National Science Foundation, which relate in the first instance to the creation of new data infrastructure which will enable the rigorous analysis which is required to improve innovation policymaking. It is important to note that the creation of data is the first, crucial step in the process. As we will see later, without being able to see program participants before and after they participated in a program, it is exceptionally difficult to understand what effect the program had on their performance. And this requires data.

2.3 Good Evaluation Requires Accessible Data

One of the other challenges for innovation policymakers is that data accessibility is far poorer in innovation policy than it is in other areas of policymaking: education, labour and health policy for example, all rely heavily on major data infrastructure in order to analyse the effects of different policies (e.g. MySchool data; Victorian Accident and Emergency Data; Household Income and Labour Dynamics in Australia (HILDA); Medicine in Australia: Balancing Life and Employment (MABEL) to name a few). Around these datasets, some of which have been around for a decade – there has built up a community of practice amongst government analysts and academics who share results, techniques, and information. This has resulted in the development of an 'evidence base' in these domains which is crucial for policymaking. Given the importance of innovation for long-term productivity growth, it is imperative that the similar national data infrastructure be established as a necessary first step in developing a more rigorous evidence base for innovation policy in Australia.

The size and quality of the evidence base is important because it is rarely the case that a single piece of evidence is enough to change policy. For example, it wasn't a single piece of evidence that was sufficient to convince the Australian Government to lower tariffs back in the 1970s; rather, it was an accumulation of evidence over time coupled with the conviction of the Government that it was the right thing to do. As Angrist and Pischke (2010) state: "… the process of accumulating empirical evidence is rarely sexy in the unfolding, but accumulation is the necessary road along which results become more general" (pp.23-4).

In more recent years, it has become clearer that access to unit-record data has become more important: the analyst must be able to see the characteristics of the individual firm in order to conduct rigorous analysis and draw strong inferences about the effects of policy. And with many new initiatives to link data and generally make it more accessible, there is every chance that technology will make it much cheaper (and safer in terms of privacy) to do so in the future.[4] Given the improved data, methods (including better research design) and computing power available today, Angrist and Pischke (2010) have argued that the profession is slowly but surely 'taking the con out of econometrics'.[5] They refer to the recent period as a revolution in the credibility of empirical economics, which was severely questioned in the 1970s and 1980s (see Leamer 1983 for example).

[4] For example, there are exciting new initiatives such as the National Opinion Research Centre (NORC) platform at the University of Chicago which potentially make remote access to confidential data easier and safer.

[5] Increased computing power on its own is not enough since datasets also continue to get larger and more complex.

2.4 Better Access to ABS Data

In many respects, this revolution has been more prevalent in labour, health, education and development economics than in industrial economics (in which innovation economic policy is included). Although the Australian Bureau of Statistics (ABS) does collect information on individual firms, access to the data has proved difficult. And it is often the case that even if access to the Confidentialised Unit Record File (CURF) of the Business Longitudinal Database (BLD) is made available, for example, crucial information has to be 'grouped' in order to protect the identity of the firm. For example, the data providers group firms as having 1-20, 21-50, 51-100 employees rather than providing information on the actual number of employees. This makes productivity calculation impossible since the number of employees needs to be included in the denominator of a productivity estimate (output/inputs).

2.5 The Philosophy of Firm Performance

There are two factors limiting the impact of innovation program evaluations. The first is the quality of evaluations. This quality is currently far below the quality of evaluations done in other policy areas – such as education, health and social policy. The second is the philosophical standpoint of some policy analysts. Many analysts will tend to put less weight on innovation program evaluations due to scepticism that such programs can have an impact. In other words, they believe in the view there is no need for governments to intervene in the market for innovation in the first place, so any programs are worthless (almost by definition).[6] However, other economists and most management scientists believe that there is a role for government since the market for innovation is subject to market failure and won't produce the optimal amount of innovation if left to its own device.[7] These (typically unstated) assumptions often lead to parallel conversations between the two groups and can form a barrier to innovation policy.

3. Characteristics of Evidence-Based Policy[8]

Evidence-based policy is a decision making process which combines deductive logic with statistical analysis to inform policy decision making.[9] Its hallmark is rigour and objectivity. Since economic theory typically predicts that policy changes will produce trade-offs and countervailing effects (that is, different groups react in different ways or are differentially affected; feedback effects occur), it is often not possible to know whether the final effect will be a net benefit or net cost to society. Moreover, theory in most cases does not indicate how large effects will be. Accordingly, logic alone cannot identify the optimal policy and empirical estimates are needed to adjudicate. Good evidence-based policy not only allows the decision maker to select the program that suits their ends but also arms them with the evidence to convince others. An evaluation makes transparent the lost benefits from pursuing one course of action over another. To quote Lindsay Tanner: "Every government

[6] This assumption can be strongly held by economists working in macroeconomics or in areas that rarely come in contact with real life companies.

[7] They argue that two stylised facts support this view: first, the persistence of differing levels of productivity by firms in the same industry; and second, the very long tail of firms who are a long way behind the leading firm.

[8] The material in this section is taken (with permission) from Palangkaraya, Webster and Cherastidtham (2012).

[9] According to Heckman (2000, p.3):"Economic theory plays an integral role in the application of econometric methods because the data almost never speak for themselves, especially when there are missing data and missing counterfactual states."

dollar wasted on a poor program is a dollar that a working person doesn't have to spend on groceries, health care and education. It is ... a dollar that the Government does not have available to spend on its policy priorities" (quoted in Banks 2009, p.20).

The characteristics of evidence-based policy are that it:

- estimates which parties are (notably) affected, the size of these effects and the net effect on societal well-being. Good economic policy should consider both pecuniary effects (e.g. productivity, income) and non-pecuniary effects (e.g. environmental and social impacts). Where possible these effects may be converted into monetary equivalents but where this is not possible, some qualitative mention should be made (such as years of life extended, air quality);
- estimates the counterfactual of a given program or policy. This involves identifying 'silent' or uninformed third parties who will be affected by a change and evaluating the impact on their choices, incomes and behaviours;
- questions habits and existing ways of doing things. All productivity change involves changes in the way work is conducted and the first step in the process towards improving productivity is to question whether established ways of operating are efficient. Inevitably, evidence-based policy tools are valued by reforming governments;
- enables policy makers to learn and refine existing programs. Radically new programs typically start life as small pilot programs and then evolve by incremental improvements. Learning from both doing and evaluation is an essential part of good program design. Lack of transparency hides failures and allows the status quo to continue;
- allows the analyst to assess whether the impact of programs is weighted towards one sub group – be it demographic, economic, or spatial. Regular evaluations of the impact of tariffs made it quite clear which industries and regions – which seemingly had no connection to tariffs – were in actual fact negatively impacted by tariffs; and
- is useful in the public sector in the absence of relevant price signals. While the private sector can use stock prices and revenue data to indicate whether a project is meeting needs or producing the desired results, the public sector must often create its own measures of impact and value. This is because the public sector does not aim to maximise profits or sales, but rather aims to maximise societal well-being. Good analysis and evidence from reputable and independent parties can win the confidence of stakeholders and the public.

Evidence-based policy is best when:

- datasets are large, flexible and reliable. These data need to be of sufficient quality to meet the end user needs. This might be fit-for-purpose aggregate statistics or fit-for-purpose micro data. The larger, more reliable and more flexible the dataset, the more able analysts are to answer a range of questions;
- the work of the evaluating organisation is open to critical challenge. Data sharing fosters an open research community and reinforces transparent scientific inquiry. It also provides expansive views as opposed to siloed information; and

- the analysts are independent and reputable. While in-house analysis, appropriately done, has value, it should not be the sole source of evidence-based policy advice. All parties to a dispute can find or buy evidence to their liking, and policymakers and the community, can find it difficult to separate the reasoned from the self-serving. The standing and independence of voices in this space is critical to winning over the confidence of people who do not have the skills or the time to make their own assessment of a policy option.

By contrast:

- it is rare that a single study is robust enough to give people confidence that it has uncovered a 'stylised' fact.[10] Understanding causes and consequences typically emerges from systematic (meta) reviews of all available research;
- the absence of objective evidence can leave policy makers beholden to interest groups, which do not represent the range of affected parties, and to speculation and sensationalism; and
- when good datasets are not made available, or people skilled-in-the-art are not available, then evaluations can proceed with sub-standard data and inferior analysis.

According to Banks (2009), all good evaluations have a number of features in common in that they:

- test a theory as to why policy action will be effective in promoting community well-being;
- treat the counterfactual seriously;
- quantify impacts where possible;
- include both direct and important indirect effects;
- set out the uncertainties and control for confounding influences;
- are designed to avoid error that might arise through self-selection or other sources of bias;
- include sensitivity tests; and
- can be tested and replicated by third parties. Wide access to research data helps prevent misrepresentations of the evidence.

4. Methods for Creating 'Evidence'

Almost all policies involve trade-offs: there are always winners and losers (costs and benefits) associated with policy implementation. We adopt a strictly utilitarian approach to understanding the effects of policy introduction: given that there are always winners and losers, the correct metric to apply to the consequences of a policy is whether it results in 'the greatest good for the greatest number' (to paraphrase Jeremy Bentham). In this Section, we consider a range of different methods employed to determine whether a policy works or not. Moreover, we consider the strengths and weaknesses of these different methods. However, it is important to note that not every method is available in every context: there are some instances where it is possible to take advantage of a natural experiment, but this is not always the case. So, when thinking about how to go about

[10] A 'stylised fact' is a simplification of regular and robust empirical findings. It is a broad generalization that which although essentially true may have inaccuracies in the detail.

program evaluation, it is important to bear in mind that the best approach is often dictated by the data that are available.

4.1 Evaluation Questions, Methods and Data

Generally speaking, there are three distinct program evaluation questions:

1. What is the effect of a program on participants and non-participants compared to no program at all?
2. What is the likely effect if the program is applied to a new environment?[11] and
3. What is the opportunity cost – that is, what are the benefits forgone from not lowering taxes or spending on other programs?

These questions require different evaluation methods. To address the first question, the common evaluation method is based on estimation of a 'treatment effect'. The underlying idea of the treatment effect approach is to mimic the hard science approach: the average outcome of persons exposed to the policy (the treated group) is compared to the average outcome of persons who are not (control group). However, policy analysts need to take into account potential influences coming from the person's social interactions which result in direct and indirect policy impacts.

The second question is more ambitious than the first. It requires answers based on estimates that are of higher degree of interpretability, transportability and comparability than the ones produced by the treatment effect approach. In other words, to answer the second question requires estimation of tightly specified economic structural models (Heckman 2000).

The third question is more difficult again because it requires defining a reasonable alternative use of funds. To answer this question, analysts typically bring General Equilibrium models into play. General equilibrium impacts of policy are of interest, but are extremely difficult to address and typically rely on controversial assumptions. These studies acknowledge that policies have effects that ripple throughout the economy, not just on the target group of interest. In this sense, the determination of whether a program 'works' captures all of the economic consequences of participation in the program. For example, a program designed to impact mothers' labour supply will probably affect mothers' decision about if (and when) to return to (casual, part-time or full-time) work but it will also have effects on the demand for childcare services, take-away meals and dry cleaning services. Thus, there are economy-wide impacts of the program that should be accounted for, which is typically done using General Equilibrium modelling.[12]

In some cases, the evaluation of a program should consider more than just whether or not an outcome has occurred. The valuation of the outcomes is also important. Different people may have different valuations of the same outcomes. Only if people's outcome valuations (i.e. preferences) are similar will there be a unique evaluation of the outcomes associated for each possible state from each possible program. This is why policy evaluation at the macro level (such as the effect of a program on GDP) may be insufficient because it ignores people's heterogeneity in the valuation of the outcomes. Where programs have more than one desired outcome – such as employment growth, export growth, productivity growth – the evaluation typically presents a cost per single

[11] (Heckman 2000, p.6).

[12] As an alternative to General Equilibrium modelling, one could compare the net present value of an innovation program with a net present value from a health or education program.

outcome. Combining these separate outcomes into a single performance indicator is an extension that involves subjective weightings from the analyst.

We argue that the first and most important part of an evaluation is question 1 (above): what is the effect of the program on participants and non-participants? Without convincingly establishing this effect, questions 2 (the effect in a new environment) and 3 (the opportunity cost of funds) are academic. Hence we would advise any Australian government to prioritise establishing the best design and practice for question 1. This report will only examine question 1.

4.1.1 Selecting the control group

One of the most difficult issues in evaluation is removing the effects of self-selection: i.e. the fact that businesses choose to participate in a program. This is essentially the problem of defining a suitable control group. If the best or most persistent and determined companies select themselves into a program, then it is very difficult to disentangle the effects of the program from the effects of persistence and determination. Most policy makers these days will not accept results from an evaluation that has not made a convincing attempt to remove the effects of selection.

There are a number of options for selecting a control group from observational data depending on the nature of the data:

- Control groups are chosen from populations that are as similar as possible to the treatment group but for some reason (which is unrelated to firm performance) have not participated in the treatment. A common method is to choose a similar firm or individual from a different location (which has a similar market environment) i.e. NSW if the treatment group is in Victoria. If we are operating a program to increase the managerial efficiency in the automotive components sector for example, we would not select as a control group firms of all sizes and from all industries. Common sense would tell us that we should first select a control group from a population that is similar as possible to the treatment group. So we might select firms in the same industry, same size group and similar technology etc. Unfortunately, the closer we are on these characteristics, the less chance we have of finding firms who have not participated in the program.
- Similar firm or individual in the same location, but we are confident that the reasons for not undertaking the treatment are unrelated to the firm's performance (e.g. in the wrong place at the right time). An example of this might be the automotive components firms in Victoria because we do not have data on similar firms from other States.
- Where this is not possible and we suspect that the more informed and active firms are selecting into the program, then the evaluator may choose to survey the managerial characteristics of both a treatment and control group at the start of the program. Note that this requires the evaluator to be involved at the start of the program.
- If this is not possible, the evaluator can simply choose a control group *ex post* and note the direction of the bias due to unmeasured confounding factors. In addition, selecting a control group *ex post* means we often miss recording valid 'controls' which were in business at the time the program operated but has ceased operations.

In addition to these common methods for constructing control groups from observational data, there are two experimental approaches. The first is to take advantage of a natural experiment and the second is to design a randomised control trial.

- Natural experiments, by chance, observe a treatment and control group that have arisen out of an unusual situation where a treatment is given exogenously to one part of a population and not another. By 'exogenous' we mean the affected people had no choice in whether or not they are part of the treatment group and are not systematically different in relevant characteristics from the control group. The main problem with this method is that natural experiments occur by chance and cannot be produced on demand. Research using natural experiments therefore tends to be driven by a happy data event rather than the importance of the question and as such is typically more common in academic studies than government reports.
- The second experimental way to create a control group is for the evaluator or program administrator to randomly allocate firms to either a treatment or a control group. If these firms are drawn from the same population (i.e. same industry of a certain size in Victoria);we have data observations on each firm before and after the program; and we have a large enough sample; then we can be confident that any unobserved differences in firm performance-related characteristics will be evenly allocated across the groups. This means that confounding effects from unmeasured factors will be accounted for.

The disagreements between evaluators who use the experimental versus non-experimental approaches to control group selection– often referred to as the 'randomistas' and the 'regressionistas' respectively – are quite marked. The strengths and weaknesses of both approaches are discussed in section 4.2 below.

4.1.2 Estimation techniques

Once the control group is selected, there are several statistical treatments that can be applied to mop up any residual pre-treatment difference in the data (observational or experimental[13]) between the treatment and control groups.

- Multivariate regression analysis. If a confounding factor (a factor that causes both assignment to treatment and impact) is measured and included in the data set, then it can be statistically excluded, to give a true measure of the impact of the treatment.
- Instrumental variable regression. If a confounding factor is unmeasured and therefore not included in the data set, then the researcher may be able to identify some indicator (known as an 'instrument') of assignment to treatment that is entirely uncorrelated to other attributes which determine outcomes. Unfortunately such instruments can be hard to find.
- Regression discontinuity. If there are thresholds employed to determine whether someone receives a treatment or not (e.g. when there is excess demand for a program). For example, in order to determine which 20 companies (out of the 100 applications) will receive some R&D assistance, the Government scores each of the applications. The

[13] Data used for evaluation can be either experimental or observational (non-experimental). Observational data may be collected via surveys, accounting records or administrative datasets (such as licensing and registrations rolls).

threshold for receiving support is a score of 70. Regression discontinuity exploits the fact that applications receiving scores of 69 are very similar to those receiving a score of 71, which provides another way of constructing a counterfactual.

- Propensity score matching. Constructs an index from multiple measured confounding factors to construct a control group (to find 'otherwise identical' organisations using observable characteristics). For example, using the fact that we know the size, location, age and industry of firms participating in a program can help us find similar firms who didn't participate in the program. This approach is a sub-set of multivariate analysis.

4.1.3 How do we compare the treatment and control?

At the most basic level, evaluators could just compare program participants with non-participants in order to work out the effect of the program. In order to show how this is done, we take the reader from the most basic comparison to a more complicated one, highlighting as we go the issues involved. Conceptually, there are a number of different ways that the treatment and control groups can be compared. We could:

i) Observe the same individual in two different states of the world at the same time. This is the ideal measure but, of course, is logically impossible.

ii) Observe two firms (one who receives government assistance and one who doesn't receive government assistance) and observe them at some point in time after the program has ended. If we could find a perfect match for each program participant then we would have a good measure of the impact of the program. We can represent this mathematically. Let θ be the program impact, *Y* be the outcome for each firm (say, change in employment or exports) and subscript 1 means participated in the program (the treatment group) and 0 means otherwise (the control group). Then:

$$Y_1 - Y_0 = \theta$$

However, it is impossible to have two identical individuals or firms. If there are systematic differences between the two, then it is difficult to disentangle the effects of the policy from differences in the individual firms (which may be unobserved). We can go part of the way towards accounting for these differences if we have good measures of which differences matter. Let $\boldsymbol{X}$ be a vector (i.e. number of variables) of measured firm differences (relating to size, technology, location, growth trajectory, export focus etc.) that we have good reasons to believe will affect the outcome. Then:

$$Y_1 = \theta + \boldsymbol{X_1}\beta + const$$

$$Y_0 = \boldsymbol{X_0}\beta + const$$

and

$$(Y_1 - \boldsymbol{X_1}\beta) - (Y_0 - \boldsymbol{X}_0\beta) = \theta$$

Where we have direct data on $\boldsymbol{Y}$ and $\boldsymbol{X}$ and can estimate β via a regression (note that good regression results require large samples of data – certainly 100s of observations, 1000s if possible).

However, we cannot rule out that there are unmeasured differences between firms – such as the work culture, the skills of the workforce or informal managerial practices – that affect performance. If we call these α then:

$$Y_1 = \theta + \alpha_1 + \boldsymbol{X_1}\beta + const$$

$$Y_0 = \alpha_0 + \boldsymbol{X_0}\beta + const$$

and

$$(Y_1 - \boldsymbol{X_1}\beta) - (Y_0 - \boldsymbol{X}_0\beta) = \theta + \alpha_1 + \alpha_0 \neq \theta$$

So our final measure will not be θ but will include $\alpha_1 + \alpha_0$ as well.

iii) Observe the treated firms before and after the 'treatment' (i.e. participation in the program). This approach allows the analyst to net out the self-selection effect. Let our data be collected not for one year but for several years where *t* represents years before the program and *s* represents years after the program has ended. We also include a variable *m* which represents the state of the macro-economy which may affect the firm's performance irrespective of whether they participated in the program or their other ***X*** characteristics. Then, for a given firm 1 we have:

$$Y_{1t} = \alpha_1 + \gamma m_t + \boldsymbol{X_{1t}}\beta + const$$

$$Y_{1s} = \theta + \alpha_1 + \gamma m_s + \boldsymbol{X_{1s}}\beta + const$$

$$(Y_{1s} - \boldsymbol{X_{1s}}\beta) - (Y_{1t} - \boldsymbol{X}_{1t}\beta) = \theta + \gamma m_s - \gamma m_t \neq \theta$$

This shows that this 'time-series' approach does not disentangle the effects of the treatment from other factors (e.g. upswing in economic activity for some people due to global forces). So the measure we get from the estimation is not a true measure of θ and is biased according to the values of $\gamma m_s - \gamma m_t$.

iv) Observe both treatment and control groups both before and after participation in the program. This is called the Difference-in-Difference approach and is a combination of the above approaches such that:

$$Y_{1t} = \alpha_1 + \gamma m_t + \boldsymbol{X_{1t}}\beta + const$$

$$Y_{1s} = \theta + \alpha_1 + \gamma m_s + \boldsymbol{X_{1s}}\beta + const$$

$$Y_{0t} = \alpha_0 + \gamma m_t + \boldsymbol{X_{0t}}\beta + const$$

$$Y_{0s} = \alpha_0 + \gamma m_s + \boldsymbol{X_{0s}}\beta + const$$

$$(Y_{1s} - \boldsymbol{X_{1s}}\beta) - (Y_{1t} - \boldsymbol{X}_{1t}\beta) - (Y_{0s} - \boldsymbol{X_{0s}}\beta) - (Y_{0t} - \boldsymbol{X}_{0t}\beta) = \theta$$

This model assumes (a) the unmeasured differences between the treatment and control groups are constant over time; and (b) the effect of the macroeconomic variable (or time-varying characteristics) is the same for both the treatment and control groups. Neither of these assumptions is necessarily true. For example, it is possible that the individual firm level characteristics or behaviours are not constant over time (i.e. they vary). For example, the fact that a firm was not selected for a program may lead them to seek out private sector services to fill their need. A solution might be to get more data on factors that may plausibly be affecting performance. Another example is when the control group may be firms in the same industry as the treatment group but located in

another State (e.g. NSW). It is possible that the macro-economic factors are different there due to special events such as the Olympics or NSW government policies. This is one of the most common problems with difference-in-difference estimations and it occurs when the treatment and control groups do not share a common trend in factors that affect the variable of interest. One way to test this is to get more data on trends for both the treatment and control groups. Another solution may be to find other control groups which can provide additional underlying trends.

As we progress each step, our evaluation requires more and more data. Not just more data on the characteristics of the firm, but more observations about the firm both before it entered the program and many years afterwards. We move from using a simple cross-section of data to long panels of data.

4.2 'Randomistas' and the 'Regressionistas'

As we have hinted above, the experimental approach in social science (randomised controlled trials) is the state-of-the-art way to overcome the self-selection problem. It mimics the approach adopted in the natural sciences by randomly assigning individuals/firms to a treatment. However, to do this, the evaluator has to get the cooperation from the program administrators. This involves the program administrator effectively flipping a coin to determine whether an individual is placed in the treatment or control group. If the sample size is large enough, the evaluator can deduce that all (pre-treatment) characteristics both observable (i.e. company size) and unobservable (i.e. the skill of senior management), will be equally distributed between the 'treatment' and 'control' groups.

The question that remains is what limitations there are with regard to the ability of the experimental approach in social science? Heckman and Smith (1995) suggest that these shortcomings are still quite acute. One obvious pitfall in the social sciences is the lack of a 'placebo': in medical trials, two groups are given pills, but one is given a pill which turns out to have no active ingredient. This approach simply can't be imitated in the social sciences: the people in the control group know that they aren't receiving the treatment (it is impossible to fool them into thinking they might be receiving a treatment when they aren't).

By contrast, the observational (non-experimental) approach uses econometrics to try and deal with the self-selection problem (see section 4.1.2 above). Selection into the program on unobservable variables (which are correlated with variables of interest) is indeed a major problem. But with the advent of more (and cheaper) data, econometricians believe that this can be effectively handled within their approach either by measuring more otherwise unmeasured characteristics or by use of instrumental variables. If these methods are valid, the marginal benefit of experiments over non-experimental methods should be diminishing over time. However, econometricians seem to be more concerned than ever about self-selection on unobservable characteristics (see Ravallion 2012). And it is true that some of these characteristics maybe psychological in nature and therefore extremely difficult to capture.

In order to shed light on the pros and cons of the two approaches, the following commonly-reported critiques of the experimental approach are stated and evaluated.

Ethical issues. Some people argue that it is unethical to simply toss a coin to determine who receives the 'treatment'. Others argue that it is only unethical to conduct the trial *if we already know that the program works*. If you don't know whether a specific policy works, it is unethical i) to do nothing; or

ii) not to conduct an experiment. However, there is concern amongst some development economists that experiments have been used in areas where we do know whether the program works: for example, medical treatments (see Ravallion 2012). On top of this, there are issues about 'informed consent' since some evaluations are conducted in developing country villages where they are not asked whether they would like to be part of an experiment. At heart is the question: Will it be unethical to randomly pick which applicants get included in your innovation program?

Practical issues. Experiments can't be used in every context. For example, it is impossible to design and conduct an experiment on macroeconomic issues such as a random shock to interest rates. Of course, such an experiment could be designed and implemented in a laboratory setting, but that is not the focus of most experiments. In addition, it has been argued that a fascination with experiments may lead researchers to avoid important policy issues that can't be solved using experiments (see Deaton 2010). For example, Angrist and Pischke (2010) state that "Critics of design-driven studies argue that in pursuit of clean and credible research designs, researchers seek good answers instead of good questions". At heart is the question: What innovation issues are you addressing? Can you devise an experiment to test different intervention approaches?

Generalisability issues. Randomised experiments are typically conducted in environments with unique characteristics which may not be representative of all possible environments. Therefore, the results observed in one setting might not be generalizable to all contexts (which is often referred to as 'external validity'). Problems of this nature arise in non-experimental analysis too. But, according to Glennester (2013), experiments tend to get criticised for this shortcoming more than other methodological approaches simply because experiments have solved most of the other methodological issues! The question is whether the external validity issues are greater in experiments than in non-experiments.

Identification issues. Identification (also called 'internal validity') is the notion that the method being used is measuring causation rather than correlation (or reverse causation). In this regard, experiments outperform non-experiments. The correct weight to be applied to internal validity versus external validity (assuming there is some trade-off between the two) is unclear: many studies tend to favour striving for greater internal validity, but it is unclear at what cost this comes. However, it is clear that in economic analyses, the issues of external validity are much more acute than say in biomedical research. In other words, a bioactive agent is likely to work in Africa if previously shown to work in England. The same is not true of most economic policy intervention since the culture, institutions and norms in the two environments are quite different.

Statistical issues: there are two stages to the process of determining the 'treatment' and 'control' groups. Take a population of units (individuals/firms) from which you want to draw the two groups. The 1st stage is to select what Deaton refers to as the 'treatment panel': those individual units which are willing to be part of the experiment (which could be in a specific location). The 2nd stage involves randomly allocating each of the units in the treatment panel to the 'treatment' or the 'control' group. One of the virtues put forward by advocates of experiments relates to the fact that they are free of (self-) selection bias. But this is only true with regard to the 2nd stage of the process noted above: in the 1st stage, it is necessary to select which units in the population will participate in the experiment and this *might not* be done randomly. For example, it might be that only some units are suitable to be included, there might be cost issues that preclude some units being involved or some units might not want to participate. The selection of suitable units to be included might be

correlated with a variable of interest in the estimation (which would invalidate the 'bias free' status of experiments).

Also note that experiments provide an average effect (not a median effect, and not a percentage of people whose position improved). So, just because a policy produces a positive effect on average doesn't mean that everyone participating in the program will experience the average effect. Of course, there is a distribution around the average: and if the distribution is spread widely (i.e. there is a high variance), the performance of a given individual could be much better (or much worse) than the average. "Just because it works on average does not mean it works for all" (Deaton 2012). Indeed, a result which showed that there was an average positive effect of a program could be dominated by a few winners (who win big) and many who fare much worse, which appears to be an issue for randomised controlled trials which report *average* effects. However, as Imbens (2010) points out (following Manski 1996), a social planner could always compare the average effects with/without treatment and the change in the dispersion of the effect with/without treatment. This then comes down to a matter of philosophy about the correct metric to use when evaluating whether a policy works.

Substitution issues. One final issue relates to the behaviour of the members of the control group. In some situations, it is possible that they will seek out alternative substitutes to the treatment (since, as we noted above, one of the weaknesses of experiments in social science in that there is no placebo given to the control group). That is, if they believe that they have been 'denied' a potentially-valuable treatment, they will seek out an alternative. This potentially dilutes the experiment since the control group has now modified its behaviour from the desired neutral set-up intended by the experiment– it has been 'pseudo-treated'.

4.3 Choosing the Evaluation Approach

Within the scope of credible evaluation methods, the decision over which method to choose depends on the following.

For *ex post* evaluations:

- The calibre of existing data – especially whether baseline and longitudinal data exist and how many measureable confounding variable data are available;
- How easy it is to quantify program outcomes;
- How the program was run – whether there is an obvious, natural control group;
- The budget;
- The lapsed time since the first cohort of program participants completed their program relative to the expected time frame for program effects.

For *ex ante* evaluations:

- The (political) potential to run a randomised trial and other constraints on capturing data from a control group;
- The budget;
- The lapsed time since the first cohort of program participants completed their program relative to the expected time frame for program effects.

All methods of evaluation have costs and benefits. While randomised control trials tend to be held as the most 'rigorous', they can be expensive to operate, difficult to negotiate and take a lengthy

period of lapsed time to undertake. By rigorous we mean how certain we are that the estimated program impact is accurate. The main advantage of other methods is the cost and convenience of being able to deliver a result quickly. While a randomised control trial, for example, may give an impact estimate that is 99% certain, a difference-in-difference estimate may be 80% certain. In some cases, the latter is all that is required for good policy.

In the debate over the rank-ordering of different evaluation methodologies, Guido Imbens makes the following point: "I do not want to say that, in practice, randomized experiments are generally perfect or that their implementation cannot be improved, but I do want to make the claim that giving up control over the assignment process is unlikely to improve matters" (Imbens 2010, p.412). In other words, it is hard to mount a convincing case that giving up randomisation will unambiguously improve the state of policy evaluation practice. So, if randomisation is possible, it should be strongly considered.

However, it is important to note that there are some constraints to the use of each methodology: that is, certain methodologies are appropriate in one context but not in others. Moreover, it may be the case that an experiment is not the most cost effective way to proceed since experiments can be costly. Of particular interest is the fact that experiments might be difficult to implement when it comes to innovation policy. Since experiments rely on randomisation of allocation of the 'treatment', it is hard to imagine how a government could *force* a company to participate in a program (note that it is essential for the process of randomisation that participation isn't determined simply by those who want to participate).

Case study: improving DSDBI industry-policy evaluation

The current project between the Department and the Melbourne Institute is a good example of a program evaluation which would be considered to provide rigorous evidence on any evaluation hierarchy. In this project, the Department has compiled information about the companies that have received support through several of its innovation programs in recent years (including their ABN, how much they received, and when they received the support). Once linked to ABS data, this can then be used to conduct difference-in-differences analysis on both a treatment and a control group, thereby providing powerful evidence on whether participation in the program has been successful.

Rigorous evaluation of policy programs requires detailed information on the firm, which has often proved problematic. After many years of negotiation, the Melbourne Institute has brokered a deal with the ABS whereby they will provide a 'test file' for two of their datasets (the Business Longitudinal Database and the BAS-BIT) which will essentially confidentialise the data. The Melbourne Institute would be able to interact with the 'test file' – which has exactly the same file structure as the original dataset, but with synthetic numbers – in order to write our statistical analysis program which would compute the effects of participating in the program (by comparing the firms before and after participation). We can then send our computer program files to the ABS to run on the real data file (so, confidentiality is safeguarded by the fact that the Institute never sees/interacts with the real data file). Access to these ABS (and ATO) datasets are the most efficient solution to the problem: rather than building new datasets, we can simply rely on the substantial investments made by the ABS (and ATO) to collect firm-level information.

4.4 Examples of Best Practice Evidence

Given the above framework (and techniques) for evaluating the quality of policy evaluations, it is worthwhile providing some examples of best-practice from around the world. One area in economics where the best-practice frontier has been moving is development economics (see Banerjee and Duflo 2011, for example, on the use of randomised controlled trials). In the innovation policy domain, the examples of best-practice are few and far between, primarily because of the difficulty associated with accessing firm-level data (i.e. unit-record data). For example, if you want to examine the impact of policy on a firm's productivity, this requires detailed information on the firm's inputs and outputs over time (before and after the policy). Such data are not typically publicly available, so the analyst is required to access confidential data (such as the information collected by the ABS) or proprietary data (such as Compustat or IBISWorld). As illustrated above, access to confidential ABS data is slowly improving.

4.4.1 Examples of high-quality evidence leading to policy reform

Below we outline two other areas of Australian policy reform which have benefited enormously from the application of high-quality evidence-based economic research: the reduction of trade barriers and social reforms. In addition, we provide examples where firm-level analysis has raised the understanding of productivity and industry dynamics in the US, an experiment in Indian textile firms which has deepened our understanding of the impact of managerial practices on performance, the effect of minimum wages on unemployment, and studies of the effects of early childhood interventions.[14] These reforms would not have been possible without rigorous, systematic evidence of their effects.

Reduction of Trade Barriers

From the 1970s to 2008, a succession of reports, papers and inquiries by academics (most particularly Max Corden, Richard Snape, Ross Garnaut and Peter Lloyd) and government bodies (Tariff Board, Industries Assistance Commission and Productivity Commission), documented the cost to the Australian economy of high tariffs; analysed the economic consequences of a reduction in tariffs; and injected objectivity into the debate and disseminated this information to the wider community. The average effective rate of assistance to manufacturing was 35 per cent in 1969-70, but had been wound back to below 5 per cent by 2012.

The process of tariff reduction was long and complicated: there were many vested interests from the labour movement to industry. The first part of the process of policy change was to present objective evidence on the actual size of effective tariffs. However, these calculations were not enough. Subsequently, the (then) Industries Assistance Commission developed quantitative models to analyse the economy-wide consequences of policy and policy changes for economic activity and employment, as well as for regions, sectors and individual industries. These models were used to make estimates of the potential gains from reducing tariffs. Work by academics and modelling by the Bureau of Industry Economics, Industries Assistance Commission and the Productivity Commission led to consultative processes and gave governments confidence to gradually dismantle trade barriers. Successive governments have used the reports and research of the Productivity Commission to raise the level of community debate on this issue. So, this reform process happened gradually rather than as the result of one perfectly designed experimental evaluation.

[14] Some of these examples are drawn from Palangkaraya et al. (2012).

Social Reform

The second example to highlight is the application of the Household Income and Labour Dynamics in Australia (HILDA) survey to inform a wide range of social and economic policies. This dataset is a longitudinal survey of members of nationally-representative Australian households: that is, it follows the same individuals in a household over time in order to understand how their workforce participation, income and other factors change over time. This dataset has been used in a variety of quantitative analyses designed to determine the effects of specific government programs. Methods used to analyse the HILDA data are primarily econometric analyses, which rely on the law of large numbers to find average effects across the representative sample of Australian households contained in HILDA. More specifically, the data contains enough information to construct appropriate 'treatment' and 'control' groups, has a very low attrition rate amongst survey participants, and has a long-enough time period to conduct 'before and after' studies. Together, these facts suggest that well-executed regression studies can produce robust counterfactuals. So, this is an instance where there availability and quality of data largely determine the appropriate evaluation method.

To date, HILDA has been used to inform policy in the following ways[15]:

- The Productivity Commission found that mothers who are not entitled to be paid maternity leave, struggle financially. As a result, the Australian Government introduced a comprehensive Paid Parental Leave Scheme for new parents who are the primary carers of a child born or adopted on or after 1 January 2011.
- The Australian Social Inclusion Board analysed trends in family joblessness in Australia and identified the main factors that had driven these trends. This research also discussed the relationship between family joblessness and income poverty and other forms of disadvantage.
- The Pension Review, as part of the broader Tax Review, used HILDA to develop a comprehensive understanding of what pensioners lives are like. This work was undertaken by the Department of Families, Housing, Community Services and Indigenous Affairs.
- The Department of Education, Employment and Workplace Relations examined the characteristics of low-paid jobs. They found that low-paid jobs were not necessarily an end in themselves, but can provide a bridge to higher paid jobs. This information was included in a submission on minimum wages to the Australian Fair Pay Commission.
- The Reserve Bank of Australia looked at the level of debt that households have entered into and their ability to repay that debt.
- The Productivity Commission investigated the role of casual employment in the workforce and found it is often a stepping stone into longer term employment.
- The Australian Institute of Family Studies considered the financial consequences of divorce for older Australians and the subsequent implications for their retirement incomes.
- The Department of Families, Housing, Community Services and Indigenous Affairs used the data to contribute to a report on child custody arrangements to the House of Representatives Standing Committee on Family and Community Affairs. They also used the data for policy development in the areas of workforce participation and retirement.

[15] We thank Michelle Summerfield for providing these examples from the brochure 'Living in Australia HILDA'.

Industry Dynamics and Productivity Growth

The third example comes from the US. Researchers have found that most productivity growth occurs from the exit of less productive workplaces and entry (and growth) of high productivity workplaces – rather than the transformation of low productivity workplaces into high productivity workplaces (see the large volume of empirical work based on US micro data conducted by researchers at the US Center for Economic Studies including Baily, Hulten and Campbell 1992; Davis and Haltiwanger 1990, 1992; Doms and Dunne 1994; and Lichtenberg and Siegel 1987). This research was only made possible via access to unit-record data on firms within the US Census Bureau.

Bloom et al. (2013) undertook a randomised control trial on large Indian textile firms to test the effect of management consulting practices on plant-level productivity. They provided free consulting on management practices (funded by Stanford University and the World Bank) to randomly-chosen treatment plants and compared their performance to a set of control plants. The population of firms was selected from a certain industry and employment size (giving 66 firms). Firms were contacted by telephone and invited to take part in the project (34 agreed). Interviews were conducted on 96 control firms which were assessed as being no different from the treatment firms in terms of relevant characteristics. They found that adopting these management practices raised productivity by 17% in the first year through improved quality and efficiency and reduced inventory. The better-managed firms grew faster and their improved managerial practices spread to their other workplaces.

Minimum Wages and Employment

One of the most well-known difference-in-difference evaluations is the study by Card and Krueger (1994) who studied the effects of raising the minimum wage on employment. To select their control group they took advantage of the fact that the minimum wage had increased in one US state and not its neighbouring state. Card and Kruger compared the difference between two pre-treatment employment effects and then differenced it from two post-treatment estimators as discussed above. Their results showed a small increase in employment in the state with the minimum wage increased. This result was met with outrage from the economics community who thought employment should fall. The employment increase in the State with the minimum wage increase makes it hard to accept the hypothesis that employment actually decreased over this time. Although this study is still controversial, it helped change the common presupposition that a small change in the minimum wage from a low base will always cause a decrease in employment.

Early Childhood Interventions

There have been numerous evaluations of early childhood interventions (such as kindergarten, literacy and numeracy coaching) using the treatment and control group methodology in order to understand their effect. The attached document provides a meta-review of these evaluations http://www.aifs.org.au/institute/pubs/resreport14/aifsreport14.pdf.

5. Reviewing Victorian Innovation Program Evaluations

There are a range of programs that have been instigated in Victoria since 1999 – as part of a series of Innovation Statements – some of which have been evaluated by external management/economic consultants. The purpose of this section of the Report is to critically review some of these evaluations to see how they compare to the gold standard and to provide some simple

recommendations on ways in which innovation program evaluation could be improve in the future. Our focus is on the economic impacts of the programs rather than the administrative efficiency as the benefits of the program are of primary interest (the costs of implementation and administration are really secondary concerns once it has been established that there are significant positive benefits).

It is important to bear in mind that the ideal standard is often impossible to attain from an *ex post* evaluation. If the evaluation isn't designed carefully *ex ante* – and the data required to perform the evaluation aren't collected – then it is simply impossible to conduct an evaluation which unambiguously determines the effects of the program. This should make it clear that the intent of this section is not to criticise the consultants who have undertaken the evaluation – in most instances, they can only do what is possible given the data they are provided (everything else is outside of their control). Moreover, in many instances, the evaluation is an 'interim assessment' which is done a few years after the program commenced which makes it even more difficult to draw definitive conclusions about the program's effects.

5.1 Impact of the Science, Technology and Innovation (STI) Initiative (2008)

5.1.1 Program Description

This initiative represents a major undertaking in Victoria's science and technology investments (in fact, it represents the largest such investment by an Australian State Government). It represents a substantial ($638m) investment in a range of infrastructure and human capital projects including the development of the Bio21 Institute, the Australian Synchrotron, and support for successful projects in the Commonwealth Government's National Collaborative Research Infrastructure Strategy (NCRIS) program. The linkages between investment in infrastructure and human capital are clear – it is much easier to attract (and retain) the best and brightest researchers in the world when you have the best facilities available.

There were 135 projects funded since 1999/2000 which were selected on the basis of achieving Victoria's objectives to become a national (and international) science leader. Evaluation of the initiative was primarily done using data collected by the Government's Outcome Monitoring Tool (OMT), interviews with stakeholders, surveys of funding recipients (53 observations in total) and case studies.[16] This was also augmented by the use of a macroeconomic computable general equilibrium (CGE) model which was used to examine the State-wide economic effects of the initiative.

There were 5 core outcome areas on which the initiative's performance was measured against:

- *Collaboration outcomes*: international and university-industry collaborations.
- *Science awareness outcomes*: science information sessions in schools, hits on websites, newsletters and e-bulletins.
- *Skills base outcomes*: attracting elite researchers (e.g. Federation Fellows), postgraduate students, and other professional development activities.
- *Commercialisation outcomes*: export contracts, licensing agreements, and patents.

[16] Note that previous reviews of this initiative were undertaken in 2003 and 2005 by a different group of consultants. The analysis reported here only discusses the 2009 review.

- *Scientific research outcomes*: scientific journal articles and discoveries.

5.1.2 Program Evaluation

The program evaluation report finds evidence that the initiative has achieved excellent results on each of these 5 core outcomes areas. With regard to international collaboration, for example, the evidence suggests that activity has increased from 253 collaborations in 2003 to 1,965 collaborations in 2005 and over 3,000 in 2009. In terms of commercialisation outcomes – which are another important source of potential outcomes – the evaluation finds that (up to June 2008) the initiative has generated 1,750 export contracts (worth $173m), 575 provisional patent applications (primarily a mix of Australian, US and PCT applications), and 97 exclusive and 604 non-exclusive licensing arrangements for IP. In an effort to benchmark these outcomes against similar projects, the report looks at commercialisation outcomes for publicly-funded research organisations in 2002 and the Collaborative Research Centres (CRCs) in 2002. Across the range of commercialisation outcomes per $m of expenditure, the STI initiative programs have out-performed other comparable programs. However, it is correctly noted that there are long and variable lags in the commercialisation of most inventions given the different risks, objectives, regulations, etc. in each commercialisation context. These factors inhibit the comparison of the nature and speed of commercialisation environments.

Moreover, the evaluation found that the initiative has generated large benefits for Victoria. For example, over the period 2001-2014, the initiative is predicted to generate an additional $1.7bn in gross state product. Given that many of the expected benefits will accrue over a longer time period, the report suggests that the reported benefits are a lower bound on the long-run effects. To model the State-wide effects, the Monash Multi-Regional Forecasting (MMRF) model was used and two scenarios were modelled, both of which found positive net benefits of the initiative (relative to the counterfactual of 'no investment'). Scenario 1 includes 'only those benefits which were presently occurring and that could be directly quantified'. The predicted benefits were simply an extrapolation of currently observable benefits. Scenario 2 included 'realistic projections of some growth in currently observable benefits'.

The macro-modelling exercise included two types of quantifiable effects:

i) **Expenditure effects**. These occur when funds are leveraged into Victoria as a result of the initiative i.e. there is a positive expenditure effect of the initiative if these funds would have been expended externally (in other states) in the absence of the initiative. For example, initiative projects that lead to successful CRC project grants that are headquartered in Victoria.

ii) **Investment effects**. These occur when the initiative generates productivity or commercial benefits for firms and society. These benefits could accrue as a result of improved skills, innovation-led productivity growth, increased knowledge adoption rates or increased commercialisation revenues (industry or university).

5.1.3 Evaluation Critique and Recommendations

The evaluation tackles a difficult issue: how to assess the macro (State-wide) effects of this large-scale initiative. This is clearly not easy to do convincingly. The approach adopted rests largely on the simulation methods employed by a computable general equilibrium model (the Monash Model). The investment effects used in the model were determined using surveys and interviews of funding recipients. Where necessary, respondents were contacted directly to determine how their

productivity estimates (and/or commercial revenues) were generated. However, it is not clear how robust these estimates of economic effects are. And indeed it is not really possible to understand why the effects in this report (an increase in gross state product of $1.7bn) are actually lower than the predicted benefits in the 2005 report of the initiative (methodology may play a part – as claimed in the later report – but it may also just be that the result is sensitive to assumptions in the model).

With regard to the collaboration outcomes, the evidence provided suggests that this has increased dramatically. However, it must be noted that this most likely seriously over-estimates the magnitude of the effect attributable to the initiative. For instance, it isn't clear that the data disentangles collaborations that were a result of the initiative from those that were would have occurred in the absence of the initiative. That is, there was no attempt to establish a counterfactual. Moreover, collaboration has been increasing across all areas of science: this contemporaneous growth in scientific teamwork (often across international borders) is difficult to separate from the effect observed in the report.

In conclusion, this report should be classified as a monitoring report rather than an evaluation. Unless a good attempt is made to establish a counterfactual, we should not call a report an evaluation. Although the effects of these large-scale projects on the Victorian economy are very hard to achieve, there must be some consideration of the counterfactual. In general, it is far more convincing to rigorously examine the effects of a few elements of the program (e.g. its effect on collaboration) than it is to poorly examine the effects of the entire program.

5.2 Interim Evaluation of the Victorian Life Sciences Statement: Healthy Futures (2013)

5.2.1 Program Description

The Healthy Futures program represents a $230m capital investment into Victorian medical research infrastructure in an attempt to enhance the wellbeing of Victorians. Via this program, a wide range of strategic capital works has been undertaken, which have provided facilities, training and enabling technologies in an attempt to foster enhanced innovation in Victoria. The specific objectives can be stated as follows:

- Leverage Victoria's competitive advantage in specific research areas;
- Grow the reputation of Victoria's research institutes in order to expand investment and generate high quality jobs;
- Continue growth in collaborations and partnerships;
- Commercialise medical research, thereby bringing benefits to business and industry;
- Hasten the translation of scientific research into clinical practice.

The *Healthy Futures* program took advantage of a 'window of opportunity' with regard to availability of capital and support for such an initiative at both the State and Commonwealth level. Given the rapid growth and quality of the Victorian medical research environment, this program was designed to realise future potential economies of scale and scope, and to leverage increased collaborative research endeavours.

5.2.2 Program Evaluation

In evaluating the appropriateness, effectiveness and efficiency of the program, three different time horizons were considered in the program evaluation report: Horizon 1 (2006-2012), Horizon 2 (medium term) and Horizon 3 (thereafter). This is designed to capture the long lags embedded in research infrastructure investments. The key pillars of the evaluation were:

i) *Literature review*. A review of relevant policy statements, reports and the literature.
ii) *Stakeholder consultation*. Face-to-face meetings with program beneficiaries including medical research institutes, government agencies and tertiary health service providers.
iii) *Data collection*. Evaluation surveys, bibliometrics, funding data, patents, licenses, invention disclosures and spin-off companies.
iv) *Quantitative analysis*. CGE models were used to estimate the indirect impacts of increased investments associated with the program.
v) *Qualitative analysis*. Cases studies of behavioural change and impacts.

In terms of financial leverage, the program was successful in leveraging more than $500m from non-Victorian Government funding sources that is unlikely to have been invested in the absence of Victorian Government (i.e. *Healthy Futures*) funds. These leveraged funds came in the form of: matching funds provided by the Commonwealth Government (e.g. WEHI infrastructure and the new Austin-Burnet Institute), co-investment from CSIRO (e.g. bioprocessing facility at Clayton), co-investment from philanthropic funds/trusts (e.g. Atlantic Philanthropies), and co-investment from universities (e.g. Monash University's Australian Regenerative Medicine Institute). On average, *Healthy Futures* projects leveraged $2.20 for every dollar of Victorian Government funds invested. These investments have created new infrastructure which is creating greater access for researchers in university and industry to the world's best equipment.

A general equilibrium model of the Australian economy – which was based on the Monash Multi-Regional Forecasting National Reform Agenda (MMRF-NRA) model developed by the Productivity Commission – was used to evaluate the macro effects of the *Healthy Futures* investments. The model identifies 53 sectors in 8 Australian states/territories. Capital stock is increased in 2 specific sectors – research and technical services, and business services – as a result of the *Healthy Futures* program of investments. A two-step approach was used in the model of the effects of higher capital stocks in medical research. Step 1 involved running a simulation of the effects of medical research capital for the research/technical and business services sectors. Step 2 involved examining the effect of a shock on research/technical and business services on medical research capital. Using this approach, it is concluded that the *Healthy Futures* program has increased Victorian gross state product (by $170m p.a.) and household consumption (by $77m p.a.). Under the following assumptions – an economic life for the investments of 30 years and a discount rate of 5 per cent– the present value of the increase in gross state product is approximately $3.2bn (in 2012 dollars) and the increase in household consumption is $1.46bn (2012 dollars).

5.2.3 Evaluation Critique and Recommendations

Aside from bibliometric data (publications and citations) and stakeholder consultation data (interviews), the program evaluation relies on the Monash Model for most of the quantitative analysis. Although this approach is valid, it is essentially a black box: it is almost impossible to critique because the details of the model are opaque. It is also highly aggregated in that it takes all of

the different components of the program and bundles them up into one. It then estimates the effects of these at a macro level. This black-box approach makes it hard to 'sell' results to central agencies who regularly receive reports with opaque modelling outcomes that favour the program of the program advocate.

Although such information is important, what is of more interest to the Government is to undertake the effects of specific programs, and to identify which aspects of the program work (and which ones don't). For example, some of the components of the Healthy Futures program leverage other funding sources (e.g. CSIRO, philanthropy, etc.). What would be interesting to know is: are these co-funded programs more successful than programs that are not co-funded? And is there a threshold above which co-funded programs are much more likely to work? That is, perhaps programs with less than 10 per cent co-funding are not more likely to be successful than programs with no co-funding. These types of interesting (and important) policy issues cannot be addressed using macroeconomic modelling approaches such as that embodied in the Monash Model. That is not to dismiss the importance of the macro-perspective, just to suggest that there is much more detailed evaluation work that could be done to help improve future policy initiatives.

5.3 Interim Assessment of the Small Technologies Industry Uptake Program (STIUP) (2012)

5.3.1 Program Description

This program had a number of specific objectives:

i) Increase uptake of 'small' technology (micro-, nano-, bio-) by Victorian business
ii) Assist businesses to improve productivity via adoption of these technologies
iii) Enable business to access capabilities with regard to small technologies
iv) Build links between providers of knowledge and the private sector

The program works via the use of vouchers that provide assistance to Victorian SMEs to access small technologies provided by participating suppliers. Within the voucher system, there are three specific types of voucher: feasibility vouchers (STFeas, up to $10k), technical vouchers (STTech, up to $50k) and trial vouchers (STTrial, up to $100k). Any company is only entitled to one type of voucher over the life of the program. As of March 2012, 59 vouchers have been issued across the various schemes with a total value of $2.5 million. Voucher recipients have come from agriculture, textiles, biotech, medical technologies and microelectronics industries.

5.3.2 Program Evaluation

There are four components to the interim assessment conducted: efficacy, effectiveness, efficiency and appropriateness. The first component – efficacy – related to an assessment of whether the program has delivered the desired outputs which included a range of short- and medium-term targets such as "at least 50 companies that accelerated the adoption of small technologies through the program's voucher system". The interim assessment indicates that the program was successful in achieving these targets, so no further critical evaluation of these will be undertaken here. Of more interest is the second assessment criteria – effectiveness – which considered some more difficult

effects such as the productivity implications of the program and the additionality effects (i.e. whether the project would have proceeded in the absence of the voucher).[17]

5.3.3 Evaluation Critique and Recommendations

Additionality is difficult to estimate, but it can be done. As mentioned earlier in this report, the problem for social scientists interested in mimicking 'the scientific method' is that no two organisations are the same. As a result, it is very difficult to compare the performance of two organisations and ascertain what caused the superior performance of one over time: was it the program that they participated in or was it some other (unobserved) factor? In the absence of an experiment, econometric analysis can be used to predict what would have happened to the organisation in the absence of the program. This is the standard way in the academic literature that policy evaluations are conducted. This typically involves taking an objective performance measure – profits, sales, employment, exports or productivity – and evaluating it before and after participation in the program. This statistical approach relies on large numbers of observations to provide robust average estimates of the effects of the program.

How was additionality of the STIUP program determined? Partly because of the small numbers of participants – which makes econometric analysis difficult – additionality was evaluated by asking program participants the following question: "Without the voucher, would the project have gone ahead anyway?" This type of direct approach is appealing on one level, but has its problems: not the least if which is that participants might be tempted to answer 'no' to this question simply because it has provided them with money/assistance they wouldn't otherwise receive. Thus, it is harder to have faith in their stated preferences.

In these types of studies, 'revealed' rather than 'stated' preferences are preferred. In other words, 'what actually happened' is preferred to statements about 'what you think happened'. This takes the subjectivity out of the equation. Using the stated preference approach, the report finds that 90 per cent of STTech voucher and 70per cent of STFeas vouchers reported that their project either wouldn't have proceeded or would have proceeded at a later date. Therefore, the vouchers either induced the project or sped up the project (the former is a 'strong' form of additionality while the latter is a 'weak' form of additionality). The 'strong' form accounted for roughly 50per cent of the total additionality effect in both voucher schemes (although it was slightly larger in the STTech scheme). Central agencies will typically put little weight on stated preference for obvious reasons. It is questionable whether there is value in collecting this sort of information at all.

5.4 Market Validation Program (MVP): Interim Assessment (2011)

5.4.1 Program Description

The Market Validation Program (MVP) is a $28m pilot program designed to facilitate demand-driven relationships between innovative SMEs and government departments. The basic notion is that there are a range of problems facing government departments for which there are currently no solutions in the marketplace. By stating these problems and providing money to SMEs to come up with innovative technological solutions to these problems, the government hopes to both solve a

[17] The productivity implications of the STIUP were not able to be considered so soon after the program was initiated. However, voucher recipients were asked about whether the scheme enhanced their innovation capabilities – which is seen as a precursor to potential future productivity – and the vast majority reported higher innovation capability scores.

domestic problem and stimulate innovation within domestic firms (which could be then be sold to other state government departments in Australia or overseas). Such demand-side (or procurement-based) innovation programs have been successful in the US (e.g. the SBIR) and have become increasingly popular in recent years around the OECD (see OECD 2011). There are three distinct stages to the MVP design:

1. *Technology Requirement Specification*: submitted by the government department (the 'host') with regard to the specific problem. Once approved, this specification is then released to the market where firms can respond.
2. *Feasibility Study*: a grant of up to $100,000 is provided to firms to develop a feasibility study on identified projects.
3. *Proof of Concept*: $1.5m is provided over 2 years to successful feasibility studies.

At the completion of these three stages, the SMEs are encouraged to commercialise the prototype using any IP that might be generated: there is absolutely no guarantee that the government will then purchase the final technology. There were 128 technology requirement specifications put forward in total. There have been two rounds of funding thus far involving 21 projects: there are 9 'Proof of Concept' projects (in Round 1, $13.4m committed) and a further 12 'Feasibility Studies' (in Round 2, $1.2m committed). Of the 9 projects that proceeded to the Proof of Concept stage, only one was completed at the time of the evaluation.

5.4.2 Program Evaluation

A framework was developed and applied to the MVP in order to evaluate whether the program should be continued. The framework was designed to:

- Define the characteristics of 'success';
- Provide a link between individual project outcomes and macro objectives;
- Track MVP outcomes over time;
- Adapt to changing conditions and lessons learned.

Aside from desk reviews of other demand-driven programs (e.g. SBIR assessments) and other relevant evidence (e.g. Commonwealth government reports on factors influencing innovation novelty), the evaluation relied on a mix of qualitative and quantitative research. On the qualitative side of things, this related to interviews with stakeholders including the MVP project team, DBI[18] managers, hosts and SME program participants. On the quantitative analysis front, an online questionnaire of hosts and program participants was conducted and the data analysed (38 observations).

5.4.3 Evaluation Critique and Recommendations

In the assessment of the MVP, it was found that the program is achieving the goals stipulated by the DBI: it is delivering new IP, exposing SMEs to new markets, facilitating new collaborations, and expanding R&D capabilities. With regard to the 'hosts' – that is, the government departments – there are other benefits: improving innovation capabilities, and delivery of new (or improved) technologies which are expected to improve productivity. It was also stated that "…90% of hosts

[18] The Department of Business and Innovation is now known as the Department of State Development, Business and Innovation (DSDBI). For convenience, we refer to the Department by its name at the time of the evaluations i.e. DBI.

indicated that projects wold not have gone ahead without the MVP support", but it seems highly unlikely that the hosts are in a position to make an accurate evaluation of the counterfactual. As an alternative, you could ask the company themselves, but they have a vested interest in stating they wouldn't have developed the project without the support. This suggests that 'stated preferences' (either by the purchaser or the provider) about whether the project would have proceeded in the absence of government support are likely to have little value in determining whether a program has been successful.

In the quantitative assessment of the outcomes of the program, there are a few interesting results which are worthy of further consideration. For example, the impact of the program on increased sales opportunities. The survey results indicated that every participant expected sales to increase in the following 5 years. However, the aggregate estimates for global sales range from $1,119m to $1,162m, which seem unrealistically high, and there is no information provided on how these estimates might have been calculated. In addition, the evaluation notes the very high success rates of projects in the program. As the authors of the evaluation report note, it would *extremely rare* for all projects to be successful as claimed: most other similar programs show more like 10-30 per cent success rates (if lucky). This observation further questions the credibility of self-reported estimates used in evaluation analyses.

In conclusion, this report should be classified as a monitoring report rather than an evaluation. One way in which this evaluation could have been improved would be to compare the outcomes of the participants with the performance of those who applied for the scheme but were unsuccessful. This would enable the construction of a counterfactual on outcomes like: would the project have continued in the absence of the funding? However, it isn't clear whether information on the non-participants was collected in this instance.

5.5 Interim Assessment of the Victorian Science Agenda (VSA) Initiative (2013)

5.5.1 Program Description

This program was designed to bring university and industry together to achieve innovation outcomes. This report assesses the initial outcomes associated with two elements of the Agenda initiative: VSA Investment Fund, and the VSA Strategic Project Fund.

5.5.2 Program Evaluation

The authors of the evaluation report are clear that this is an interim assessment and that, although there have been some initial benefits, much of the expected benefit will occur in years to come. This long time-frame makes program evaluation difficult. There are three criteria used to evaluate the program:

i) Appropriateness: the program's rationale and design
ii) Effectiveness: the program's benefits relative to its objectives
iii) Cost effectiveness: the program's cost relative to outcomes

A comprehensive evaluation would also consider a fourth criterion – Opportunity Cost Effectiveness – which would evaluate the program's effectiveness next to the other possible programs that the money could have been spent on. That is the ultimate test of a good policy: has it generated positive

returns that are larger than those that i) could have been earned in the absence of the program; or ii) by the next-best alternative program? Unfortunately, this is extremely difficult to evaluate.

The methods used to evaluate the program on each of the three stated criteria were online surveys, interviews with grant participants, and desktop review. On the effectiveness criterion, the report notes the following outcomes:

1. 14 lead agencies reported new or substantially improved products as a result of VSA participation, and 11 lead agencies reported generated new or substantially improved processes as a result of VSA participation.
2. 52 patents have been applied for and 6 licenses have been granted.
3. 18 (of 25) lead agencies reported that the VSA collaboration was successful 'to a great extent'.
4. Prizes and awards have been awarded to VSA program participants.

5.5.3 Evaluation Critique and Recommendations

Like most studies in the innovation literature, these outputs are actually intermediate outputs (not outcomes). The Government's interest in innovation is not a patent grant *per se*, but the productivity improvements, the better health outcomes and the competition benefits that the patent might induce (a point which the report notes). Of course, the benefits take some time to work their way through the system and it is hard to demonstrate the causal effect of the program on these long-term objectives. Nevertheless, some initial estimates of the economic, social and environmental benefits of the program were provided: the economic benefits reported to date were approximately $63m which are expected to rise to over $234m over the next 5 years. However, there is no information provided in the interim assessment of how these numbers were calculated, so it is unclear as to how robust these estimates are. In terms of administration costs (and cost effectiveness), DBI spent $1.91 to manage/administer the $62.5m program (which represents 3.1% of the total cost).

The report concludes with a tick for the program and provides some possible areas for improvement. From an economic perspective, some statements in the conclusion sound weak: for example, the statement that "The OECD and Commonwealth Government have found that the connectedness between public researchers and business in Australia is suboptimal" seems incorrect. Both agencies have certainly observed that Australia has much lower levels of university-industry collaboration than other countries, but it is not clear whether this is above or below the 'optimal' level of collaboration since we simply don't know the private (or social) rates of return to collaboration. Assuming there is no measurement error, it certainly seems likely that Australia could do more collaboration than it currently does, but we don't know what blockages there might be or whether the government needs to do anything to stimulate more collaboration. That requires a much more sophisticated analysis than we can presently undertake given the paucity of data in this domain in Australia.

5.6 Summary of Evaluations

This section of the report provided a detailed summary of the main findings from the evaluations of recent Victorian innovation support programs. In Table 1, a summary of the data collated, methods used and presence of a counterfactual is presented. The Table is organised according to whether the

evaluation was designed ex ante (Yes or No); the type of data that were used in the evaluation (survey, case study or observational); the analytical method used in the evaluation (experiment, regression, simulation or qualitative); and whether a counterfactual was estimated (Yes or No). This Table is meant as a guide to aid in the evaluation of the evaluations recently conducted at the behest of DSDBI, rather than a definitive account of the characteristics of the perfect evaluation.

Consistent with the framework developed in this Report, the ultimate goal of an evaluation is to consider the *causal* effects of an intervention which typically requires consideration of a counterfactual. But the appropriate method used to perform an evaluation is contingent on the environment in which the policy intervention took place, the quality and availability of existing data (whether it is longitudinal or cross-sectional, the attrition rates of survey participants, etc.) and the ethical/moral/practical considerations surrounding the implementation of specific evaluation methods (e.g. randomised controlled trials). In some situations, it just isn't feasible to implement an evaluation method which will provide clear evidence of a causal effect, but it is something that we should at least aspire to. As a result, it isn't possible to categorically provide a checklist of data and methods that is required to produce iron-clad evidence of causal effects. Instead, this must be evaluated on a case-by-case basis.

Although there are some serious challenges in trying to establish the causal effects of specific programs, it is clear from the summary that the evaluations undertaken fall a long way short of the mark: in fact, not one of the evaluations can be said to have produced robust evidence of the causal effects of the program. In other words, there is still a lot of room for improvement. However, this shouldn't be seen as a direct criticism of the authors of these evaluations: in most instances, they were only working within the existing constraints of the available data and the budget available to conduct the evaluation.

Table 1: Summary of Program Evaluation Evidence

Program	Design Ex ante	Data Type			Analysis Method				Counterfactual
		Survey	*Case Study*	*Observational*	*Experiment*	*Regression*	*Simulation*	*Qualitative*	
STI Initiative	N	Y	N	Y	N	N	Y	Y	Y?
Healthy Futures	N	Y	Y	Y	N	N	Y	Y	Y?
STIUP: Interim Assessment	N	Y	N	N	N	N	N	Y	N
MVP: Interim Assessment	N	Y	N	N	N	N	N	Y	N
VSA Initiative	N	Y	Y	Y	N	N	N	Y	N

Notes: Y? means that an attempt at constructing a counterfactual was made, but that it had some major limitations.

6. Conclusions

Evaluating the nation's economic and social policies is important if we are to continue to live in a prosperous nation. Moreover, ensuring that we spend taxpayers' money wisely and prudently is an important part of the covenant between government and the people. Once we accept that evaluation is important, the question is: how should we go about doing it? In this Report, we have outlined the state-of-the-art as it relates to evaluation methods (and the rationale underpinning them). The simple answer to the question posed is that there is no silver bullet that can be applied in all contexts. Although there are lots of good reasons to support randomised controlled trials – indeed, they are the benchmark in science – there are also lots of important limitations to their use in social contexts. These include some ethical issues and some practical limitations. That is, there is a range of interesting policy questions that randomised controlled trials probably cannot answer. These issues should not be overlooked when designing and implementing innovation policy program evaluations. However, it is probably worth invoking Voltaire's famous aphorism "Perfect is the enemy of good" (Voltaire, undated) when evaluating evaluation methods.[19]

The evaluation methods considered here are then applied to the standards used to evaluate a range of existing Victorian Government innovation programs. Overall, the standard used to evaluate these programs falls a long way beneath the gold standard set by the randomised controlled trials and the silver standard set by *ex ante* data collections. Rather than producing clean quantitative estimates of the causal effects of the government interventions, the reports by and large rely on simple interviews of participants and back-of-the-envelope calculations rather than rigorous, systematic analysis. This means they are really monitoring (or auditing) reports and not evaluations.

This is not to criticise the authors themselves – the task of program evaluation is difficult, and is made even more difficult by the fact that they have been asked to evaluate the program *ex post* and undertake an evaluation in a time-frame that does not allow for the necessary data collection. In addition, where data are not available and the programs have macro-, long-term effects, authors have typically relied on CGE modelling which lacks transparency and can be too aggregated to be realistic. Concrete, transparent evaluations of well-defined programs are probably more useful for the Government as building blocks rather than large, ill-defined and opaque evaluations. Moreover, the Government could do more to carefully design *ex ante* a data collection process that would lead to an evaluation with a sensible counterfactual. The credibility of the authors and the commissioning agents would be raised if reports that are essentially monitoring or descriptive reports are not called 'evaluations'.

Nevertheless, in the short-medium term, there is much more that could be done to improve the program evaluation process: continue to work with the ABS to obtain access to unit-record data; think carefully *ex ante* about how to design and implement program evaluation; continue to build capability within the Government with regard to undertaking evaluation; avoid spending money on *ex post* evaluations that will only produce subjective and unreliable evidence; and develop additional linkages with academics to promote the adoption of best-practice evaluation methodologies. The issues and actions noted here are not unique to Australia: many countries around the world are grappling with them. So, we should also continue to see out best practice methods and technologies

[19] This phrase is loosely translated from Voltaire's poem La Begeule: the exact text is "Dans ses écrits, un sage Italien Dit que le mieux est l'ennemi du bien" which is translated as "In his writings, a wise Italian says that the best is the enemy of the good".

and attempt to incorporate them into our evaluations. One such example is the technology used by NORC at the University of Chicago which is designed to provide remote access to secure data (see http://www.dataenclave.org/index.php/home/welcome). There is no reason why Australian data collectors couldn't implement similar technology in Australia, thereby promoting access to confidential unit-record firm data for those in government and academia. This would undoubtedly have a positive impact on the quality of future policy evaluation in this country.

A way forward

One possible solution to this is that stakeholders – State and Commonwealth government departments of innovation, the Productivity Commission, and academics – join forces to construct data infrastructure that can enable rigorous policy evaluation to be done quickly and easily. Rather than implementing a new survey – as was done in order to create HILDA, at great expense to the taxpayer – this could be done by trying to link together pieces of existing administrative data held by various organisations (ideally, the linking would be done by a unique identifier such as an Australian Business Number (ABN)). For example, information on which firms participated in Government innovation programs could be linked to patent and trade mark applications (IP Australia), sales and export data (Australian Bureau of Statistics' Business Activity Statement) and company registration data (Australian Securities and Investment Commission). Via fairly standard techniques – such as the construction of 'treatment' and 'control' groups of firms – this would enable analysis of the effects of government innovation programs on firm survival, productivity, and sales/export growth.[20]

References

Angrist, J.D. and Pischke, J.-S. (2010). "The credibility revolution in empirical economics: How better research design is taking the con out of econometrics", *Journal of Economic Perspectives* 24(2): 3-30.

Banerjee, A.V. and Duflo, E. (2011). *Poor Economics: A Radical Rethinking of the Way to Fight Global Poverty*, Public Affairs Publishers.

Banks, G. (2009). "Evidence-based policy making: What is it? How do we get it?", ANU Public Lecture Series, presented by ANZSOG, Canberra, February.

Baily, M.N., Hulten, C., and Campbell, D. (1992). "Productivity Dynamics in Manufacturing Plants", *Brookings Papers on Economic Activity: Microeconomics*, 187–249.

Bloom, N., Eifert, R., Mahajan, A., McKenzie, D. and Roberts, J. (2013). "Does management matter? Evidence from India", *Quarterly Journal of Economics* 128, 1-51.

Davis, S.J. and Haltiwanger, J. (1990). "Gross Job Creation and Destruction: Microeconomic Evidence and Macroeconomic Implications", in *National Bureau of Economic Research Macroeconomics Annual*, Cambridge, MA: MIT Press, 123–168.

Davis, S.J. and Haltiwanger, J. (1992). "Gross Job Creation, Gross Job Destruction and Employment Reallocation", *Quarterly Journal of Economics* 107, 819–864.

[20] This is exactly the approach the Melbourne Institute is adopting to evaluate the effects of DSDBI's innovation programs (Project #1 of the Research Partnership with the Melbourne Institute).

Deaton, A. (2012). "Searching for answers with randomized controlled trials", presentation at NYU Development Research Institute, March 22, 2012.

Department of State Development, Business and Innovation (DSDBI) (2013). *Science and Technology Programs: Monitoring and Evaluation Protocol*, June, Melbourne.

Doms, M.E. and Dunne, T. (1994). "Capital Adjustment Patterns in Manufacturing Plants", Discussion Paper 94-11, US Department of Commerce, Bureau of the Census, Center for Economic Studies.

Gans, J.S. and Leigh, A. (2009). "Born on the first of July: An (un)natural experiment in birth timing", *Journal of Public Economics* 93: 246–63.

Glennester, R. (2013). Presentation at "Evidence-Based Policy-Making: Meeting the Challenges", 5th July 2013, Canberra.

Heckman, J. (2000). "Microdata, Heterogeneity and The Evaluation of Public Policy", Bank of Sweden Nobel Memorial Lecture in Economic Sciences, December 8, Stockholm, Sweden.

Heckman, J.J. and Smith, J.A. (1995). "Assessing the case for social experiments", *Journal of Economic Perspectives* 9(2): 85-100.

Imbens, G. (2010). "Better LATE than never: Some comments on Deaton (2009) and Heckman and Urzua (2009)", *Journal of Economic Literature* 48 (June): 399-423.

Lane, J.I. (2009). "Assessing the impact of science funding", *Science* 324, 1273–75, 5 June.

Leamer, E. (1983). "Let's Take the Con Out of Econometrics," *American Economic Review*, 73(March), 31-43.

Lichtenberg, F.R. and Siegel, D.S. (1987). "Productivity and changes in ownership of manufacturing plants", *Brookings Papers on Economic Activity*, 3, 643–673.

Ludwig, J., Kling, J.R. and Mullainathan, S. (2011). "Mechanism experiments and policy evaluations", *Journal of Economic Perspectives* 25(3), 17-38.

Manski, C. (1996). "Learning about treatment effects from experiments with random assignment of treatments", *Journal of Human Resources* 31(4): 709-33.

NIH (2009). "National Survey to Evaluate the NIH SBIR Program Final Report", National Institute of Health, (Authors: Jo Anne Goodnight, Susan Pucie, Stephanie Karsten, Lynne Firester, Georgine Pion, April Smith, Maura Kephart), Washington DC, USA.

OECD (2011). *Demand-Side Innovation Policies*, Directorate for Science, Technology and Industry, OECD Paris.

Palangkaraya, A., Webster, E. and Cherastidtham, I. (2012). "Evidence-based policy: Data needed for robust evaluation of industry policies", A Report for the Australian Department of Industry, Innovation, Science, Research and Tertiary Education, Canberra.

Productivity Commission (2010). *Strengthening Evidence Based Policy in the Australian Federation*, Volume 1, Roundtable Proceedings, Productivity Commission, Canberra.

Ravallion, M. (2012). "Fighting poverty one experiment at a time: A review of Abhijit Banerjee and Esther Duflo's *Poor Economics: A Radical Rethinking of the Way to Fight Global Poverty*", *Journal of Economic Literature* 50(1): 103-14.