Australian Government
Department of Industry and Science

Office of the Chief Economist

# Randomised Controlled Trials and Industry Program Evaluations

WWW.INDUSTRY.GOV.AU/OCE

MELBOURNE INSTITUTE®
of Applied Economic and Social Research

MELBOURNE INSTITUTE® of Applied Economic and Social Research
Australian Government Department of Industry and Science
Office of the Chief Economist

For further information on this research paper please contact:

Evaluation Unit

Department of Industry and Science

GPO Box 9839

Canberra ACT 2601

Email: Evaluation.Unit@industry.gov.au

## Disclaimer

The views expressed in this report are those of the author(s) and do not necessarily reflect those of the Australian Government or the Department of Industry and Science.




## Creative Commons Licence



ISBN: 978-1-925092-69-1

# 1. Introduction

The objective of this Report is to scope the potential for experimental methods and randomised controlled trials (RCTs) to be used to evaluate the Department of Industry and Science's innovation programs. This will involve a review of the extant literature on the rationale, methods and outcomes of RCTs around the world. Although many of the applications considered in the literature are clearly focused on economic development and poverty alleviation issues, there is an emerging literature on industry policy which will be considered. The potential to employ RCTs in the Australian context will then be examined given the impacts demonstrated in the international arena. The scope of the analysis will cover a selection of Department of Industry and Science programs, including the Entrepreneurs Infrastructure Program (EIP), and the Manufacturing Transition Programme (MTP). The end product of this project will identify what portfolio programs may be amenable to an RCT approach and outline potential design options for a randomised controlled trial evaluation which will help inform the Department about the value of participation in new initiatives such as Nesta's Innovation Growth Lab ("IGL").[1]

# 2. Background: Rationale and Benefits of Policy Evaluation

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. 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; hence the need for evaluation. 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]

Rigorous evaluation of existing and new program proposals is fundamental to building an evidence base to inform public policy. It is important to determine what works, what doesn't work, and why, to build our knowledge on the design of effective interventions. A key challenge in evaluation is determining and isolating 'what works' since most programs are instituted in complex settings where there are many factors that influence their success (e.g. the effects of 'good programs' could be overwhelmed by factors such as a global economic downturn). Although this is arduous, there are ways in which a counterfactual can be calculated.

---

[1] According to their website, "Nesta is an innovation charity with a mission to help people and organisations bring great ideas to life". They were originally established with an endowment from the UK Government.

[2] There are many examples of 'evaluations' that are really descriptive reports since there is no counterfactual considered.

Microeconomic evaluations use a control group to separate the effects of the program from other confounding influences – e.g. the fact that businesses *choose* to participate in a program (participation is not random). 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. RCTs are not the only way to estimate a counterfactual; but, if designed and executed well, they represent an extremely clean estimate of the effects of a program intervention.

One of the main benefits of the RCT is that it enables the separation of the effects of a program from any confounding influences (especially selection influences). However, because of the small numbers involved in most business support programs, there have been very few RCTs of firms around the world (an exception is Bloom et al. 2013 on Indian firms). The major source of benefit associated with the use of RCTs comes from the improved allocation of resources. That is, by providing clear evidence on the effectiveness of a given program, they provide policy makers with information that enables them to put resources into programs that meet their economic and social objectives (rather than those that don't). In the long run, this is hugely beneficial to Australian productivity growth because resources are allocated in a more efficient way.

In an RCT applied to an industry program, the evaluator or program administrator randomly allocates firms to either a treatment or a control group. If these firms are drawn from the same population (e.g. same industry of a certain size in Australia) and we have data 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. However, there are some important limitations of RCTs. For example, it is possible that spillovers from program participation (i.e. knowledge flows from participants to non-participants) might dilute the precision of the conclusions. However, this issue is constrained in instances where the program includes direct financial assistance and where spillovers might take a long time to materialise.

One of the key objectives of this Report is to consider the case for promoting the use of RCTs in Australian industry policy, and to develop a culture of serious, evidence-based policy development over the longer term. Although the primary focus is on industry policy, we acknowledge and explore the usefulness of RCTs in other areas of public policy – such as vocational training and energy efficiency – as well. By reviewing the existing literature on the usefulness (and limitations) of randomised controlled trials – particularly their application in industrial policy contexts – the Report aims to deliver insights from cutting-edge research knowledge to the Australian environment.

# 3. Framework: RCTs and Hierarchies of Evidence

In this Report, we develop a framework that can be used to assess the suitability of the Department of Industry and Science's programs for evaluation using RCTs. This framework will stipulate the key factors that need to be considered when designing a trial and what program features are amenable to RCTs. Our framework is utilitarian in nature: although there are always winners and losers, the metric we apply to examine the effects of a policy is whether it results in 'the greatest good for the greatest number' (to paraphrase Jeremy Bentham). This framework will cover the following:

- the steps involved in assessing a program's suitability for RCT evaluation;
- ethical considerations in the conduct of RCTs;
- external validity, i.e. to what extent the findings from RCTs can be generalised;
- limitations of the findings from an RCT;
- other important considerations in RCT design such as:
  - controlling for differences in firms, i.e. what needs to be considered in order to be confident that the random allocation to either group can control for all relevant characteristics that will interfere with statistical inference.
  - the level of dosage of the intervention required to adequately detect and measure the effect of the intervention.
  - the length of time or duration of the trial needed to measure the effect.
  - the scale of the RCT i.e. the number of observations included in the study to ensure that the estimated treatment effect is accurate.
  - the timing of the trial i.e. if a trial involves small businesses, the operation of these businesses (especially self-employed ones) can be very seasonal and the choice of quarter to observe small businesses could matter.

## 3.1 Evaluation Question Types

Before embarking on a discussion of the strengths and limitations of using RCTs for program evaluation, it is necessary to outline some basics about the types (and objectives) of evaluation questions. Generally speaking, there are three distinct program evaluation questions:

1. What is the effect of the program on participants and non-participants compared to no program at all?
2. What is the effect if the program were to be applied to a new environment?[3] and

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

3. What is the opportunity cost of the program – that is, what are the benefits foregone from either i) not lowering taxes or ii) spending on another program?

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 harder 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 address this, analysts typically bring General Equilibrium models into play. 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 the uptake of solar panels in Australian households (e.g. via a Government rebate) will undoubtedly have an effect on households' consumption of electricity off the grid which might reduce jobs in coal-mining regions (the net effect on jobs is unclear since jobs will also be created in the solar panel industry), as well as an effect on consumption of other goods/services since the household's total energy bill might now be smaller.

## 3.2 The Self-Selection Issue

One of the most difficult issues in conducting a program 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 self-select into a program, then it is very difficult to disentangle the effects of the program from the effects of persistence and determination. The best way to deal with this is to construct a control group of firms from the population that is similar (using observable characteristics such as size, location, industry etc.) to the treatment group of firms but for some reason (unrelated to firm performance) has not participated in the program. This is often done with techniques like propensity score matching, which systematically finds 'nearest neighbours' to the control group using a set of observable characteristics.[4]

[4] However, this approach is not always possible. In these instances, it may be appropriate to construct a control group *ex post* and note the direction of the bias due to unmeasured confounding factors.

The performance of the treatment and control groups can then be evaluated using a range of different techniques, which may involve estimating specific parameters of interest or simpler approaches such as difference-in-differences (see below for more on this). With the advent of more (and cheaper) data, econometricians believe that the self-selection problems can be effectively handled within their approach either by measuring more previously-unmeasured characteristics or by use of instrumental variables. Despite this, econometricians seem to be more concerned than ever about self-selection on unobservable characteristics, particularly those that are psychological in nature and therefore extremely difficult to observe (see Ravallion 2012).

In addition to these approaches which use observational data, there are two experimental approaches to constructing a counterfactual: natural experiments and RCTs.

- Natural experiments: in this approach, a treatment and control group are naturally constructed when a treatment exogenously occurs to one part of a population and not another. By 'exogenous' we mean the affected people had no choice in whether or not they participated and are not systematically different from the control group. The main problem with this method is that natural experiments occur by chance and cannot be produced on demand.
- RCTs: in this approach, firms are randomly allocated to either a treatment or a control group. As long as these firms are drawn from the same population – and we can observe a large number of firms before and after the program – then we can be confident that any unobserved differences (e.g. the skill of senior management) in firm performance-related characteristics will be evenly allocated across the groups. This means that confounding effects from unmeasured factors are accounted for.

## 3.3 Strengths and Weaknesses of RCTs

The experimental approach in social science (i.e. RCTs) is the cleanest way to overcome the self-selection problem. Despite their benefits, there are still concerns about their usefulness in social science contexts. Heckman and Smith (1995) suggest that these shortcomings are 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).

Some issues raised in commentaries about the strengths and weaknesses of RCTs are stated below.

**Ethical issues**. Some people argue that it is unethical to simply toss a coin to determine who receives the 'treatment'. RCT proponents counter 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 to find out what works. 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 issue 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.

**Practical issues**. RCTs 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. 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".

**Generalisability issues**. RCTs 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 ('external validity'). Problems of this nature arise in non-experimental analysis too. But, according to Glennerster (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.

**Identification issues**. Identification ('internal validity') refers to the idea that the method being used enables the analyst to draw inferences about the phenomenon under consideration. For example, is the model suitably defined to make it possible to draw inference about causation rather than simple 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.

**Interaction (contamination) issues.** There is the potential in RCTs for the treatment and control groups to be contaminated either i) because of information flows from treatment to the control group; or ii) because the act of receiving a treatment changes the very nature of the competition between agents in the treatment and control groups. For example, in instances where there are only four potential participants in a regional assistance program – and three of these firms end up receiving assistance – it is likely that the estimated treatment effect will be overstated. This latter point is particularly relevant to industry RCTs. Although it is not always possible to solve these problems perfectly, it is possible to geographically separate participants in treatment and control groups so that information is harder to transmit (although this is imperfect given the cost of communication nowadays) and to increase the sample size of the groups so that the estimate of the treatment effect converges on the true treatment effect.

**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 first stage is to select a 'treatment panel': those individual units which are willing to be part of the experiment. The second 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 second stage of the process noted above: in the first stage, it is necessary to select which units in the population will participate in the experiment and this might not be done randomly (Deaton 2012).

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. 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.

**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 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'.

While RCTs tend to be held as the most 'rigorous' (i.e. how certain we are that the estimated program impact is accurate), they can be expensive to operate, difficult to negotiate and take a lengthy period of time to undertake. While an RCT may give an impact estimate that is 99 per cent certain, a difference-in-difference estimate may be 80 per cent 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, 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 pursued. However, it may be the case that an RCT is not the most cost-effective way to proceed. In addition, RCTs might be difficult to implement in some contexts.

## 3.4 Basic RCT Designs

The classic and simplest way of introducing randomisation into an evaluation is to do so at the initiation of a program as is done in clinical trials. This ensures that the process is likely to produce the most reliable results possible. However, this is not always possible because the frontier in the application of RCTs evaluation methods is in social settings rather than laboratories. There are a range of alternative approaches that have been developed – primarily by people working in development economics – with regard to introducing randomisation into both new and existing programs: including over-subscription, phase-in, within-group randomisation and encouragement design (see Duflo, Glennerster and Kremer 2007). In this section, we cover the basic approaches to conducting RCTs and consider some of the alternative methods that are used to introduce randomisation into the evaluation process. For an overview of the strengths and weakness of the different mechanisms, see Appendix A.

**Simple '1 treatment' approach.** This is the simplest approach where there is simply 1 treatment and 1 control group.

**Multiple treatments.** It is possible to introduce multiple treatment groups into the analysis – and then the performance of the treatment groups can be compared or that treatment group(s) can be compared against a control group. One interesting approach is to have three different groups: one that receives Treatment #1, one that receives Treatment #2 and a third group that receives both Treatments #1 and #2. A control group receives none of the treatments. So, this approach provides a great deal of flexibility with regard to the evaluation of the effectiveness of different programs.

**Over-subscription.** In some instances where it isn't possible to introduce randomisation at the start of a program – or where it is not politically palatable to do so – it may be possible to introduce randomisation (in the form of a lottery) for programs where demand outstrips supply. In other words, if there is only a finite number of slots available in a program (due to the scarcity of funds), then it is possible to allocate some (or all) of the slots on a lottery basis.

**Phase-in.** There are situations where practical or financial constraints may mean that it is simply not possible to implement a new program all at once, so it needs to be phased in over time. Determining who should be phased in at different stages of the program can be done via randomisation. This can be an effective approach to randomisation, but it does suffer from some drawbacks. For example, it won't work in instances where there is likely to be contamination between the groups in different phases such as might occur when people modify their behaviour in Phase 1 because of the expectation of being part of the program in Phase 2. In that case, participants in Phase 2 are not an appropriate group for comparison in Phase 1.

**Within-group randomisation**. It is not always possible to provide services to one group and not to another group, as is required in the clinical trial approach. In this case, it is common to use within-group randomisation which effectively means that different members of a group will receive the treatment. For example, in a school situation where a new program is

intended to be implemented, it could be that students in Grade 3 at School A will receive the treatment whereas students in Grade 4 at School B will receive the treatment. On equity grounds, it is deemed acceptable that Grade 3 students at School B don't receive the treatment as long as students in Grade 4 do. That is, what matters is that *all of the schools* receive equal treatment.

**Randomising within a bubble**. In situations where scoring is used to evaluate proposals on a merit basis, this approach is particularly useful. It relies on the fact that 'merit' is often difficult to perfectly measure and so many of the proposals that are just under the merit threshold score could actually be considered to be 'potentially fundable' (which is sometimes referred to as a "bubble"). Take an example where the merit score threshold is 70: if there is statistical noise (i.e. error) in the way the scores are determined, it is entirely possible that an application with a score of 69 is equivalent in quality to another which scores 70. Yet, only the application with a score of 70 is funded. By randomly providing funds to some of these applications that scored just below the threshold in the treatment group, you get a richer set of comparisons since you can consider the treatment group, the potentially fundable group plus a reference group (that is not funded).

**Weighted randomisation**. One of the common concerns about a pure randomisation approach to allocating resources is that it doesn't take into account the quality of the applicants. This is particularly a concern in merit-based programs where the basic idea is to allocate monies to people/firms with the best ideas/projects. Of course, there is always statistical noise in the way we evaluate the merits of different proposals (i.e. there are measurement issues which mean we don't know the 'true' underlying quality), but nevertheless merit-based programs typically involve making serious investments into scoring applications. The main benefit of the weighted randomisation approach is that it puts a 'weight' on higher-ranked proposals, thereby increasing the likelihood that they will be granted some money (which satisfies some equity considerations). And noise can potentially be reduced by the introduction of stratified random assignment, which can identify homogeneous sub-groups within a population.

## 3.5 Other Quantitative Evaluation Approaches[5]

Although RCTs are extremely useful evaluation techniques – since they provide a very clean way of estimating a counterfactual – they are certainly not the only approach to evaluating the effects of a specific policy or program. If the experimental approach is not used to construct the control group, 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

[5] This section borrows material from an earlier report Jensen, P. and Webster E. (2014) *Evaluating Innovation Programs,* which was prepared for the Victorian Department of State Development, Business and Innovation.

to performance) have not participated in the treatment. There are two methods:

  - Choose a similar firm or individual from a different location (which has a similar market environment). 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.
  - Choose a 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 component 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 have ceased operations.

Once the control group is selected, there are several approaches that can be applied to mop up any residual pre-treatment difference in the data (observational or experimental[6]) 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. Useful 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 twenty companies (out of the 100 applications) will receive some R&D assistance, the Government scores each of the applications. The threshold for receiving support is a score of 70. Regression discontinuity exploits the fact that applications receiving scores of 69 are very similar

[6] 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).

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.
- Difference-in-differences. Observe both treatment and control groups both before and after participation in the program. Then calculate the difference in the differences to determine the net effect of the program. This assumes (a) the unmeasured differences between the treatment and control groups are constant over time; and (b) the effect of 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.

# 4. RCTs and Industry Policy Program Evaluation

In this Section, we provide i) an overview of the rationale for applying RCTs to industry policy; ii) a detailed critique of the small number of existing industry programs that have been evaluated using RCTs; and iii) a careful exploration of the potential to apply experimental methods to Department of Industry and Science programs. These programs have been identified in discussion between the Melbourne Institute and the Department and include the following: Entrepreneurs Infrastructure Program (EIP), and the Manufacturing Transition Program (MTP).

Each program will be assessed according to: its suitability to evaluation by RCT; consistency with program guidelines; ethical appropriateness; timeliness; and cost. For each program identified as feasible for RCT, we will consider various design options, which may vary in size and scope. These designs would be of sufficient detail for the Department of Industry and Science to consider potential implementation of the relevant design in collaboration with Nesta's IGL.

## 4.1 Rationale for RCTs in Industry Policy

One of the major challenges for those interested in the development of effective industrial policy has been the dearth of statistical evidence underpinning it.[7] For too long, industrial policy has failed to embrace the

[7] As a side point, note that Stiglitz, Lin and Monga (2013) argue that industrial policy is actually more pervasive than most people acknowledge. In fact, they argue that all developed nations engage in industrial policy to varying degrees, particularly if you consider indirect ways in which one industry might be given preferential treatment. For example, they contend that the way in which US banks are lent money at 1 per cent but are able to buy Treasury bonds at 4 per cent is a way of supporting the banking industry. Similarly, the ways in which depreciation allowances are imposed indirectly favours some industries because of the different capital life of infrastructure

rigorous evaluation methods that have been adopted in other areas of policy including social, labour, education and health policy. This has held back the evolution of effective industrial policies and enabled critics to continue to push the view that industrial policy is simply 'industry welfare'. This is somewhat surprising for a number of reasons. First, the absolute value of evaluation in industrial policy is arguably higher than in other policy areas because the rate of support (i.e. $ per unit) is much higher. On average, the level of support for a firm is higher than the level of support for a household, which means there is an imperative from a public finance perspective to undertake analyses to ensure that the money is spent prudently. Second, the ethical issues associated with random allocation are (arguably) less acute for firms than they are for households.

Given this, there appears to be a sound basis for the consideration of the use of RCTs in industrial policy in Australia: there is a pressing need for a rigorous evidence base, and there is an international appetite for the evolution of industrial policy. On both of these fronts, RCTs can play an important role. However, they are not without their problems, so it is worthwhile taking stock of how they have been implemented in other countries and considering how the lessons learned in these contexts might translate to the Australian situation. Note, however, that the experiences accumulated thus far with RCTs in industrial policy are somewhat limited in number, but more results will come to light in the near future once the results from the first tranche of projects funded by Nesta's Innovation Growth Lab come to light in the next year or so.

## 4.2 Checklist of Requirements for High-Quality RCTs

Although there are no definitive rules that must be applied when thinking about whether a program can be subjected to an RCT or not, there are definitely some guidelines that should be considered if the RCT is to produce robust evidence[8]. In this section of the Report, we provide a brief overview of these guidelines. This should provide some guidance when considering which specific Department of Industry and Science programs could be subjected to an RCT (an issue that is addressed in detail in the next section of the Report).

The primary issues to consider are:

1. Randomisation must be feasible and conducted at the appropriate level (e.g. individual, firm, industry, etc.). In a well-designed RCT, consideration will be given to the most appropriate level of randomisation. Some examples where this is important include: whether there is likely to be contamination of the control group e.g. in a program which is designed

in different industries. Stiglitz, Lin and Monga (2013) argue that "…the question is not whether any government should engage in industrial policy but how to do it right" (p.9).

[8] This section is based on the checklist provided in the Coalition for Evidence-Based Policy (2010) and other guides to RCT design and use such as Duflo, E., Glennerster, R., and Kremer, M. (2007) and the Better Evaluation website on RCTs (http://betterevaluation.org/plan/approach/rct).

to provide information to SMEs with regard to potential export opportunities – if this information is shared with other SMEs who are then included in the control group, then contamination has occurred.[9] This is less of a concern in industry policy because the vast majority of prospective programs would require analysis at the firm level (rather than the individual or industry level).

2. Sample size should be large enough to provide statistically significant results. If the results weren't statistically significant, this could be because a) the sample size wasn't large enough; or b) the intervention had no effect. Obviously we want to be able to distinguish between these two possibilities. There are ways in which this can be done. For example, statistical techniques referred to as 'power analysis' can be used to ascertain whether the sample size was large enough.[10] These tests should be done prior to the study as part of the design phase.

3. Comparison of the control and treatment groups suggests they were similar in key dimensions (i.e. don't just rely on randomisation to achieve this). Although the randomisation process – coupled with a large sample size – is designed to ensure that the control and treatment groups are equivalent (on average), it is important to compare the descriptive statistics of the two groups to ensure that this is true. Simple t-tests of differences between the means is sufficient. This is particularly important given the possibility of non-participation bias (see point v below).

4. Exclusion of certain groups from the randomisation process should be clear and disclosed. In certain instances, it may be necessary to exclude certain groups from the randomisation process. If this is done, it should be carefully articulated and disclosed since it could bias the results (for the simple reason that exclusion means than you are drawing the treatment/control groups from a constrained population which may not be representative of the total population).

5. Non-response/participation/completion bias should be considered. Ideally, there would be some analysis which would look into systematic differences between those who participated in a program and those who were invited to participate in a program but did not. Recall that it is hard to imagine a program where a firm can be forced into participating, so this potential for 'non-participation bias' is considerable. So, this participation bias is potentially significant in just about all industry policy RCTs and should be properly investigated.

---

[9] Similarly, a study of new classroom curriculum would not simply be able to allocate students to two different classrooms because this would make it impossible to distinguish between the effects of the curriculum and the teacher. To do this would require random assignment at the classroom-teacher level and would require a sufficient number of groups to ensure that average teacher quality was the same across both groups.

[10] Statistical 'power' is determined by: the sample size, the minimum detectable effect size, the outcome variable's underlying variance, the proportion in treatment and control groups, and the intra-cluster correlation (where relevant). See Duflo, Glennerster and Kremer (2007) for more on how to calculate the power of the statistical analysis.

6. Non-compliance is considered and evaluated. This refers to the situation where a treatment group can be contaminated by participation by members of the control group. For example, one village is allocated a specific treatment and members from another village (which is part of the control group) simply travel to the treatment village in order to get access to the treatment. This is a classic example of non-compliance since it effectively undermines the randomisation process.
7. Outcome measures were valid and correlated with the underlying outcomes of interest. In instances where outcomes measures are self-reported, they should be cross-validated with official statistics (if possible). In most industry studies, this is easier than in other areas of social policy because the outcomes of interest are typically objective (profits, exports, survival, etc.) – however it may be harder with regard to outcomes like 'innovation' which is generally proxied by patent applications or R&D expenditure, although these are not perfect measures of the underlying concept.
8. Time frame was sufficient to capture the effects. Often times, specific policies might take time to show an effect so the choice of time frame should be in line with the expected lag between implementation and outcome. For certain programs – e.g. training and skills formation – it might take several years before the effects of participation in the program show up in the outcome of interest (e.g. employment).

## 4.3 International Industry Policy Programs Evaluated Using RCTs

As discussed above, there has been an increasing groundswell of support for more rigorous evaluation – specifically, the use of RCTs – in international industry policy in recent years. However, the adoption of RCTs by government agencies is still in its infancy and so there are still precious few examples of RCTs which we can analyse. In this section of the Report, we review the handful of international examples that we are aware of.

### Innovation Vouchers and Creative Credits (UK)

This is a business-to-business (B2B) voucher program that was designed to foster new innovative partnerships between SMEs and creative service providers in the UK. It used what it referred to as an RCT+ methodological approach which combined an RCT coupled with longitudinal and mixed methods in order to provide both an estimate of the size of the effect of the policy plus an examination of the underlying causes (Bakhshi et al. 2013). The logic underpinning the scheme is founded on the notion that collaborating with external partners may improve the innovation performance of firms. There are a number of mechanisms which might facilitate this effect: i) it may enhance the firm's networks thereby opening up new ideas and/or market opportunities; and ii) it may allow firms to search their technological environment more thoroughly, thereby improving access to technologies developed by other firms.

The policy experiment itself was conducted over approximately a 1-year period (September 2009–October 2010) in the Greater Manchester region.

Over the course of the program, 672 SMEs submitted eligible applications and 150 creative credits were allocated in two waves, six months apart. The allocation process was done randomly. Each of the Creative Credits was worth £4,000 and firms were also required to contribute a minimum of £1,000 to the project. Post award of the credit, firms were encouraged to identify a potential creative partner (which could not be a firm they had previously worked with) and to develop a collaborative project. In order to try and assist with the identification of possible partners, a web-based marketplace of creative firms was made available to all eligible firms.

Four sequential surveys were undertaken of treatment and control groups (the latter were made up of eligible SMEs that were not randomly allocated to the treatment group). Survey 1 was a baseline, Survey 2 was undertaken at 6 months (just after the SMEs in the treatment group had completed their projects), Surveys 3/4 were undertaken 6/12 months after the projects were completed. Although some payments were made to promote participation in the surveys, significant attrition occurred. In the control group, only 52 per cent of those firms that responded in Survey 1 remained by the time Survey 4 was completed. In the treatment group, 78 per cent of firms remained. Analysis of the firms who dropped out suggested that they weren't systematically different to those firms that remained in the sample (i.e. there was no evidence of attrition bias).

In addition to the quantitative component of the analysis, a number of qualitative surveys were performed with both the SMEs and their collaborative partners. Again, there were substantive attrition issues (at least in terms of the number of firms leaving the sample): some of these non-responses were due to the firms going out of business while some were simply due to a refusal to participate any further. In addition, Survey 3 was accompanied by two group workshops in which fourteen firms from the treatment group participated. These workshops served to contribute to the development of the questions used in Survey 4.

The initial focus of the analysis was on the important issue of additionality: how many extra business relationships between SMEs and collaborative partners were formalised as a direct result of receiving a Creative Credit? Their results suggest that there was a large increase in the probability that an SME went ahead with its collaborative project – of the 301 firms, 12 per cent in the control group went ahead with their project despite not receiving a Creative Credit, while 96 per cent of the firms that received the Creative Credit went ahead with their project. This result was similar to other schemes employed in the Netherlands and Austria. One of the ways in which Bakhshi et al. (2013) were able to extend the analysis was that their use of qualitative techniques to augment the quantitative analysis enabled them to dig deeper into the causal mechanisms underpinning this effect. They identified two factors that caused the effects: by enabling SMEs to market their company's wares more widely; and by helping them accelerate their business opportunities. That is, many of these SMEs would have waited until they had accumulated more money before proceeding with the project.

In terms of the causal effects of the Creative Credits program on sales and innovation, there was a statistically significant effect at the end of the first 6

month period: firms that were part of the program were more likely to be undertaking product/process innovation and exhibited an increase in the distribution of sales. However, these effects had evaporated by the end of the twelve months: that is, there were no statistically different outcomes between the treatment and control groups at the end of twelve months. One caveat that is noted in the study is that the additionality effects of the program may have been under-estimated due to the effect of the economic downturn that the UK was experiencing at the time.

## Innovation Vouchers (Netherlands)

An early study using RCTs in industrial policy was the 2004 Dutch innovation voucher pilot scheme which was analysed by Cornet et al. (2006). The scheme itself was designed to increase the level of interaction between SMEs and public research organisations in an effort to try and promote greater technology transfer and knowledge dissemination. The vouchers were all allocated randomly – in the first round the scheme, a total of 1,044 SMEs applied for a voucher but only 100 were successful (in the 2$^{nd}$ round of the scheme, there were 400 vouchers on offer). The voucher is a credit worth €7,500 which the SME can use to spend at designated universities, polytechnics and the Dutch national Organisation for Applied Scientific Research (which is similar to CSIRO). The rationale underpinning the scheme was that there was a shortcoming (or market/coordination failure) relating to the level of interaction between SMEs and universities. That is, there are some barriers to interaction such as the firms' limited ability to absorb (and commercialise) knowledge, inefficiencies in the capital market which prevent firms from undertaking R&D, and weak incentives for university researchers to seek out potential industry partners (see Canton et al. 2005).

To measure the effects of the scheme, the researchers observed three important dimensions of SME-university interactions:

1. the number of knowledge transfer projects;
2. the size of the knowledge transfer projects; and
3. the timing of the knowledge transfer projects.

The analysts are careful to identify the limits of the study and they enunciate a range of questions that they can't address. For instance, they acknowledge that they can't examine the long-term effects of the voucher scheme – to do so properly would require another 2–5 years (or so) of additional data which wasn't available at the time of analysis. And it is clearly the long-term effects that we should be interested in – that is, whether the scheme resulted in long-term partnerships that effectively increased the speed of knowledge diffusion. Note also that the analysis considers the number, size and timing of the interactions, but not the quality of the interactions per se. Moreover, it is hard to make generalisations from this study as it is a pilot study and it is quite possible that the firms that choose to participate in the pilot rather than the large-scale project (at some later date) are quite different from the population of SMEs. Note that this is because there is some self-selection in the process of approving the vouchers: although they were randomly allocated, firms first had to nominate their interest in applying.

Once the voucher has been awarded to an SME, they proceed with commissioning a designated research institution with a specific, applied research problem. The voucher scheme does not require the SME to contribute any funds themselves to the project (i.e. there is no requirement for matching funds). But if the work that is commissioned by the SME costs more than the allocated voucher amount – €7,500 – then the SME must pay the balance. The administrative burden of the scheme was kept at a minimal level – there was no requirement to submit a project plan, not was the SME required to disclose the question it asked or the institution it approached. Also note that the three technological universities in the Netherlands – Delft University of Technology, Eindhoven University of Technology and University of Twente – all voluntarily agreed to double the amount offered as part of the voucher scheme.

The analysts gathered information on the first round of the pilot scheme using both the application form submitted by each of the applicants and a follow-up telephone survey of a sample of both participants in the scheme (i.e. voucher winners) and non-participants in the scheme (i.e. applicants who were not successful in getting a voucher). The application form provided information on the size of the SME (turnover, employees), industry and region, whereas detailed information on the specific research assignments was only collected on firms that participated in the survey. A total of 600 SMEs were approached to participate in the survey: all 100 voucher winners and a random sample of 500 others. In terms of survey responses, there were 71 voucher winners who responded and 242 other SMEs (i.e. response rates of 71 per cent and 48 per cent respectively). Comparison of the treatment and control groups indicates some small differences between the two groups.

In terms of the results of the program, 62 out of the 71 voucher winners reported an assignment with a public research organisation, whereas 20 of the 242 'voucher losers' did. Their modelling suggests that 9 out of 10 vouchers are used, but approximately 1 in 10 firms would have commissioned an assignment even if they had not been awarded a voucher (which suggests an additionality of 8 out of 10). In terms of the value of the assignment made, it is impossible to draw any conclusions since only 1 of the 20 'voucher losers' who reported an assignment also reported the value of the assignment. Similar concerns arise in the attempted calculations of a quantitative effects associated with the timing of the assignments.

## 4.4 Nesta's Innovation Growth Lab

One of the most exciting developments in recent times has been the launch of Nesta's Innovation Growth Lab, which was announced in late 2014. The initiative – which is financially supported by the Kauffman Foundation – aims to promote the use of RCTs in innovation and entrepreneurship research via the creation of a new global laboratory which will (co) fund policy experiments around the world. The lab has two specific aims: to generate actionable insights for decision-makers and to give academic researchers the opportunity to rigorously examine the effectiveness of innovation/entrepreneurship policies.

Some examples of projects that have been funded in the first round of IGL projects include the following (NB: these examples are taken directly from the IGL website. See here for more details):

1. The Effects of Mentoring in Entrepreneurship Education (Lynn Wu, Chuck Eesley). *What are the most effective methods to match entrepreneurs with mentors within the context of (online) entrepreneurship education?*
2. This trial seeks to understand the causal effects of mentorship as an aspect of (online) entrepreneurship education. Do mentors with a more diverse network from the mentee/student have a different impact in entrepreneurship education compared with those with a relatively similar network, and do these impacts depend on the type of strategic approaches the founder chooses? Specifically, the trial will test whether mentees using non-predictive logic and flexibility have positive outcomes compared to those with predictive logic and persistence. The trial will then test how a networked mentor can complement or substitute these strategic approaches. The trial should also show how these different factors affect the result in terms of class engagement, satisfaction and even real world outcomes such as raising funding and product releases.
3. A Randomised Control Trial to Identify the Effect of Tech Incubators on Startups (Max Nathan, Henry Overman, Silva Olmo). *Is there an effect of incubator spaces on the survival of startups and their economic performance? And if so, why?*
4. Working with one of the largest tech incubators in the UK, this trial will deploy a multi-site RCT in two different cities. After pre-selection, entry into the incubator will be randomised for 100 firms per site. The experiment will then explore post-treatment outcomes including survival, recombination, and changes in post-treatment revenue, employment and level of external finance raised. Using interviews and surveys we will also explore whether different parts of the treatment vary in their effectiveness (e.g. mentoring versus peer to peer interactions).
5. Business-science links and technology transfer (Albert Banal-Estañol, Inés Macho-Stadler, David Perez-Castrillo). *What is the impact of different types of knowledge transfer activities on the number and quality of business-science interactions?*
6. Motivated by the “European Paradox” (top-notch academic research but much weaker business-science links), this trial will test the impact of two interventions to raise awareness of academic research and connect it to businesses. 300 researchers will be allocated into 3 groups, one group which gets promoted through an online platform that showcases their business-relevant academic research and a second that gets promoted through an online platform and active offline promotions in businesses, such as meetings, presentations and business R&D management. The third group serves as a control group. The RCT will test if researchers receiving passive or active support increase the number and the quality of their business-science interactions (such as contract research, joint research, consultancy, etc.).
7. Mini-Sales Accelerator with Corporate Match Making (Linda Hickman). *Can a more focused, mini-accelerator program be an effective and low-cost approach to increase startups’ growth?*

8. This trial sets out test a novel “low-cost” accelerator program with 30-40 post-launch startups earning revenues under £1million. The program will be solely focused on sales and business growth, including a corporate match making component. It will thereby aim to address a key challenge facing many startups: not being able to sustain business growth after an initial market launch, often due to lack of sales skills and access to clients. The program will consist of three two day modules (rather than the standard 10 weeks accelerator model). The aim is to identify if a structured mini-accelerator program significantly improves sales revenues for randomly selected attendees in contrast to non-attendees from the same applicant pool.

## 4.5 Innovation Performance Contracts

The Netherlands has been quite active in regard to experimentation with industry policy in recent years. In 2012, they set up an expert working group to examine the state of the art with regard to impact evaluation which had a broad range of senior people from academia (e.g. Free University of Amsterdam) and government (e.g. Statistics Netherlands, Ministry of Economic Affairs). The main objective of the working group was to analyse ways in which the direct impact of different Ministry of Economic Affairs’ policies could be systematically evaluated. The report produced by the working group (Impact Evaluation Expert Working group 2012) wasn’t focused solely on RCTs but it considered RCTs in its analysis, recognising their potential importance as a means of establishing systematic evidence of policy impact.

Of the 6 programs that they considered, the Expert Working Group put forward one which might be suitable for evaluating using an RCT: the Innovation Performance Contracts (IPC) program.[11] The IPC program aims to promote collaboration amongst SMEs who are typically ‘appliers and followers’ (i.e. imitators with ‘new-to-the-company’ innovations) rather than (new-to-the-world) innovators. The basis of the scheme is to provide a subsidy to these SMEs to collaborate with other innovative firms with no repayment obligation, but a commitment to co-financing by the SME. The subsidy offered covers 40 per cent of the total project cost, up to a maximum of €25,000 per entrepreneur, which can be used for project costs including wages, materials and the like (but not travel expenses, overheads, etc). And the minimum requirement for collaboration is 20 per cent.

In the current system, all applications for the scheme are ranked by the agency and then the highest ranking applications are accepted (as long as the budget permits) – there is no threshold such as a minimum score. Of the 50 applications received in 2012, 2/3 were granted. The Expert Working Group were asked to consider future designs of the IPC allocation system. They considered the following tradeoff: on the one hand, randomised

[11] Note that this scheme wasn’t actually evaluated as part of the Working Group’s report; rather, its suitability for evaluation as an RCT was considered. It is included here because the process of considering what characteristics of a program make it suitable for an RCT is important for our purposes.

selection provides simple robust impact assessment, while on the other hand a tender process is preferable from a policy perspective.

The Expert Working Group considered a range of different design mechanisms to allocate the funds for the IPC program including randomisation, ranking, first-come-first-served, surveys, and other miscellaneous selection mechanisms. The strengths and weaknesses of each were carefully considered and discussed. For example, unweighted randomisation potentially weakens the incentives for firms to put in a high quality proposal because the allocation process is random regardless of the application quality. In addition, the first-come-first-served mechanism might encourage low-quality firms to apply quickly in order to ensure that they got some money. In order to balance the competing aims, the group recommended the following two-stage process:

1. Use a tender process to solicit applications which can then be sorted into two groups: potentially accepted and rejected. Apply randomisation to the potentially accepted group in order to facilitate an impact evaluation.
2. Use weighted randomisation in order to promote fairness in the process so that those rated as the best applications in the tender process have a higher chance of success. But this does move the process away from 'pure' randomisation which might mean the treatment and control groups aren't statistically equivalent.

## Does Management Matter?

Although the RCT example examined in Bloom et al. (2013) isn't strictly speaking an example of 'industry policy', it has direct relevance for industry policy since it addresses the complex issue of whether differences in management practices across firms can explain differences in firm productivity levels. Given that the promotion of firm productivity is at the heart of most industry policies, this should be of direct relevance. And it clearly has direct relevance to any industry policy related to the training and development of business management skills. Although the context is set in a developing country, India, the design and implementation of the program is interesting and the results may also have implications for Australia.

The source of variance in productivity across firms and countries is a long-standing puzzle in economics. Even in contexts where firms in the same industry are essentially producing identical products (e.g. ice, cement), there are huge observed differences in productivity. This is a puzzle because economic theory suggests that firms should be able to mimic their competitors to get close to the production possibility frontier. That is, the very process of competition is predicated on the notion that firms continually minimise their costs by improving practices and processes. In areas where firms are essentially producing homogeneous goods, there should be some convergence in productivity as firms either improve performance or go out of business.

One possible explanation for the observed variance in productivity levels is that the quality of management differs across firms. But economists have not found that argument compelling since, for one thing, it is not clear what the

barriers to good management are. In this study, Bloom et al. (2013) use an RCT to help solve the puzzle. They took seventeen large, multi-plant firms (100-1,000 employees) in the Indian textile industry in 2 towns around Mumbai and randomly provided 5 months of extensive management consulting services for free[12] (from a leading international management consulting firm) to some firms (i.e. the treatment group) but not other firms (i.e. the control group). This randomisation allowed them to examine the effect of improved management practices on firm performance.

The RCT was not without its challenges, however. For example, of the 66 firms that were originally identified as 'in-scope' (and were contacted about participating), only 34 expressed an interest in participating and seventeen finally agreed to participate (which involved a commitment to provide senior management time to the project). This potentially introduces a selection bias into the analysis – which is a serious concern for RCTs in industry policy due to the fact that firms can't be forced to participate in programs (and therefore 'pure' randomisation is difficult to achieve). The researchers undertook two steps to assess the extent of any selection bias. First, they compared observable characteristics (assets, employees, etc.) of treatment and control groups and found no statistically significant differences between them. Second, they surveyed the population of textile firms in the greater Mumbai area and found no observable differences between the treatment group and the non-participating firms that responded to the survey.

There are a range of other methodological issues that are discussed carefully in the paper – such as the impact of the rather small sample size used to conduct the experiment. Given that the experiment had been designed and implemented so carefully, the results from the experiment are really quite clear and profound. The researchers conclude that receiving the management consulting services increased productivity by 17 per cent in the first year via i) improved quality, ii) improved efficiency, and iii) reduced inventory. Over a 3-year window, receiving the management consulting services led to the opening of more production plants. With regard to why the firms hadn't already adopted such practices, the authors find that there were substantial information barriers which suggest the firms aren't aware that they are badly managed.

## Other Related RCT Examples

As alluded to earlier, the pool of case studies on industry policy RCTs is quite shallow. The importance of RCTs in industry policy has only really just been recognised and there are a range of new experiments that are currently underway which will see the evidence base developing in the coming years. However, there are lots of related policy areas in which RCTs have been embraced more readily which also cast light on important industry policy issues. For example, in development economics, there are loads of recent RCTs which have considered the effects of issues such as micro-finance,

[12] The cost of the provision of the management consulting services was in the millions of dollars and was covered by research grants received by the analysts and the fact that the management consulting firm charged 50 per cent of their normal commercial rates because of the nature of the project.

export intensity, and business skills training (e.g. see McKenzie and Woodruff 2012). While it is not the purpose of this Report to provide a comprehensive overview of these cases, a number of the most relevant (not all of which are from development economics) are presented below.

**UK Growth Voucher Program**. This program – which is being tested in a pilot RCT project – is a £30 million business support program that promotes business growth by providing a subsidy to help pay for external advice. The pilot involves 25,000 firms, and will examine whether access to a subsidy encourages businesses to seek external advice, and if such advice results in better business outcomes. This government program helps small businesses get strategic business advice on:

- finance and cash flow;
- recruiting and developing staff;
- improving leadership and management skills;
- marketing, attracting and keeping customers;
- making the most of digital technology.

Businesses will be randomly chosen to get a voucher of up to £2,000 to help finance strategic business advice. The voucher can pay for up to half of the cost of the advice.

**Exporting and Firm Performance**. Atkin and Khandelwal (2011) discussed the design of an RCT in Egypt to evaluate a trade facilitation program. The program has three main components: matching firms in the handloom weaving sector to US buyers, providing assistance to improve design quality, and offering business skills training. A set of firms was randomly selected and invited to receive a combination of the three services. This RCT will enable them to: determine if increasing market access through exporting has an impact on firm performance; examine which are the key firm-specific factors that lead to export success; and explore whether export market access results in greater and less volatile income in a context of political instability. In a related study by Atkin, Khandelwal and Osman (2014), they identify the impact that exporting (Egyptian rugs) has on firm performance using an RCT and find that it increases profits by 15–25 per cent and results in large increases in productivity and product quality.

# 5. Suitability of Department of Industry and Science Programs for Experimental Evaluation

There is no definitive template, checklist or blueprint that can be used to determine whether a particular program is suitable for experimentation via an RCT. However, there are some obvious issues that any RCT must contend with if it is to end up producing credible, robust results. These can be used a guiding principles to aid in the process of determining whether specific industry program can be evaluated using an RCT. In this section of the Report, we consider a range of policies that, in consultation with the Department, were agreed to be potential candidates for randomisation.

These can be categorised as either eligibility programs, competitive programs or voucher-based programs and they include the Entrepreneurs Infrastructure Program (EIP) and the Manufacturing Transition Program (MTP). Before doing this, however, we briefly provide an overview of the key questions that guide our discussion, which include the following:

1. **Is randomisation feasible?** There are often practical, ethical and political issues which are impediments to randomisation. These need to be carefully considered before proceeding with a plan to introduce an RCT.
2. **Are there any selection issues?** Given that firms can't be forced to participate in an industry program, selection issues may be serious. There must be enough data about the relevant population that the sample is drawn from in order to be able to ensure that firms that do participate are not different from the population. If not, the resulting analyses won't be credible.
3. **How large is the sample size?** If it isn't large enough, this may call into question the results for the simple reason that large numbers are required to ensure that the treatment and control groups are equivalent, on average. Moreover, it makes generalisation (to other contexts or environments) problematic.
4. **Is the sample representative?** This is less of a concern that in non-experimental contexts, but it nonetheless can be a problem if the sample size is small and the distribution of the population is highly skewed e.g. unlike other economic units (e.g. households), an industry might be composed of a few very large firms and thousands of small ones. This could be an issue for causal inference.
5. **Is non-compliance a potential problem?** This can occur if there is contamination of either the treatment or control groups which may seriously mis-estimate the true impact of a program since there isn't clean separation of the treatment and control groups.
6. **Are outcomes observable?** Most firm performance outcomes are easy enough to measure conceptually (e.g. productivity, profitability, survival) but sometimes there are serious data shortages which may mean that the only way to track outcomes is via direct survey (which can be very expensive, although the commitment to collect the 10 core data items will mitigate this concern). Moreover, there are some outcomes, like those associated with innovation, which may be hard to measure conceptually and practically.
7. **Is the timeframe suitable?** Sometime an expected effect can take a long time to materialise so it can be necessary to make sure that the evaluation allows enough time for this to occur.

## 5.1 Entrepreneurs Infrastructure Program (EIP)

This program incorporates three distinct components – business management, research connections, and accelerating commercialisation – and is the Government's flagship firm-level program for business competitiveness and productivity. The total package is worth $484.2m and it is to be delivered through one single business service, which is supposed to

streamline the way businesses access government information and support services. It effectively provides a one-stop shop for any business looking to find out more about ways in which the government might be able to help improve their performance.

Despite the fact that the three components have quite different goals, there is a unifying big-picture objective for these components – business competitiveness and productivity. And the three components do have some similar characteristics e.g. they all provide specialist advisors who go out to businesses to help improve their performance. For example, the business management component – which is specifically designed to help free up managers/CEOs to spend more of their time on running the business (rather than working in the business) – provides business evaluation services and matched funding of up to $20,000 to engage others to assist in implementing projects recommended in the evaluation.

Given that all three components provide these tailored advisory services, and that this feature is potentially suitability for randomisation, this is the focus of our analysis here.[13] These tailored advisory services might be suitable for randomisation because advisers could potentially be randomly assigned to some businesses and not others, thereby creating suitable treatment and control groups. If there was excess demand for these services, then the randomisation could be applied after a certain threshold had been be reached since a lottery is just as good a way as any alternative method (first come, first served) at allocating a slot in the over-subscribed program. Indicators of the success of the advice would be sales revenue, productivity, profits, or business survival (although note that the Business Management component of the program doesn't focus on start-ups, so the latter indicator may be less suitable in that context).

One obvious challenge with randomisation of EIP advisory services would be disentangling the effects of the program from the effects associated with the quality of the advisor. In other words, as long as there is variation in the quality of the advisors which affects the observed changes in the business' performance, it will be very difficult to separate the effect of the advisor from the effect of the program. This wouldn't be a problem if the same advisor was providing services to every business – but this clearly isn't the case. There are a multitude of different advisors across the country providing services. This issue is more of a concern here than it was in the Bloom et al. (2013) situation – where consultants were providing services to Indian textile businesses – because it was the same consulting firm that provided services, so there would have been less variation in the quality of advice provided.[14]

---

[13] Other aspects of the Entrepreneurs' Infrastructure Programme – such as the linking/networking activities of the Commercialising Ideas component of the programme – were also considered for randomisation. The problem with evaluating networking activities is that they typically take a long time to come to fruition. That is, any effect occurs with some considerable lag.

14 In regression analysis, it is impossible to control for the effect of an individual unit (whether it be a person, a firm or an industry) via the inclusion of a dummy variable. But part of the appeal of the RCT is that you don't need fancy econometric models in order to produce robust results.

## 5.2 Manufacturing Transition Program (MTP)[15]

This is an initiative designed to provide support in the form of $50m worth of grants to help manufacturing businesses in Australia become more competitive and sustainable. The focus of the grants is capital investment projects that can be demonstrated to assist business i) move (or expand) into higher value manufacturing activities; and/or ii) build skills in higher value activities or new markets. A non-exhaustive list of 'high value' activities and their likely manufacturing sector is provided. Grants will cover up to 25 per cent of the total project cost, with a minimum of $1m and a maximum of $10m.

To be evaluated, all applications must be eligible in terms of the level of expenditure (>$4m), the types of activities, and the expenditure items. The eligible applications are then evaluated on merit according to five criteria: i) the extent to which the project represents a transition or expansion to higher value-added activities (25 points); ii) the projects' net economic benefit (25 points); value for money (20 points); demonstrated capacity to conduct the project (20 points); expected productivity improvements (10 points). The applications are assessed by an independent committee of experts and successful applications must score highly on each of the five criteria.

This type of program could conceivably be subjected to evaluation by randomisation in many different ways. However, for the sake of brevity, our focus is on a couple of different approaches which provide the best potential, which are discussed below. The most difficult challenge facing any type of randomisation in this instance relates to the self-selection issues: the only companies who will apply for the scheme are those who would like to move to higher value-added activities (let's call them 'transition firms'). Ideally, we would like to know how the sample of 'transition firms' who apply for the grants differs from the population of 'transition firms' (which includes all those who might try to make this transition without government support). But this is probably quite difficult to ascertain, so we might need to fall back on information about the population of all firms in that manufacturing sector (whether they are attempting to transition or not).

**Option 1: Weighted Randomisation**. The first option to consider is weighted randomisation, where the scores that an eligible application receives from the independent evaluation committee are used to provide weights in the randomisation process, as in the Dutch Innovation Performance Contracts case discussed above. This would ensure that the applications that are the highest ranked are more likely to receive the grants, which would be politically (and ethically) sound. As noted before, a simple randomisation process which simply allocates all applications as either 'successful' or 'unsuccessful' can weaken the incentive for businesses to put in a high-quality application since it doesn't take into account 'quality'. So, there are very good reasons to adopt the weighted randomisation approach. The main downside is that it might mean that the treatment and control groups are no longer statistically equivalent – that is there are no observed (or unobserved)

[15] Other similar (but now terminated) programs include Clean Technology, Green Building Fund, and Automotive Diversification.

differences between the two groups – which would be a problem for the estimation of the treatment effect. Recall that the main benefit of the random assignment approach is that, as long as the number of observations is large enough, it ensures that each group is equivalent.

Given that applications vary in the amount requested and that there is a fixed budget for the program, this would introduce an administrative wrinkle into the process (since it would be unknown *ex ante* how many applications will be received and how many are needed to exhaust the funds). But this could be done such that each application that is successful according to the randomisation process is awarded a grant *until the funds are exhausted*.

Following the randomisation process, all applicants could be categorised into the following groups –unsuccessful and successful – which would form the basis of the treatment and control groups. In terms of the sample size, there is a lot of uncertainty since we don't know *ex ante* how many businesses will apply, but it could be expected to be quite a large pool given the uncertainty of the future of many parts of the Australian manufacturing industry. What we do know – given that the grants must be between $1-10m each and that there is $50m to allocate – is that the total number of businesses in the 'successful' group will be between 5 (if all grants funded are $1m) and 50 (if all grants funded are $10m). From a purely statistical perspective, it would be preferable to fund a larger number of small grants since that would increase the number of businesses in the treatment group, thereby enhancing the statistical power of the analysis.[16]

An *ex post* evaluation of the performance of the treatment and control groups can then be undertaken. The choice of suitable indicators of performance could include metrics like profitability and productivity since the program is designed to have net economic benefit and lead to higher value-added products, both of which should result in higher profits and productivity. At a more micro level, metrics such as 'new product launches', 'new job creations', and 'new (export) markets' could also be used since all of these factors are mentioned as important components of the program's objectives.

**Option 2: Randomisation of Potentially Fundable Applications**. This approach is slightly different to the approach mentioned above in Option 1, but it has many of the same desirable characteristics. In particular, it is preferable to the simple randomisation process since it takes into account the quality of the applications. It works in the following way: using the assessments of the independent evaluation committee, applications would be categorised as 'always funded', 'possibly funded', and 'never funded'. The randomisation would only be applied to the applications in the 'possibly funded' category which are applications with merit but for which there might not be sufficient monies to fund otherwise (this is the approach that is often referred to as 'randomisation within the bubble'). There is a good rationale for adopting this approach – it stems from the fact that the scores by the independent evaluation committee are noisy signals of the true underlying

[16] A proper power analysis would need to be undertaken if the Department of Industry and Science was seriously considering an RCT of this program. This is clearly outside the scope of this Report.

quality of the application. The larger is the size of the committee, the closer the score will be to the application's underlying true quality. In situations like this where the true underlying quality is not perfectly observed, randomisation of those applications in the 'potentially funded' category makes perfect sense.

So, as in Option 1, this approach would have the desirable feature that the quality of the applications is taken into consideration since all of the highest-ranked applications would be funded. And by randomising some of the grant recipients in the 'potentially funded' category, there is an opportunity to do an *ex post* comparison of the performance of the various groups which will provide convincing evidence about the merit of the program. In fact, this approach provides a rich set of different groups that can be compared *ex post* using the metrics discussed above in Option 1.

# 6. Conclusions

This Report examined the potential for experimental methods and randomised controlled trials (RCTs) to be used to evaluate the Department of Industry and Science's innovation programs. In doing so, a review of the extant literature on the rationale, methods and outcomes of RCTs around the world was undertaken. Although the literature on RCTs and industry policy is still emerging, some interesting examples were discussed and critiqued. The broad conclusion from this discussion is that there is enormous potential to introduce more rigorous evaluation of industry policy in Australia and RCTs should play an important role in this. RCTs are certainly no policy panacea – as the examples presented here have shown that there are many challenges associated with the design, implementation and interpretation of findings based on RCTs – but they nevertheless provide a conceptually clear and methodologically sound way of figuring out what works (and what doesn't) in a complex policy environment.

# Appendix A Mechanisms for random assignment in industry programs

Table A1: Mechanisms for random assignment in industry programs

| *Mechanism* | *Strengths* | *Weaknesses* | *Most applicable* | *Example* |
|---|---|---|---|---|
| Weighted randomisation | Rewards high quality applicants<br>Provides incentives to produce good applications | Results must be analysed to remove impacts of weighting applied at the front end | Merit based programs with strong assessment processes that are able to produce meaningful "scores" | Dutch innovation performance contracts |
| Randomisation by region | Easily understood by stakeholders<br>Counterfactual data maybe available for certain regions (local government areas, States, statistical regions) | Differences between regions may make analysis difficult<br>Contamination may result if businesses or people move between regions to access the intervention | Implemented in programs with an eligibility stage | Distributing grants based on randomly selected local government area |
| Simple '1 treatment' approach | Very easy to understand<br>Very easy to interpret results | Not always possible to implement<br>May not be suitable for merit-based schemes | In simple settings | Most clinical trials<br>Many J-PAL microfinance examples |
| Multiple treatments | Provides great flexibility to compare different programs | Requires larger sample size | Where there are numerous related interventions | |
| Over-subscription | Provides evaluation opportunities without sacrificing political objectives<br>Random assignment is more equitable than first-come-first served | Only applies to cases where demand outstrips supply | Where program is highly popular<br>When randomisation is better than alternative allocation mechanisms | Innovation voucher schemes |
| Phase-in | Accommodates programs with staggered implementation | Potential contamination between groups in different phases | Prospect of contamination is low | Treatment and control groups in non-contiguous areas<br>J-PAL's deworming study |
| Within-group randomisation | Provides equity without sacrificing randomisation | Only applicable in a limited number of different settings | In school settings where groups are easy to identify and separate<br>Where policies are implemented at a higher level than the individual | |

| *Mechanism* | *Strengths* | *Weaknesses* | *Most applicable* | *Example* |
|---|---|---|---|---|
| Randomising within a bubble | Richer set of evaluation comparisons<br>Perfectly suited for merit-based schemes | Can be subject to a critique that any randomisation of merit based programs is unfair. | Any merit based scheme<br>In settings where scores are very noisy (i.e. a lot of error) | Innovation grants<br>R&D grants |
| Encouragement design | Provides information on likelihood of participation and program effects | Can be expensive depending on size of incentive | Where participation rates are low<br>Participation is desirable | |

Source: Melbourne Institute, 2015

# References


Angrist JD and Pischke JS (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

Atkin D and Khandelwal A (2011) “The Use of Experimental Designs in the Evaluation of Trade-Facilitation Programmes: An Example from Egypt”, p.107-122 in *Where to Spend the Next Million? Applying Impact Evaluation to Trade Assistance*, ed. Cadot, O., Fernandes, A.M., Gourdon, J. and Mattoo, A., The World Bank: Washington

Atkin D, Khandelwal AK and Osman A (2014) “Exporting and Firm Performance: Evidence from a Randomized Trial”, unpublished mimeo, November

Bakhshi H, Edwards JS, Roper S, Scully J, Shaw D, Morley L and Rathbone, N (2013) “Creative Credits: a Randomized Controlled Industrial Policy Experiment”, Nesta Working Paper, June

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

Canton E, Lanser D, Noailly J, Rensman M and van de Ven J (2005). “Crossing borders: When science meets industry”, Netherlands Bureau for Economic Policy (CPB) Document No. 98

Coalition for Evidence-Based Policy (2010) “Checklist For Reviewing a Randomized Controlled Trial of a Social Program or Project, To Assess Whether It Produced Valid Evidence”

Cornet M, Vroomen, B. and van der Steeg, M (2006) “Do innovation vouchers help SMEs to cross the bridge towards science?”, Netherlands Bureau for Economic Policy (CPB) Discussion Paper No. 58

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

Duflo E, Glennerster R, and Kremer M (2007) “Using randomization in development economics research: A toolkit”, *Handbook of Development Economics*, 4, 3895–3962

Glennerster, 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 JJ and Smith JA (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

Impact Evaluation Expert Working Group (2012). “Dare to Measure: Evaluation Designs for Industrial Policy in The Netherlands”, Final Report, November

Jensen PH and Webster EM (2014) *Evaluating Innovation Programs*, Report prepared for the Victorian Department of State Development, Business and Innovation, Melbourne, June

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

McKenzie D and Woodruff C (2012) “What are we learning from business training and entrepreneurship evaluations from around the world?”, unpublished mimeo, World Bank, Washington

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

Stiglitz JE, Lin JY and Monga C (2013) “The rejuvenation of industrial policy”, World Bank Policy Research Working Paper 6628, September