impact canada

MEASURING IMPACT BY DESIGN

A Guide to Methods for Impact Measurement

2019

Canada

MEASURING IMPACT BY DESIGN

A GUIDE TO IMPACT EVALUATION METHODS

ISBN??

ISBN
la lalalaaa llalalaaaaaa
something here... lalala

I WILL PUT THE ISBN STUFF HERE

Praesent tristique dolor eu ultricies aliquam. Donec egestas, massa non auctor ornare, elit mauris tincidunt justo, non mattis nisi nisi in ligula. Etiam ultricies sem nibh, non commodo sem mattis sed. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Duis tellus turpis, porttitor at feugiat pretium, scelerisque at odio. Nam massa ante, tristique tempor neque id, iaculis pellentesque tortor.

MESSAGE FROM THE PRIVY COUNCIL OFFICE

"We are aware that program spending is an investment that we are making on behalf of, and directly for Canadians, and we need to place a greater emphasis on understanding what differences these investments make in improving the lives of citizens."

We are pleased to share the Impact and Innovation Unit's (IIU) first set of guidelines for impact measurement, in support of its work under Impact Canada. This document is intended to be both an accessible introduction to the topic, as well as a reference for those involved in the design, delivery, procurement or appraisal of impact measurement strategies for Impact Canada projects. Drawing on best practices, Measuring Impact by Design was written to guide its readers to think differently about measuring impact than we have traditionally done within the federal public service.

In its role leading Impact Canada as a whole-of-government effort, the IIU works with an ever-expanding network of partners to deliver a range of innovative, outcomes-based program approaches. We are aware that program spending is an investment that we are making on behalf of, and directly for Canadians, and we need to place a greater emphasis on understanding what differences these investments make in improving the lives of citizens. That means we need a better understanding of what works, for whom, and in what contexts; and we need a better understanding of what kinds of investments are likely to maximize our sought after social, economic and environmental returns.

Good impact measurement practices are fundamental to these understandings and it is incumbent upon us to be rigorous in our efforts. We recognize that we are still building our capacity in government deliver on these approaches. It is why we built flexibility within Impact Canada authorities to use grants and contributions to fund research organizations with expertise in the kinds of techniques outlined in this guide to help. We encourage our partner departments to consider taking up these flexibilities.

Measuring Impact by Design is one of a number of supports that the IIU plans to provide to deliver on its commitment of improving measurement practices for Impact Canada. We look forward to continued collaboration with our partners in the delivery of these important outcomes-based approaches across the public sector.

Matthew Mendelsohn
Deputy Secretary

Rodney Ghali
Assistant Secretary

TABLE OF CONTENT

INTRODUCTION

Improving upon the delivery of results for Canadians is the *raison d'etre* of Impact Canada, a whole-of-government initiative launched in Budget 2017 and led by the Impact and Innovation Unit (IIU) in the Privy Council Office. Impact Canada accelerates the use of innovative and experimental approaches across the Government of Canada. With this as its core purpose, Impact Canada takes an outcomes orientation to everything it does. This focus on outcomes is what links each of the Impact and Innovation Unit's (IIU) core business functions. These include the application of new tools like behavioural insights in program design, the use of innovative finance to improve implementation, and the application of leading edge impact measurement methods to evaluate the extent to which those outcomes were actually achieved.

As part of its mandate to co-design innovative and experimental approaches with Government of Canada departments, the IIU works with departments to put in place a rigorous evaluation plan for any initiatives launched under Impact Canada through a project planning process. To this end, this guide emphasizes the use of rigorous impact measurement techniques that could be applied in many circumstances under the Impact Canada umbrella. *Measuring Impact by Design* is likely to have particular relevance for program and service delivery efforts that lend themselves to establishing a 'counterfactual', to understand what was achieved, over and above what would likely have happened anyway. In this way, the methods outlined in this guide can be used to better determine the causal impact that can be attributed directly to the initiative in question. Such approaches, when executed appropriately, can provide more robust evidence and greater confidence that the observed results are actually attributable to an initiative itself, rather than other factors.

Experiments or Randomized Controlled Trials (RCTs) have a long history. RCTs began to take root in the mid-1920s in agricultural science, and have since come to be used by a wide range of evaluators and social science professionals in many domains (Jamison, May 2017, p. 15). They have experienced an increase in popularity over the last decade as their use, particularly in the area of international development, has expanded.

"A central theme of this document is that, with the right planning, most programs can find an experimental or quasi-experimental method of impact measurement that requires very little or no adjustment to its normal mode of operation (Gertler, Martinez, Premand, Rawlings, & Vermeersch, 2016, pp. 188-193)."

Quasi-experiments are generally of more recent origins. Nevertheless, we are now seeing that they too are becoming increasingly common as impact measurement designs. With this guide, the IIU hopes to accelerate their adoption in Canadian contexts. As a category of methods, quasi-experiments provide a balance of rigour and pragmatism, and as such, they are well suited for the task of impact measurement. They are rigorous enough to give us plausible estimates of our impacts, and pragmatic in the sense that they can be deployed in circumstances where randomized methods are not feasible. A central theme of this document is that, with the right planning, most programs can find an experimental or quasi-experimental method of impact measurement that requires very little or no adjustment to its normal mode of operation (Gertler, Martinez, Premand, Rawlings, & Vermeersch, 2016, pp. 188-193).

KEY OBJECTIVES

This document is not a step-by-step 'how to' for executing experimental or quasi-experimental evaluations. Rather, it aims to be a key resource for policy and program staff to better understand what the main approaches are, their advantages and disadvantages, and to identify the conditions under which they can be used. Advanced training in research methods or statistics is not required for users of this document.

Its key audiences are frontline staff who oversee the design and delivery of Impact Canada initiatives. These staff are uniquely placed to proactively plan for impact measurement as a program is being designed. Proper planning at this early stage is a critical success factor for impact measurement. Without this kind of foresight, important opportunities to capture baseline data, or to create comparable groups at a program's outset can easily be missed. These kinds of oversights are common, and not easy to deal with once they occur. The IIU has established a process for working with and supporting departments to build a robust evaluation approach as a core element of the program design. Each approach is customized to the needs of the department and the circumstances and outcomes being targeted for each initiative.

The guide is structured to provide a brief overview of some key concepts related to impact measurement, as well as a non-technical overview of the main set of experimental and quasi-experimental methods for measuring impact. It aims to present their logic in an accessible way, and illustrates key concepts with examples where possible.

"The guide is structured to provide a brief overview of some key concepts related to impact measurement, as well as a non-technical overview of the main set of experimental and quasi-experimental methods for measuring impact."

Key objectives of this document are to:

- Improve program and policy specialists' knowledge of basic concepts related to impact measurement;
- Create greater awareness of the range of experimental and quasi-experimental methods available for measuring impact;
- Empower frontline program and policy staff to advocate for better impact evaluations, particularly for pay-for-success initiatives under the Impact Canada umbrella;
- Provide a key resource to enable program and policy staff to be critical consumers of impact measurement plans and reports;
- Build an awareness of which kinds of impact measurement methods are best suited to which program contexts.

THE IMPACT CANADA INITIATIVE

Announced in Budget 2017, Impact Canada is a whole-of-government effort that will help departments accelerate the adoption of outcomes-based approaches to deliver meaningful results to Canadians. Outcomes-based approaches represent a new way of managing grant and contribution funding that shifts the traditional emphasis on process and outputs, towards one where payments are tied to the achievement of measurable economic, environmental, and/or social outcomes.

Impact Canada promotes the use of a range of innovative funding approaches, including:

- Challenges – Issuing prizes for whoever can first or most effectively find a solution to a defined problem, and/or making use of structured, open competitions to solicit proposals to fund the best ideas with the potential to solve thematic problems.

- Pay-for-Results – Using customized instruments to-shift the focus towards issuing payments based on funding recipients achieving positive and measurable societal outcomes (e.g., social impact bonds, pay-for-success mechanisms).

- It also encourages the application of behavioural insights and other evidence-based approaches to improve program and service delivery.

Impact Canada is supported by a Centre of Expertise housed within the Impact and Innovation Unit. The team has extensive experience in executing novel and innovative programming in government in the following areas:

- **Innovative Funding and Partnership Approaches** – staff have been directly responsible for implementing innovative funding approaches in government programs including social impact bonds, impact investing, and launching large-scale challenges to crowd-source solutions to pressing problems. These experiences all involved implementing multi-sectoral approaches that have drawn together government, private sector, philanthropic and non-profit sectors to achieve shared outcomes.

- **Impact Measurement** – the team has worked with partners to co-design and co-develop evidence-based approaches to ensure improved program outcomes. The staff leverages a range of evaluation and impact measurement approaches and works with partners to advance new impact and outcome measurement methodologies.

- **Behavioural Insights** – the team has extensive experience supporting the execution and delivery of experiments and projects incorporating behavioural insights methodologies. The team's work includes the application of evidence-based principles and approaches, running of small to large scale experiments, and the strategic application of behavioural science to policy development in direct support of the Government of Canada's core mandate and commitments.

WHY MEASURE IMPACT?

WHY MEASURE IMPACT?

At its core, impact measurement is an impartial means of informing decisions through robust, scientific methodology. The main reason that we measure a program's impact is to allow us to determine whether a program achieved its desired outcomes or not. In addition, when well designed, impact evaluations can help to inform a number of related questions:

- What kinds of programs are best suited to deliver the most value to society?
- What kinds of programs should we discontinue or avoid?
- How are the benefits of a program distributed across different groups of society?
- Are there parts of a program that are of more value than others?
- How does a program achieve its impacts?
- How might we improve a program to achieve a greater impact?
- What kinds of programs should we scale up or expand?

In the broader landscape of public sector innovation, impact measurement is only becoming more and more relevant:

- Measuring impact is about being accountable, which is a basic principle of good government. We should be rigorous in our efforts to demonstrate how public spending is linked to improved outcomes for Canadians. Our growing emphasis on results and delivery within the public service is clear evidence of this.

- Deputy Heads are now required to devote a percentage of program spending to experimentation. Impact measurement is a key means by which we can meet this commitment.

- Pay-for-success models like social impact bonds are an emerging means by which departments are looking to improve results for Canadians. These approaches are central features of Impact Canada. If we want to pay for success, we need to clearly understand when we have achieved it and when we have not. That is the central aim of impact measurement. A rigorous method for evaluating the impact of a given initiative is critical to provide both initial investors as well as those who are paying for success confidence that the program's results are being measured accurately and fairly.

- There is a growing demand for economic forms of evaluation, like (social) return-on-investment, or cost-benefit analysis. Impact measurement is a core and essential step in these approaches. In order to know what our return on investment is, we first need a rigorous understanding of what our impact has been.

- Decision-makers are increasingly looking to innovate because they want to discover 'what works' when addressing long-standing or difficult social problems. Overtime, the accumulation of multiple impact evaluations in specific areas can support this.

- Overall, we are working in a context which is increasingly looking to innovation as a means to improve results for Canadians. For this reason, measuring our impact needs to become an embedded part of how we work.

IMPACT MEASUREMENT DEFINED

Because the delivery of results for Canadians is at the heart of Impact Canada, understanding the extent to which programs make a difference in the lives of Canadians is of central importance. Having an impact has a specific meaning. An ***impact*** is any change in outcome that was ***caused*** by a program or policy investment.

In practice, the terms *'outcome'* and *'impact'* are often used interchangeably. While these are closely related concepts, their meanings are distinct. An outcome is any social, environmental or economic benefit that a policy or program is interested in maintaining or improving in some way. An example might be labour force participation. Outcomes measurement answers largely descriptive questions, like *what is the current rate of labour force participation? How does labour force participation vary by region? Has it varied over time?*

Impact is defined as a change in social, environmental or economic outcomes (positive or negative) that are directly attributable to an in-

An *impact*, by contrast, is the extent to which a program causes a change in an outcome. Impact is defined as a change in social, environmental or economic outcomes (positive or negative) that are ***directly attributable to an investment***. To build upon the labour force participation example given above, a question of impact might be: *to what extent did a job training program **change** the labour force participation rates of trainees?* An easy rule of thumb here is to think of ***outcome*** as a noun, and ***impact*** as a verb.

The distinction between outcome and impact is important, because very often, we need to know more than descriptive information about the outcomes we are interested in. Often we want to uncover the extent to which our actions (policies, programs or interventions) impact (or change) outcomes. This is the role of impact measurement. With this in mind, impact measurement, therefore, goes a step beyond the measurement of outcomes, and attributes the extent to which a policy or program creates a change in an outcome or outcomes of interest. Another way of saying this is, it is a way of isolating the extent to which a program changes outcomes from the effects of anything else which might also change that outcome.

Using this definition, the following sections will provide a brief overview of some key concepts related to impact, before moving into a description of the main set of impact measurement methods that we should be aware of.

THE PROBLEM OF IMPACT MEASUREMENT

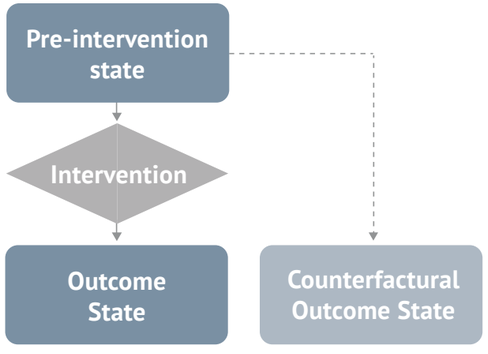
When we thinking about measuring the impact of a program, very quickly it becomes obvious that there is a problem we need to overcome: it is not always obvious that the results we observe are actually the product of the interventions we are interested in.

"it is not always obvious that the results we observe are actually the product of the interventions we are interested in."

Imagine a group of individuals are placed in a program to improve their levels of physical activity. After three months, we observe they are (on average) more active than they were before the program. Was the program successful? In this scenario, it is impossible to tell whether the program itself was responsible for this change in outcomes, or whether that change was wholly or partially due to other causes.

- Suppose the program began in April and ended in July. It is plausible that the participants simply became more active with improving weather.
- Suppose another organization offered the same group of people a different program. It is also plausible that that program had the desired impact, and perhaps ours had none at all – or even a negative impact.
- Suppose a major study on the benefits of healthy living was promoted in the media during that time. We cannot necessarily disentangle any program effect from the effect of this new information our participants might have received.

Figure 1



This illustrates very well why the earlier distinction between outcomes and impact is critical. Outcomes very often change, and often for reasons we may not know about. we want to understand a specific program's impact, simply measuring changing outcomes is not enough. That strategy will not allow us to attribute that change to any single cause like a program. We need to find a way to compare what did occur, with what would have occurred. Those are what we call the ***factual*** and ***counterfactual*** scenarios. The factual scenario is what actually occurred (participation in the program). The counterfactual scenario is what *would have* occurred in the absence of the program. By definition, the counterfactual scenario cannot ever be observed, because it is defined as what did not happen. So, the challenge of impact measurement is to find some way to reconstruct what *would have occurred* in the absence of a program, so we can compare those two scenarios, and determine our true impact.

*"The factual scenario is what actually occurred (participation in the program). The counterfactual scenario is what **would have** occurred in the absence of the program."*

Measuring *impact* is the way that we can isolate the effects of a program on an outcome (or outcomes) of interest. This is akin to asking: how much of the observed change in outcomes is attributable to the program itself? As we will see later, the best approach we can take is to compare two groups, alike in every way, except that one participates in a program, while the other does not. We refer to the former as the ***treatment*** or ***test group***, and the latter as the ***control*** or ***comparison group***. Because the only thing that differs between these two groups is their participation in a program, any differences in outcomes that we observe between them can be attributed to that intervention, and not anything else. This is what provides us evidence of impact. Later sections of this document will provide an intr oduction to the main set of methods used to do this in order to determine a program or policy's true impact.

THE CETERIS PARIBUS CONDITION

Ceterus Paribus translates from Latin as 'other things equal'.

Understanding the counterfactual outcome state requires comparing the outcomes of a treatment group with those of a control group. In doing so, we must ensure that the control group satisfies what is known as the ***ceteris paribus*** condition. *Ceterus paribus* translates from Latin as 'other things equal'. By randomly assigning individuals to two groups, or creating matched groups where we ensure participants are matched on key variables, we minimize the likelihood of systematic differences between the two groups, and we can assume that anything that may affect the outcome has the same influence on both groups. It follows then that when one of these groups does not participate in the intervention, their outcome state represents what the outcome state of the treatment group would have been in the absence of the intervention. Furthermore, the *ceteris paribus* condition is meant to ensure that the two groups would react to an intervention in the same way.

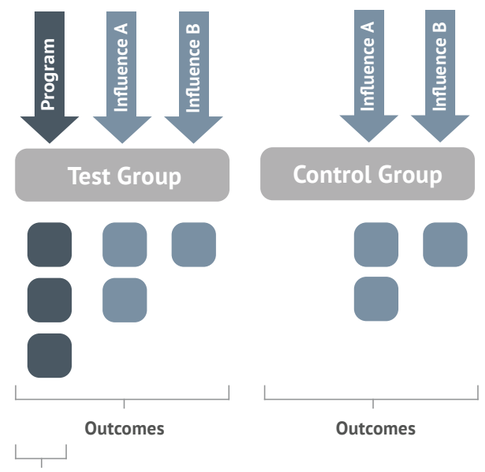
Gertler et al (2016, p.52) provide three helpful criteria to consider when assessing whether *ceteris* is in fact *paribus* between a treatment and control group:

1. There should be balance between the two groups in terms of average characteristics, both observable (e.g., age, gender) as well as unobservable (e.g., motivation, ability);

2. The treatment (program or policy) cannot directly or indirectly affect the comparison group in any way (for example units in the comparison group cannot receive the treatment); and;

3. The outcomes of the comparison group should change in the same way as those of the treatment group, if the comparison group were exposed to the program.

Figure 2



When these criteria are met, we can be satisfied that *ceteris* is *paribus* between the two groups, which is to say that they are alike in every way before the treatment, except that one group participates in the program, while the other does not. This is known as a ***valid*** control group. It allows us to attribute the difference in outcomes between the two groups to the program itself, because there is nothing else that is different between them. This ensures a high level of internal validity in the impact evaluation and its results.

INTERNAL AND EXTERNAL VALIDITY

There are two key aspects of impact measurement that must be considered in any evaluation:

i. **Internal validity** refers to the extent to which an evaluation successfully captures a causal effect of an intervention on the outcome(s) of interest.

ii. **External validity** refers to the extent to which an estimated impact from one study can be generalised to, for example, other regions, population groups and so on.

It is key to understand that internal and external validity are independent concepts and it is possible to have a study that has high internal validity but low external validity and vice-versa. The focus of this guide is on the establishment of internal validity. A later section is dedicated to issues related to external validity.

"Internal validity refers to the extent to which an evaluation successfully captures a causal effect of an intervention on the outcome(s) of interest."

THE MAIN THREATS TO THE INTERNAL VALIDITY OF AN IMPACT EVALUATION

When the *ceteris paribus* condition is met, we can compare the outcomes of our treatment and comparison groups, and retrieve what is known as an ***unbiased treatment effect*** estimate. This is the same thing as saying the results have high internal validity. Impact evaluations which do not use valid comparison groups are vulnerable to a range of biases or threats to internal validity (Campbell, 1957). The most common and most important causes of bias are:

- Selection bias,
- Reverse causality,
- Systematic measurement error,
- Time effects,
- Non-compliance,
- Placebo effect,
- Behavioural effects, and
- Observer bias.

The aim of impact measurement is to select a design which eliminates, or at least minimizes as much as possible these potential sources of bias. When the sources causing bias are not addressed properly, the internal validity of an impact evaluation can be compromised, and the result needs to be interpreted with caution.

SELECTION BIAS

Selection bias occurs when there are systematic differences in characteristics (either observable or unobservable) that exist between the treatment and control groups as a result of the process of assigning individuals to either group. For example, in a job training program this can happen if more motivated people put themselves forward for the program, or if program administrators select more able participants. In this case the treatment group would be fundamentally different from the control group before the start of the

[1] For a more detailed explanation of the theory of random assignment and why it is important, see (Shadish, Cook, & Campbell, 2002, pp. 248-251)

job training program – specifically, the treatment group would be more motivated and/or of greater ability than the control group due to the selection process. This often means that over time, the treatment group will naturally have better job outcomes than the control group regardless of the effectiveness of job training. It is likely that in this scenario comparing the outcomes for the treatment and control groups would lead to an overestimate of the impact of the job training program. Where selection bias is an issue the control group is not valid, and the treatment effect estimate (any difference in outcomes) is assumed to be biased.

The reason why we have to pay attention to the characteristics of the treatment and control group is that some of these characteristics are very likely to have an impact on outcomes of interest independent of the program being evaluated. Selection bias is unfortunately very common in evaluation research. Programs that allow self-selection, where participants opt-in of their own volition, potentially all suffer from selection bias, because those who choose to participate are necessarily different from those who opt not to. Although it may happen by chance in some situations, in practice we can never assume that those reasons are irrelevant to the outcome.

Selection bias can result in either an overestimate of impact, or an underestimate of impact:

- It is common that those who are most likely to be successful are the ones who opt-in or whom are selected by program administrators. In these cases, it is generally assumed that simple comparison of the observed outcomes between the treatment and control groups will lead to an ***overestimate*** of impact.
- In other cases, program administrators may purposefully choose people with lower chances of success (e.g. lower levels of ability or motivation) then a simple comparison of the observed outcomes between the treatment and control groups will lead to an ***underestimate*** of impact.

The problem with trying to address selection bias is that while many characteristics/factors are observable (e.g. gender, age, employment status, income, education, health status), many are not. In practice, observable characteristics are easy to deal with in impact evaluation. We can compare these kinds of attributes between our two groups and if differences in characteristics exist, then we know that the two groups are not comparable.

"Selection bias occurs when there are systematic differences in characteristics (either observable or unobservable) that exist between the treatment and control groups as a result of the process of assigning individuals to either group."

It is the unobservable characteristics that are more challenging to deal with, and they can take two forms. First, they can be the kinds of things that are difficult or impossible to observe at all. Things like personality traits, motivation, ability/intellect, preferences are generally unobservable, or at least they are very difficult to measure accurately. The second kind of unobservable characteristics are those that might in fact be observable but are (for whatever reason) not collected in the available data. For example, some things which are considered sensitive might not be captured, like income or sexual orientation.

TIME EFFECTS

Time effects (also known as history effects) are a major issue that can compromise the validity of simple before-after comparisons of outcomes. They represent part of what would have happened in the absence of an intervention, and so are not the impact of the interventions we are interested in measuring the impact of. Whenever we measure an outcome before an intervention takes place and then the same outcome after the intervention, we need to be aware that many things can change over time in the broader environment, and that many of these things can affect the outcome independently of the program. Our earlier example of a physical activity program taking place over the summer illustrates the problem time effects can pose. It is likely that physical activity levels ebb and and flow naturally over time with the changing weather. We need to be careful to not attribute these natural changes in outcomes to programs we are evaluating, because we can easily under- or over-estimate the true effect of the program in doing so.

In these situations, the availability of a control group allows for designs that can isolate time effects from the effects of the intervention, using techniques described later in this document. As discussed the control group must be similar to the treatment group, in which case the impact of time effects would be the same across the treatment and control group and hence would be controlled for in the study.

Another related source of bias is what is known as '***regression to the mean***'. Regression to the mean is a phenomenon where outcomes that are 'extreme' (unusually high or unusually low) tend to naturally trend toward the average value of that outcome over time. An example would be a tutoring program for very poorly performing students. Their outcomes (in this case, grades) will tend to improve naturally, simply because there is not much room for them to worsen (they can really only improve anyway). These instances might lead us to overestimate the impact of our tutoring program, because it is likely those grades would rise naturally (to some extent) without the program.

REVERSE CAUSALITY

Linked to the selection issue above, estimates can and will be severely biased if the target population of the intervention has been chosen (or selects itself in) based on the pre-intervention outcome level. For example, if a special forces police unit is deployed in the most crime-ridden regions, a simple comparison of crime levels across the treatment and control regions will reveal that the intervention is associated with higher rather than lower levels of crime. A greater police presence may generate higher rates of reported crime, so it may appear that the intervention (police presence) is exacerbating the problem (crime) even as it is having the opposite effect. Similarly, if a health intervention such as an exercise program tends to attract people who are sportier and healthier to begin with, we would derive biased estimates of impact.

In these cases, it is the outcome level that determines the intervention and not the other way around. It can be helpful to think of reverse causality as a 'chicken or egg' scenario. This should not be mistaken for a causal effect. In such situations, longitudinal data which observe outcomes over time enables some techniques that better approximate causality.

SYSTEMATIC MEASUREMENT ERROR

Measurement error refers to the differences that exist between the true and recorded values of a measure. These differences can be either random or non-random/systematic. For example, individuals often tend to over-report their income, meaning the recorded value of their income is greater than the amount they actually earn in reality.

"Measurement error refers to the differences that exist between the true and recorded values of a measure. These differences can be either random or non-random/systematic"

Measurement error that is *random* is not problematic, because errors randomly occur for both test and comparison subjects, and these anomalies will in theory balance out between the two groups. It is *systematic* measurement error that is problematic.

NON-COMPLIANCE

Participants can fail to comply with their group assignment status. This includes those in the treatment group who refuse to participate or drop out before completing it (this is known as attrition), as well as participants in the control group that find a way to benefit from the treatment. The likelihood of non-compliance depends heavily on the nature of the intervention and how easy it is to monitor the participants' behaviour. An intervention that entails watching a 20-minute health and safety video in a lab setting is likely to have high compliance. For an alternative intervention where the video is sent to the participants via email, compliance might be lower. If the intervention is a 6-month training course, attrition might become a problem.

Non-compliance does not bias the treatment effect estimate if it is random. Unfortunately, non-compliance is rarely random, and evidence from many areas including job training, substance abuse training and psychotherapy supports this (Shadish, Cook, & Campbell, 2002, pp. 323-324). For example, less motivated people or people with health problems are more likely to drop out of a job training course. Non-random, non-compliance like this is a problem because it systematically alters the composition of the treatment or control group, meaning that they are not similar (comparable) anymore. In this case, the resulting impact estimate is likely to overstate the true effect, because those that remain in the treatment group are, on average, more motivated and healthier than those in the control group due to the attrition.

"Unfortunately, non-compliance is rarely random, and evidence from many areas including job training, substance abuse training and psychotherapy supports this (Shadish, Cook, & Campbell, 2002, pp. 323-324)."

Ensuring perfect compliance is usually not possible, either for ethical reasons (you often cannot impose an intervention by force on somebody who does not consent) or for practical reasons (it would require constant monitoring of many subjects for an extended period of time), or reasons which cannot be avoided like the death of participants. One practical workaround is to use an encouragement design, which is described in a later section.

PLACEBO EFFECT

The placebo effect arises from the *experience* of receiving the treatment rather than from the treatment itself. This is common in the health field where it is well-known that patients can react to sugar pills when they believe that it might be the actual medication. This is normally controlled for by a technique referred to as a 'blinded experiment' wherein the participants cannot tell whether they have been assigned to the treatment or control group; usually because both receive some sort of treatment experience (e.g. the actual pill or an identical looking sugar pill).

"The placebo effect arises from the experience of receiving the treatment rather than from the treatment itself."

The placebo effect is often difficult to deal with in social programs, as it requires mock or placebo treatments to be administered. In the job training program example, in effect this would entail entering the control group into a carefully designed placebo program. One way to get around this is to use a multi-arm trial, in which both the treatment and control groups receive treatment but the treatment group gets some additional support. However, note here that the estimated impact refers to the impact of the program options rather than the program itself. In circumstances where a program is being modified in some way, another approach to dealing with this is to use the original program as the control for the modified program, thus enabling a comparison of the new approach with the status quo.

It is also worth noting that in some policy settings, placebo effects are often part of the desired effect and do not need to be controlled for. For example, a training course may increase recipients' performance either through the acquisition of new skills from the content of the course, or through an increase in motivation from the mere fact of participating in the course, but both of these would represent a desired outcome.

BEHAVIOURAL EFFECTS

Related to the placebo effect is the issue of how participants can change behaviour as part of an impact evaluation. These arise from the fact that people know when they are in the treatment or control group when no placebo is used. Two key behavioural effects are:

- **Hawthorne effect:** participants of the program change their behaviour because they are being observed. The name originated from a study done at Hawthorne Works, an electric factory in Illinois, where increases in workers' productivity as a result of improved lighting and maintaining clean workstations were later found out to be actually due to their awareness of being research subjects and the increased attention to their work resulting from that. In summary, outcomes can improve for the treatment group even if the program has no effect and hence the Hawthorne effect leads to a biased overstatement of impact.

- **John Henry effect:** the comparison group (non-participants) change their behaviour when they know that they are not receiving the treatment. Out of spite, comparison participants may be motivated to do better than usual to improve their outcomes to show program administrators that they can do just as well without the program. Here, outcomes can improve for the control group and the John Henry effect leads to a biased understatement of impact.

As with the placebo effect, if possible, conducting blinded experiments, where subjects do not know which group they belong to and face equal conditions (both groups are being monitored with equal level of attention), is recommended to avoid behavioural effects like the Hawthorne and John Henry effects.

Best-practice in the medical sciences is a double-blind protocol. This is where both the treatment and control groups and the program administrators (people that give out the pills) are unaware of which pills (real or placebo) they are taking or giving out. Blinded experiments in the social sciences are a usually a lot more difficult to achieve – for example, even if we are able to roll out a mock or placebo job training intervention for the control group the program administrators will know which program they are delivering and may give away hints to the participants in the comparison group. In reality, these kinds of placebos are resource intensive and generally not practical.

"When using the Impact Canada Evidence Scale it is important to keep in mind that rigour is in the implementation, not the name"

OBSERVER BIAS

The evaluator (observer) might consciously or subconsciously expect different outcomes from the treated and untreated subjects. This may result in either biased assessment by the evaluator when recording measurements (especially if the assessment is subjective, the assessor may instinctively judge that the treated group performed better because he/she knows that they were treated), or he/she may give different signals/hints to subjects of the two groups, which may influence their behaviour. Where observer bias can be an issue, the typical solution is to conduct a double-blind experiment, where neither the participants nor the people collecting the data know whether a participant is in the treated or comparison group. Allocation to the two groups is decided by a third party, which is uninformed or unfamiliar with the characteristics of the participants.

THE IMPACT CANADA EVIDENCE MATRIX

THE IMPACT CANADA EVIDENCE MATRIX

Hierarchies of evidence (or evidence pyramids) have grown in popularity over the past few decades. These rank impact measurement methods according to their internal validity, and generally rank RCTs at the top, cascading down through quasi-experimental methods, with the most basic (before-and-after) designs at the bottom. This approach has a lot of merit, however it can create a misperception that RCT methods are a universal 'gold standard'. This can discourage the application of other rigorous methods that might be more appropriate in a given circumstance. In addition, RCTs may not always be administratively feasible. Grants and contributions programs, for example, sometimes do not fund RCTs as this is seen to overlap with research funding.

While internal validity is clearly the primary criterion of concern in the choice of an impact measurement method, there are other relevant concerns as well:

- **Ethics:** there are often circumstances where it is (or is perceived) to be unethical to randomize assignment, or withhold services from a group for the purposes of creating a comparison group. Many service providers operate with limited resources, and deliberately target those with the greatest need. On the flip side of this, it can be unethical to randomly assign certain things. Smoking causes cancer, for example, but there are no experimental studies of this because it is unethical to randomly assign tobacco.

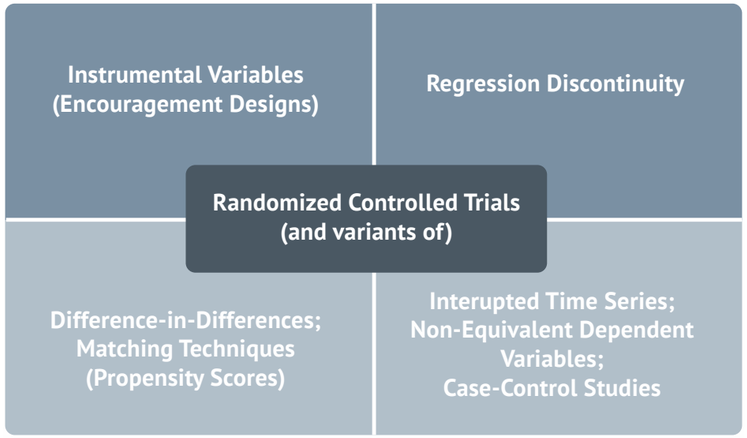
- **Feasibility:** there exist many things which we may be interested in understanding the impact that they have on outcomes, but which cannot be randomly assigned. Gender or personality traits are good examples here. We can't *manipulate* these things, and therefore we cannot randomly assign them.

- **Non-excludability:** some kinds of programs or interventions roll out in a way that precludes the ability to create a comparison group of any kind, because everyone is exposed to the intervention simultaneously. Legislation, regulation, taxation are each good examples of this. We cannot study the impacts of the legalization of cannabis on Canadians with a randomized controlled trial, because the law applies universally to all Canadians – there is no comparison group.

So, it does not necessarily make sense to imply that there exists any 'gold standard' that applies in all instances. In those circumstances where establishing control groups is feasible, randomized approaches will usually be most effective as these methods do have the highest level of internal validity.

For those questions of impact that simply cannot be answered with randomization, and so non-randomized methods (quasi-experiments), though they have lower internal validity, are often our best choice. For Impact Canada, our approach is to strive to use the most rigorous (internally valid) method we can, given the operational realities of the program we are measuring the impact of. This includes the use of both experimental and quasi-experimental techniques as appropriate.

Figure 3



Rather than a traditional pyramid or hierarchy, the IIU employs an ***evidence matrix***, presenting a menu of impact measurement options, and illustrating three key principles:

- The lack of an obvious hierarchy indicates that no single method is universally preferred in all circumstances.
- That being the case, randomized methods, including RCTs, stepped wedge/waitlist trials, and multifactorial designs, occupy a unique place in the menu of options as the category of methods with the highest internal validity.
- The strength of a method's ranking in terms of its internal validity is generally indicated by colour, with darker shades indicating higher internal validity.

"A key assumption with any ranking of impact measurement methods, is that rigour is in its implementation, not its name."

A key assumption with any ranking of impact measurement methods, is that rigour is in its implementation, not its name. A well-executed matching scheme may very well out-perform a poorly executed regression discontinuity, so it is important to be vigilant about quality. As these methods can be complex to design and implement, the IIU encourages, where practical, the support of experts who can provide guidance and advice to ensure that impact evaluation approaches and the conclusions derived from them are accurate.

CHOOSING THE RIGHT DESIGN

Based on the theory of causality and the definition of impact measurement outlined in this document, we have classified the methods described below into the ***Impact Canada Evidence Scale***. It reflects an ordering of the strength of the methods with respect to their internal validity (i.e. deriving robust estimates of impact). As discussed above, it should be noted that this ranking does not take into consideration issues related to external validity (i.e. the generalizability of the study's results). This is discussed at the end of this guide.

In order to make the ranking easier to visualise and comprehend, we categorize it into six different validity levels,. For each level in the scale a list of the methods that fall into the respective category is provided. Note that there are numerous variants of some of the methods (e.g. it is possible to combine matching with DiD), but they will fall into the same level ranking as the method to which they are related.

Following best practice in this area (which aligns with the generally accepted evidence threshold in many governments and public sector organisations in OECD countries), we set the minimum threshold at Level 2. Use of methods at Level 0 and Level 1 is not recommended for impact measurement as they do not meet a basic standard for impact measurement, which is high internal validity. In general, attempts should be made to attain as high a level as possible, but the resources used in the evaluation must be proportionate to the size of the program at hand.
When using the Impact Canada Evidence Scale it is important to keep in mind that rigour is in the implementation, not the name. The robustness of a method only applies on the condition that certain basic assumptions are met. Some of the more complex approaches, such as instrumental variables in particular, can be very volatile, and an estimate using an invalid or weak instrument can generate significant bias and be considerably less reliable than even a Level 1 study design.

"When using the Impact Canada Evidence Scale it is important to keep in mind that rigour is in the implementation, not the name"

A final critical point to note here is that the level of the study is in no way correlated to the level of effort or resource required to conduct the study and so budget should not be a consideration when determining the study to conduct. For example conducting a good difference-in-differences (DiD) or case control study can be more time consuming and expensive than conducting an experiement, which require planning upfront but less statistical analysis of the data than DiD and case control studies.

The sections that follow introduce the key design features of each impact measurement method, indicate the kinds of contexts in which they are likely to be most useful, and provide some key considerations to watch out for when appraising the quality of impact evaluations using these methods. These are followed by a brief section addressing the unique contributions of qualitative methods to impact measurement.

IMPACT CANADA EVIDENCE SCALE

LEVEL	DESCRIPTION	METHOD
1	The intervention is randomly assigned or 'as good as randomly assigned' through natural processes such that the treatment and control groups are on average identical at baseline, in terms of both the observed and unobserved variables. This means that the groups are identical (in expectation) and both will follow the same trajectory in outcomes in absence of the intervention. Therefore, any differences in the observed outcomes after the intervention are due solely to the intervention. These study designs produce unbiased estimates of causality with the highest levels of internal validity.	Randomized Controlled Trials (RCTs)/Experiments
2	The intervention is randomly assigned or independent variation in the intervention can be isolated after applying the respective estimation methods. However, there might be some confounding factors related to administering the intervention (such as placebo or observer effects) that have the potential to bias the resulting treatment effect estimates. Convincing evidence was presented to support the argument that the identifying assumptions of the estimation method hold.	Phased Introduction Instrumental Variable (IV) Estimation Regression Discontinuity Design
3	The intervention is not randomly assigned. The estimation method is able to control for selection bias arising from observable confounding factors and at least some of the bias arising from unobservable confounding factors. Reasonable evidence was presented to support the argument that the identifying assumptions of the estimation method are likely to hold. We cannot be sure of inferring a causal effect. However, level 2 and 3 methods have been shown to perform well (Concato, Shah, & Howritz, 2000) and represent the dominant form of evaluation in policy analysis.	Difference-in-Differences Interrupted treatment and NEDV
4	The intervention is not randomly assigned. The estimation method is able to control for selection bias arising from observable confounding factors, but not the bias arising from unobservable confounding factors. There is a high chance that the identifying assumptions of the estimation method do not hold. We therefore cannot be sure of inferring a causal effect. However, level 2 and 3 methods have been shown to perform well and represent the dominant form of evaluation in policy analysis.	Matching Case control studies
5	The intervention is not randomly assigned, but data on both treated and untreated units is available. The estimation methods do not account for selection bias. At best these methods are able to show a correlation between the intervention and the outcome, but correlation is not causation, because this could be due to a whole host of other reasons: the outcome may be causing participation in the intervention (reverse causality); outcomes may be driven by the particular characteristics and situation of the treated group (selection bias and regression to the mean); the outcomes could be caused by a number of other events (history effects). These methods provide minimal levels of internal validity and will usually be extremely biased. The bias will usually be upward, which means that social impact assessments using these study designs will systematically overstate the social impact of the intervention. Therefore, these methods should not be used in measuring social impact and instead should only be used for building hypotheses and for providing supporting evidence for higher-ranking studies over the Level 2 validity threshold.	Difference of means Before-after estimation Benchmarking with aggregate data
6	There is no comparison group (either a control group or the treatment group before receiving treatment) to compare the treated observations against. Alternatively, the conclusions of the study are not based on individual-level data, but on theoretical arguments or subjective opinions which are not supported by evidence and do not acknowledge the potential threats to validity. These methods should not be used in measuring social impact on their own and instead should only be used for building hypotheses and for providing supporting evidence for higher-ranking studies.	Statements of causal claims without support from data or other studies

CAVEATS AND DISCLAIMERS

1. The ICI Evidence Scale ranking is only based on the internal validity of the methods and does not account for generalizability of results (external validity).

2. The scale applies to single evaluations. Some evidence scales in the medical sciences use higher levels of rigour based on whether the results have been replicated elsewhere in follow-up studies. Replication serves to add further confidence to a study's findings.

3. The context of the study also matters. The context may mean that certain assumptions do not hold or that the budget, resource and initial conditions may not be suitable for a given methodology. If the assumptions are not upheld or the methodology cannot be applied properly, then the highest-ranking methods will effectively be no better than any other method in the scale. Since the context of the study matters, evidence scales like the one above should be seen as a guide to the merits of the different methods, rather than as a hard and fast rule.

METHODS FOR ESTIMATING IMPACT

METHODS FOR ESTIMATING IMPACT

The following sections provide an overview of the most popular impact measurement designs.They describe the methods in detail, structured in decreasing order of robustness according to their capacities to identify an adequate counterfactual. The ordering of different methods within the same subsection is arbitrary. A summary overview of these methods, along with their advantages and disadvantages, is provided in an appendix for ease of reference. A subsequent section discusses the important complementary role that qualitative methods play in impact measurement.

THE RANDOMIZED CONTROLLED TRIAL AND EXPERIMENTAL APPROACH

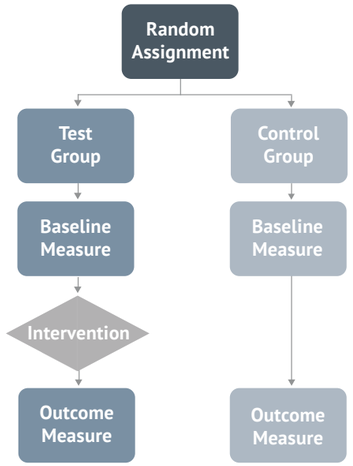
The ***experiment*** (a term often used interchangeably with ***randomized controlled trial***) is the best way of meeting the *ceterus paribus* condition. When well executed, experiments are the methods with the highest internal validity, meaning they are very well suited as impact measurement methods where they can be implemented.

The key feature of experiments is random assignment to test and control groups. Random assignment to either test or control conditions means that each individual is assigned *only* according to chance. This does two important things:

1. It ensures that the two groups are balanced (in expectation) in both observable and unobservable ways. Both groups will be as similar as we can achieve. Because the two groups are not inherently different from one another, it means there is likely nothing about either group that might influence the outcomes differently.

2. It also ensures that any threats to validity are also distributed randomly across the two groups. This means that any other factors that might change outcomes are also balanced between both groups, and these effects 'cancel out'. Any observed changes in outcomes can therefore be attributed directly to the program, to the exclusion of anything else.[2]

Figure 4



This means that random assignment is a highly effective way to rule out alternative explanations for changing outcomes. With random assignment, any change in outcomes observed between the two groups cannot be attributed to differences between the two groups because there are no such differences, or to alternative causes because any that exist impact the outcomes of both groups equally.

[2] For a more detailed explanation of the theory of random assignment and why it is important, see (Shadish, Cook, & Campbell, 2002, pp. 248-251)

The other key design features of experiments are the 'controls'. We can think of control features as preserving the benefits of random assignment. Controlling a study means preserving the composition of the test and control groups such that they do not change over time. This usually means:

- Ensuring there is no ***attrition*** in either group. We know that when attrition (drop-out) occurs, it is usually not random, so it is very possible to start out with randomly assigned groups, but end with groups which are no longer comparable due to attrition.

- Ensuring that there is no ***contamination***, meaning that all of those in the test group actually get exposed to the program/treatment/intervention, and none of those in the control group are exposed.

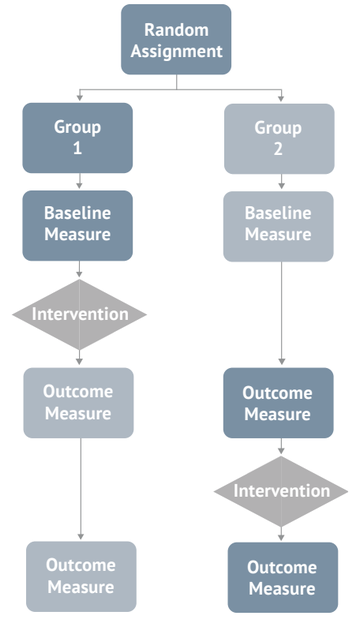
- Guarding against ***behavioural effects*** (described above). Individuals in controlled studies can change their behaviour if they have knowledge of their treatment status. Similarly, those running the program, or those conducting the impact evaluation might behave differently toward those in test versus control groups. These behavioural 'shifts' can often impact the outcomes that we are interested in, meaning any changes in outcomes we might see can not wholly be attributed to the program itself when these occur. The best impact evaluations are blinded, meaning no one involved (participants, program administrators, evaluators etc…) has knowledge of which group is the test, and which is the control. In practice, this is difficult to achieve in impact evaluations (discussed above).

It is important to be aware that experiments are more than just randomized evaluations – other design features to preserve the random assignment over the life of the impact evaluation are necessary. We should be aware that an experiment with poor controls may not meet the ceterus paribus condition, and results of such impact evaluations should be interpreted with caution.

VARIATIONS OF RCTS

The traditional RCT approach is to randomly assign participants to the control and treatment groups at the same time and to measure outcomes post-treatment. However, this is not always possible and in these cases, there are ways to use randomized assignment more 'creatively' to achieve much of the same effect. Phased introduction, multi-arm RCTs, and encouragement designs are a few common approaches to using randomized assignment for evaluation in pragmatic ways. These kinds of variations of traditional RCT methods can be attractive because they can integrate very well into natural modes of operation of many programs and in many service-delivery contexts.

Figure 5



PHASED INTRODUCTION

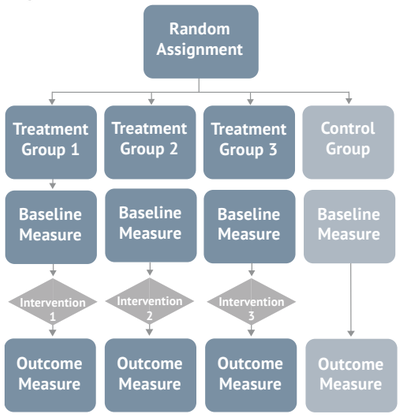
Also known as a ***stepped-wedge design***, ***waitlist*** or ***pipeline comparison***, this is an adaptation of the RCT, in which an intervention rolls out to participants sequentially in two or more phases, and the participants are randomly assigned to the phases. This approach can be applied to programs that are short-term in nature, with outcomes which are observable soon after the intervention ends. This method is often employed when there are ethical objections to withholding treatment from some, as in the end, all participants will have been offered treatment.

The main feature of this design is that the control group receives the intervention later than the treatment group as opposed to not receiving it at all. One can then identify the treatment effect either by comparing treatment and control groups prior to the control group having received the intervention, or simply as the effect of having been exposed to the treatment for longer (for interventions which last over time). Assignment to the earlier or later treatment phases must still be random, that is, independent of any relevant individual characteristics, in order for the study to be robust and derive estimates with high internal validity.

MULTI-ARM / FACTORIAL DESIGNS

Multi-armed RCTs simultaneously test multiple program options. In this design, participants are randomly assigned into one of several programs, enabling a comparison of results between the options being tested. Notably, it is possible to use a multi-arm RCT that does not include a control group, but rather compares the outcomes of various programs against one another. This can also be useful in instances where it is not possible to deny eligible participants treatments. Alternatively, where there is hesitance to establish a true control group (no program at all), it is possible to include an 'arm' in which participants benefit from some minimal-level program intervention which can be used in the analysis to approximate a 'no program' status.

Figure 6



WHEN CAN RANDOMIZED METHODS BE USED?

Methods in this category are characterised by the fact that the program administrator/evaluator can determine assignment of participants to the program/intervention upfront. Circumstances where this might be the case could include:

- Programs that have excess demand. Where there are more eligible potential participants than space in the program itself, the fairest or most ethical approach to allocating spaces is likely random selection.
- Where a program is delivered in phases or cohorts. This enables 'stepped wedge' designs, and participants can be randomly assigned to phases.
- Where a program is being 'tweaked' or variants of a program are being tested. Participants can be assigned to versions, enabling multi-arm/factorial designs.

THINGS TO CONSIDER

- The intervention must be randomly assigned.
- The participants must comply with the group they have been randomly assigned to (they must not take the intervention if they were assigned to the control group and vice versa).
- The evaluator must ensure that the impact estimate is not due to the simple fact of being observed or being aware of one's belonging to the intervention or control group.

QUASI-EXPERIMENTAL APPROACHES

It is often the case that programs cannot, or are unwilling to undertake a process of random assignment, making an experimental design impossible to execute. In these instances, there are ways to estimate the counterfactual using quasi-experiments. Quasi-experiments are defined as designs which are "characterized by non-random assignment" (Dunning, 2013, p. 19), but which attempt to replicate many of the benefits of an RCT. The key feature of quasi-experimental approaches is, therefore, that they cannot determine the assignment of the program/intervention in any way.

"Quasi-experiments are defined as designs which are "characterized by non-random assignment"

Quasi-experiments vary widely in approach, but they all have the common objective of attempting to replicate the rigour of the RCT as closely as possible in scenarios where random assignment is either not feasible or not desirable.

In this document they are divided into 'upper-tier' and 'lower-tier' quasi-experimental methods.

- The upper-tier methods are by their nature able to control for a wider range of factors, including unobserved characteristics. Under the right conditions, they can derive more robust estimates of cause and effect at levels close to the RCT.
- The lower-tier quasi-experimental methods can only control for observable characteristics and are therefore only valid under the assumption that only observable factors affect the outcome and treatment status.
- Nonetheless, as with any set of methods, there are advantages and disadvantages, and even the upper-tier methods have fundamental assumptions which can be violated. Both upper-tier and lower-tier methods can produce unbiased impact estimates in certain circumstances and strongly biased estimates in others.

UPPER-TIER QUASI-EXPERIMENTAL METHODS

The following quasi-experimental methods are classified as upper-tier and superior to other quasi-experimental methods because they can account and control for both observable and unobservable differences between treatment and comparison groups. This means that when well executed, they serve well as impact measurement methods.

INSTRUMENAL VARIABLES

The instrumental variables (IV) approach has its origins in the field of economics, where it has been in use for approximately fifty years. In recent times, it has become increasingly popular within the broader social sciences, and in impact measurement in particular.

IV takes advantage of an external random source of variation in a variable to identify its causal effect on an outcome variable. For example, it is hard to isolate the causal effect of income on happiness, because happier people may earn more money (reverse causality). However, lottery wins are random and represent a source of random variation in people's income and we can use this random variation in income (due to lottery wins) to estimate the causal effect of income on happiness.

In the example above, lottery wins are an example of an instrumental variable (IV), used to *instrument* for income. So what is an IV? In plain language, an IV can be understood as anything that influences the factor whose effect we are interested in measuring, but is itself free from biases such as selection, time effects and reverse causality. This usually happens when the IV occurs randomly (like the lottery wins above) or is randomly assigned by the evaluator (see Encouragement Design below).

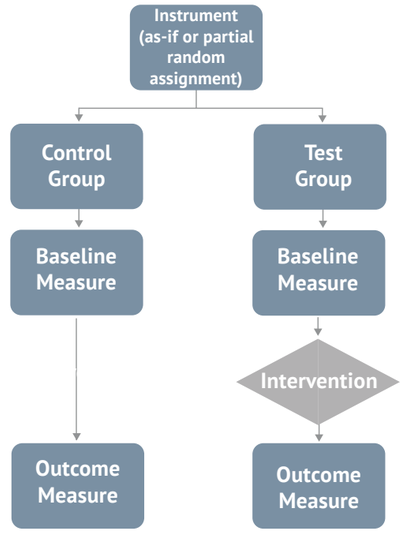
There are two fundamental assumptions required for an instrument to be considered valid. A good instrument must:

1. affect an individual's probability of participating in a program; but,
2. not have any direct impact on that participant's outcomes, except through its influence on that individual's participation.

The first assumption is known as ***first-stage relevance***, and an IV which violates it is said to be a *weak* IV. The second is known as the ***exclusion restriction***.

If these assumptions hold, as with an RCT, all other factors (both observable and unobservable) apart from the treatment itself are controlled for and so we can derive a treatment effect with high internal validity using IV. This is why IV sits in the upper-tier quasi-experimental method category. However, if these assumptions do not hold, estimates can be severely biased. IV should therefore not be used without a plausible argument in favour of these two assumptions. Statistical tests to check the validity of the assumptions should always be performed and combined with appropriate theoretical reasoning.

Figure 7



IV is common in what are known as 'natural experiments'. These are 'natural' in the sense that some form of randomization occurs naturally, and not through the efforts of an evaluator. As an example, a well-known study in India examined the impact of schooling on educational attainment by using proximity to bus stops as an IV. Because the locations of bus stops were determined randomly, and since access to public transport was a key determinant of school attendance, this made school attendance effectively random

"IV takes advantage of an external random source of variation in a variable to identify its causal effect on an outcome variable."

for a large number of children in India. In this case, close proximity to a bus stop is an instrument, which was highly correlated with school attendance, so children who lived in close proximity to bus stops could be compared with those who did not in order to understand the impact of schooling on educational attainment.

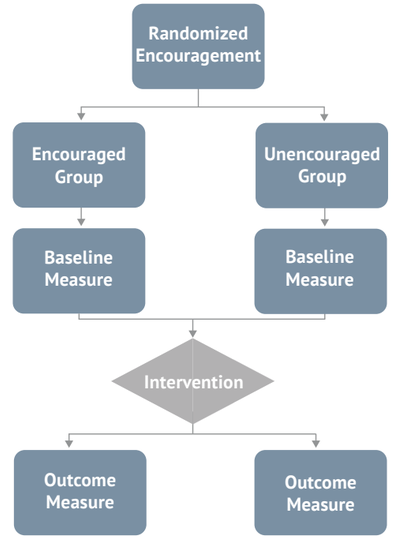
ENCOURAGEMENT DESIGN

Encouragement designs are a special case of IV that are particularly useful in impact measurement. In cases where the intervention/treatment itself cannot be randomly assigned (perhaps because delivery of the intervention is not under full control of the evaluator), or where the likelihood of people dropping in or out of the treatment is high, an alternative option is to randomly assign 'encouragement' to participate in the intervention[3,4]. In practice, with encouragement designs, all potential participants are made aware of the availability of a program, but a random selection of these are offered extra incentive to enroll. This can be in the form of phone calls or email reminders, vouchers or other financial/non-financial incentives for participating, which will make the encouraged group more likely to enroll in the program.

Even if assignment to the intervention is non-random, the existence of a random source of variation within it – the encouragement – allows for the use of an IV approach to extract an unbiased impact estimate. In the case where assignment to the intervention was random but later mired by attrition or non- compliance, an IV can be used as well where assignment itself is seen as encouragement to participate in the intervention.

One caution to note here is that the results of a randomised encouragement design are less generalizable than a traditional RCT because they are specific only to those people who change their behaviour as a result of the encouragement (known as the complier group). However, encouragement designs are often useful in impact measurement because they relate more closely to what is in the control of government agencies. Often, government departments/ agencies can only encourage people to do things rather than mandate them.

Figure 8



3 For a more detailed discussion of the theory of random assignment, see (Shadish, Cook, & Campbell, 2002, pp. 248-251)
4 This method is also known as a *downstream randomisation* design.

Suppose a policy maker is interested in understanding the impact of after-school tutoring on educational attainment. She could randomly assign the tutoring to treatment and control groups and measure outcomes. However, given the ethical issues that may arise in doing this, one solution could be to offer tutoring to all eligible students, but randomly assign encouragement to parents to enroll their children. This is similar to the technique used in a well-known study (Bogatz & Ball, 1971) that wanted to assess the impact of watching Sesame Street on educational outcomes for young children. Here one group of parents was encouraged to let their children watch Sesame Street and that encouragement was used as an 'instrument' to estimate the causal effect of Sesame Street on educational outcomes, relative to the group that was not encouraged.

WHEN CAN INSTRUMENTAL VARIABLES BE USED?

IV approaches can be used either:

- Where a natural source of variation exists, like a lottery, or 'as if' random exposure to a program; or,
- In the context of universally offered programs which are voluntary. This means no one is excluded from participation, and participants select in. This enables use of encouragement designs.

THINGS TO CONSIDER

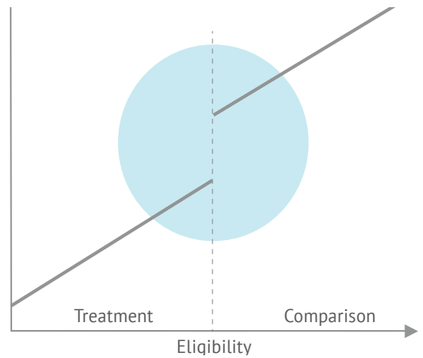
- The exclusion restriction must hold.
- The instrument must be relevant. A weak instrument would lead to a highly unstable estimate and magnify the bias from violation of the exclusion restriction.
- (In the case of encouragement design) The individuals who comply with the encouragement must not be markedly different from the rest of the population.

REGRESSION DISCONTINUITY DESIGN

Regression discontinuity design (RDD) is applicable for programs that determine eligibility of participants using some (quantifiable) threshold criterion. Examples might be household income level, age, or grade point averages, where individuals who score above or below a certain value are targeted for treatment. In reality, it is very common for programs and social services to determine eligibility in this manner, so the opportunities to use this method to understand impact are great. Some argue that RDD is a highly under-utilized method (Moscoe, Bor, & Barnighausen, 2015; Shadish, Cook, & Campbell, 2002, p. 208).

Take for example a program of additional support at school for poorly performing children where eligibility for the program is assessed by the students' most recent test scores in mathematics. Suppose the support program is offered to children who scored less than 40%. In reality, the group around that threshold (e.g. those that score between 37%-43%) are actually very similar. However, due to the somewhat arbitrary cut-off point those scoring 37%-39% receive the treatment while very similar students scoring between 40%-43% just miss out.

Figure 9



RDD therefore operates on the assumption that the 'just ineligible' and the 'just eligible' are effectively comparable as groups, because the slight differences in their respective eligibility status are likely to be due to random variation in their circumstances. In the hypothetical scenario above, the difference between scoring 39% and 40% or 41% on the day of a test is likely to be due to chance, such as how the student felt that day or how successfully they guessed some of the answers.

In effect, we have an RCT around the eligibility threshold. This means we can assume that all characteristics (both observable and unobservable) are balanced between the treatment and comparison groups, and therefore any differences in outcomes we observe can be attributed to the program/treatment itself. If the assumptions for RDD hold, then impact estimates will have high internal validity. This is why RDD is an upper-tier quasi-experimental method.

WHEN CAN REGRESSION DISCONTINUITY BE USED?

Regression Discontinuity Designs are useful for programs that:

- Determine eligibility based on some quantifiable criterion, like a grade, age, or a score of some sort; and
- That criterion is not correlated with anything else that might explain outcomes; and
- There are enough 'cases' (participants) clustered around the criterion to enable an analysis (usually for this reason, sample sizes for RDD are larger than for classic experiments).

THINGS TO CONSIDER

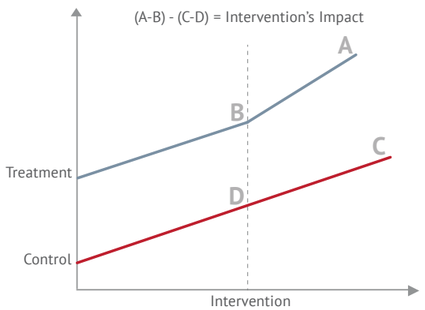
Matching requires large data sets to increase the likelihood of finding good matches for program participants.

- If there are other things that change abruptly at the cut-off, RDD will not be able to distinguish between the effect of those changes, and the effect of the program. For example, an RDD that is used to test the effects of a tax credit on the financial well-being for people aged 65 and older might be problematic if, for example, people also tend to retire around that age, or also become eligible for various other seniors benefits at that age.
- The choice of "bandwidth" around the cut-off point is important. This is to say, there need to be enough cases on either side of the chosen cut-off to allow for a meaningful statistical analysis. However, bandwidths that are too wide can mean that the two groups on either side are no longer comparable.
- RDD can give robust, internally valid estimates, but because they analyse only a smaller portion of cases around the cut-off, they are more limited in terms of external validity. In the mathematics support program example above, the impact estimate that is derived from RDD is likely less applicable to children in the class who are very low performing (students who scored less than 30% for example).

DIFFERENCES-IN-DIFFERENCES

The difference-in-differences (DiD) approach uses before-and-after data from a treatment and comparison group to compare the changes in the outcomes between them. Difference-in-differences relies on the ***parallel trends assumption***, which is the assumption that the trend in the outcome of the comparison group is a good representation of what the trend in the outcome of the treatment group would be in the absence of the intervention (Gertler, Martinez, Premand, Rawlings, & Vermeersch, 2016, p. 134).

Figure 10



If the parallel trends assumption holds, then DiD is able to estimate a causal effect with high internal validity controlling for all observable and unobservable characteristics. The validity of the parallel trends assumption can only really be tested by looking at historical trends in the outcome in the treatment and comparison group. If the outcome has been trending very similarly for a long period of time in the treatment and comparison groups before the program, this provides evidence in support of the parallel trends assumption. The more historical trends data we have, the more evidence we can obtain to support or refute this assumption. However, we cannot rule out the possibility that the parallel trends assumption does not hold even if historical trends are near-identical. For example, an unpredicted new policy may come into effect, suddenly altering the comparison group's trend in outcomes.

DiD takes the following steps:

1. Find a comparison group that has similar historical trends in the outcome variable.
2. Observe outcomes, before and after the program, for treatment and comparison groups.
3. Calculate the difference in outcomes over time within each group. This is the first difference.
4. Calculate the difference between the two differences calculated in step two. This is the second difference, and the estimate of the impact of the program.

WHEN CAN DIFFERENCES-IN-DIFFERENCES BE USED?

Difference-in-differences can be used in cases where:

- Random assignment is not feasible.
- A comparison group can be identified, and historic trend data confirms its outcomes have moved in tandem with the treatment group; or,
- It is reasonable to assume that the 'parallel trends assumption' holds.

LOWER-TIER QUASI-EXPERIMENTAL METHODS

MATCHING

The following set of methods require the analyst to assume that a valid counterfactual can be constructed based only on observable characteristics. By nature, these methods are not able to account for any unobservable differences that might exist between the treatment and comparison groups. For this reason, they are considered weaker than the methods described above.

Matching techniques are ways to construct a comparison group based on observable characteristics. Using the observable characteristics that are available in the data, the aim is to find at baseline (before the program starts) a 'match' for each participant among the non-participants, and in this way build a comparison group that looks as similar to the treatment group as possible.

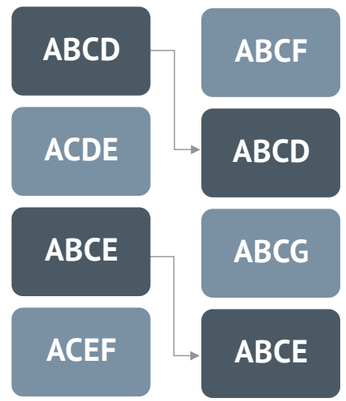
This 'match' must have the most similar values for a set of confounding variables, such as age, gender, income, education etc. Intuitively, a match is a 'copy' of a treated unit which resembles it in all relevant ways except for treatment status. If a well-matched comparison group can be created, then the difference in levels of outcomes between the treatment and comparison group can provide a good estimate of the effect of the program. The underlying theory here is that, because good matches have similar background characteristics at baseline, they would also have similar levels of potential outcomes, and therefore their outcomes can be used as the counterfactual case.

One key difficulty with matching is that the more variables we attempt to match on, the more difficult it becomes to find a close match in the comparison group. This issue is known as common support – that the treatment and comparison groups span the same range of values for all the relevant control variables. Otherwise, for some units it will be impossible to find a good match, because there will be no units from the opposite group which are similar enough. The image to the right illustrates this problem: if each letter represents a characteristic, it can be difficult to find an exact match on the comparison side, as the number of characteristics to match on increases.

In addition to this, the most important assumption for the validity of matching estimation is that differences between treatment and comparison groups are observable (i.e. that there should be no unobserved sources of selection into the program). This is a very difficult assumption to satisfy. In practice it can be difficult to know exactly what these variables are, and evaluators are often limited by what is known from previous research, as well as what is available in existing data.

On this basis, matching falls into the lower tier of quasi-experimental methods because it can only be done on the basis of observable characteristics. This is a strong assumption to make, and matching runs the risk of ignoring important unobservable differences between treatment and control groups which might be driving outcomes independently of the intervention or program.

Figure 11



PROPENSITY SCORE MATCHING

As mentioned previously, matching based on a greater number of baseline characteristics is desired, however, this can create what is known as the ***problem of dimensionality***. As the list of variables to match on grows, it becomes harder to find a suitable match for each unit. Matching can quickly become impractical for large datasets with many control variables.

"...matching based on a greater number of baseline characteristics is desired, however, this can create what is known as the problem of dimensionality. As the list of variables to match on grows, it becomes harder to find a suitable match for each unit. "

The approach of propensity score matching is designed to address this problem. Rather than matching on characteristics themselves, propensity score matching takes the approach of estimating an individual's propensity (likelihood or probability) to enroll based on a set of observable characteristics that predict enrollment, assigns a summary score between zero and one, and matches those cases with cases in the non-participant group that have identical or near-identical scores. That 'matched' group represents the comparison group. In practice, this allows for a comparison of participants to non-participants who have similar propensities to participate, but for whatever reason, have not.

Propensity score matching is essentially trying to replicate the unique benefits of randomization. When we randomly assign participants to two groups, what we are doing is ensuring that they have equal probabilities of being in the treatment or the comparison group – they are balanced in this way. By matching based on propensity scores, we are doing something similar: creating two groups which have similar probabilities of being in either the treatment or comparison group. This is another – albeit imperfect – way to create balance between the two groups to make them comparable.

WHEN CAN MATCHING TECHNIQUES BE USED?

Matching techniques are versatile, and can be appropriate in a range of contexts where random assignment is not feasible. They are suited to situations where:

- Administrative data on participants and non-participants exist, or could be created.
- There is good knowledge of the factors that predict program participation (for propensity matching).
- It is reasonable to assume that no unobservable characteristics predict participation, or at least, if they do, to a minimal degree.

THINGS TO CONSIDER

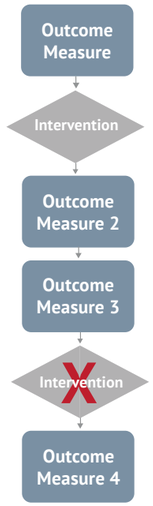
Matching requires large data sets to increase the likelihood of finding good matches for program participants.

- A distance metric must be defined which determines how 'close' any two observations are, given that they vary across multiple variables (different age, different income and so on).
- The overall treated and comparison subsamples must be comparable (e.g. it would be a problem if all the treated individuals were low income, whereas all comparison individuals were wealthy. In that case, matching could not be performed).
- All variables that cause bias must be identified and included in the matching algorithm. Arguments must be made to support the case that there are no omitted variables relevant to the outcome.

REMOVED/INTERRUPTED TREATMENT DESIGNS

Instead of having a control group that is never subjected to treatment, the main idea behind removed/interrupted treatment designs is that data are collected for the treatment group over multiple time periods, where treatment starts and stops several times. The evolution of the outcome during the time when treatment is stopped then serves as a counterfactual for the intervention. If the outcome values reproduce this cyclical pattern, we have reasonable grounds to believe that the treatment has a causal effect. Only other variables which also follow this pattern are potential confounders; we can attempt to control for them in a regression. By design, this method is inapplicable for one-shot treatments such as infrastructure projects.

Figure 12



WHEN CAN REMOVED/INTERRUPTED TREATMENT DESIGNS BE USED?

These designs are well-suited for circumstances where:

- It is not possible to construct a comparison group.
- It is possible to measure outcomes multiple times before and after the intervention.
- The intervention is of the type that can be removed easily (e.g., a rule or a fee, rather than an educational intervention which may have lasting effects).
- The intervention has relatively immediate rather than delayed effects.

THINGS TO CONSIDER

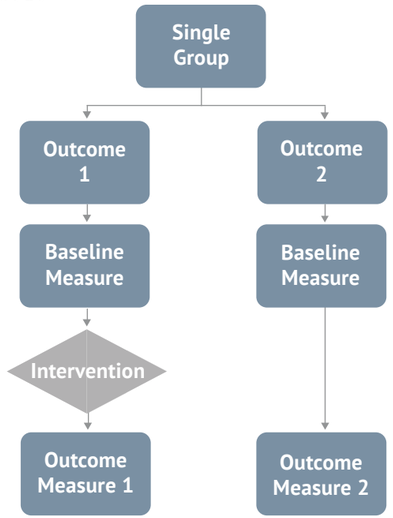
- The intervention must have been activated and deactivated multiple times.
- The outcome variable must respond sufficiently quickly to the intervention. If the effect of the intervention happens with a delay, it becomes much harder to isolate the time periods affected by the intervention from the unaffected ones.

NON-EQUIVALENT DEPENDENT VARIABLES

Another way to make up for the absence of a control group is to find another variable which evolves in a similar fashion to the outcome variable of interest but is not affected by the treatment. The trends in that variable can then be used as a counterfactual.

Reichardt (Reichardt, 2011) gives the hypothetical example of an educational television show for children. The show teaches the letters of the alphabet, one per week to pre-schoolers. After the thirteenth week, once the children have been taught the first but not the second half of the alphabet, they are tested on their knowledge of all twenty-six letters. In this scenario, the true impact of the television program is the difference in scores between the children's knowledge of the first thirteen letters, and their knowledge of the last thirteen letters. This is because we can assume that in the absence of the program, the children's knowledge of the first thirteen letters would have evolved naturally along with the last thirteen letters.

Figure 13



Ensuring that the control variable is not affected by the treatment is a tough requirement which often cannot be verified beyond theoretical reasoning. Furthermore, the control variable must be measured on the same scale as the treatment variable in order to be a suitable counterfactual.

Refinements of this design consist of selecting multiple control variables which are expected to be affected by the treatment to different degrees. The evaluator can then check if the resulting changes in the outcome and control variables match the expected pattern. Where possible, such approaches would greatly benefit by using robust-causal impact estimates of the treatment on the control variables from other studies.

WHEN CAN NON-EQUIVALENT DEPENDENT VARIABLES BE USED?

Non-equivalent dependent variables are a design feature that can be added to any design to improve causal inference, where:

- There exists a variable prone to all of the same things that the outcome of interest is, except the treatment itself.

THINGS TO CONSIDER

- There must be a variable which is measured on the same scale as the outcome variable, but it not affected by the intervention.

CASE CONTROL STUDIES

Case control studies originate in the field of epidemiology. An easy way to understand the logic of case control studies is that they are 'retrospective', or operate in the reverse manner of experiments as described above. Recall that an experiment randomly assigns individuals to two groups, exposes one to an intervention, and then observes the difference in outcomes between the two groups to determine whether the intervention was effective or not. By contrast, in a case control study, we work backwords, observing the outcome, and detecting what its cause (likely) was. These are typically used where the outcome of interest is 'binary' (one either has or does not have it, like a particular disease, or unemployment).

"..., in a case control study, we work backwords, observing the outcome, and detecting what its cause (likely) was. "

These methods are useful in two circumstances. First, where it is unethical to randomly assign – such as to understand the effects of tobacco on cancer. No experiments of this nature exist. Much of the research determining this is based on case control studies, which compared people with and without the outcome (in this case, the negative outcome of cancer), and tried to determine what was different between them that might explain the outcome (in this case, exposure to tobacco smoke).

The other instance where this method might be useful is where the outcome we are interested in is rare, which makes random assignment infeasible, because in an experiment, we would risk assigning individuals to the control condition who might be most likely to achieve the outcome, rendering our efforts useless.

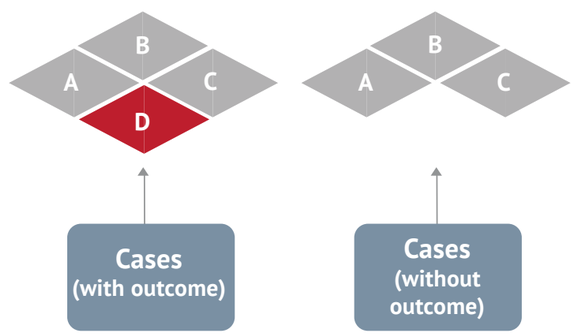
Case control studies rank relatively low in the Impact Canada evidence scale, but are a practical choice for generating initial evidence to build a case for conducting a more robust evaluation.

WHEN CAN CASE-CONTROL STUDIES BE USED?

Case control studies can be used:

- Where the outcomes of interest are rare (e.g., development of a rare disease, elite athlete training programs) and;
- A plausible theory of what causes that outcome exists so that exposure to it can be compared between the cases and the controls.

Figure 14



NON-EXPERIMENTAL EXPLORATORY QUANTITATIVE METHODS

Non-experimental methods are characterised by the fact that they cannot determine the assignment of the program/intervention in any way and that they do not try to replicate an experiment. It is in the latter context in which they differ markedly from quasi-experimental methods. In this sense, non-experimental methods ignore the important issue of the counterfactual. These methods, therefore, do not use control groups or comparison in a robust way and have low levels of internal validity. They are therefore best used as exploratory methods that can then build the case to obtain/provide funding to undertake a more rigorous evaluation using RCTs or quasi-experimental methods.

DIFFERENCE OF MEANS

The simplest approach used for treatment effect estimation is the difference between the average values of the outcome variable for the treatment group and for the untreated group. It is simple in that it does not require information on any further characteristics of the units in the sample. Because of this, it is highly prone to bias, such as selection bias and reverse causality. For this reason, the mean difference estimation is often referred to as the "naïve estimator" because it assumes that the treatment and control groups are equivalent at baseline. Note that only when the treatment and control groups are on average the same at baseline (e.g. in a perfect RCT), is the naïve estimator a valid impact measurement approach.

BEFORE-AFTER COMPARISON

Before-after studies are a common evaluation design. They take the difference in average levels of the outcome for the treatment group before and after the program as a measure of impact. In other words, they assume the impact of the program is the change in the outcome for the participants over time. This is a highly problematic method as the before-after change in outcomes will capture, alongside the impact of the program, the effect of everything else that has affected the treated group in the same period of time. In other words, this method suffers from time effects (history effects) bias. Any positive change in the outcomes over time can be due to a whole host of reasons other than the program itself: the participants grew older, gained more experience, the local economic situation changed etc. With this method, these additional factors are impossible to separate (as discussed above a valid control group is needed to do this). As a consequence, the before-after estimator does not constitute acceptable evidence of a causal impact of an intervention/program.

BENCHMARKING WITH AGGREGATE DATA

In situations where a valid control group does not exist, evaluators sometimes use benchmarks to project counterfactual outcomes. For example, if we know the average proportion of people employed in a treated group after a job training program we could use the national average levels of employment for a similar cohort of people as a benchmark to proxy a control group. These benchmarks can be national or regional averages for the outcome in question. This allows the calculation of a "treatment minus benchmark" effect. These generally have poor internal validity and do not indicate causality, as the treated sample might have different characteristics from the benchmark. Adjusting for these differences (i.e. controlling for differences between the treatment group and the group of people whose data make up the benchmark figure) can help to improve the rigour of this approach. This can only be done if there are unit-level data in the benchmark data (e.g. for the job training program we have data on individuals in the region covering their employment status and background characteristics such as age, gender, education, health, etc...). In general, benchmarking should only be used to provide a very rough initial estimate when exploring the outcomes of an intervention.

QUALITATIVE METHODS

Qualitative methods have an important role to play in understanding the processes behind impact. While the quantitative methods set out in this document can derive estimates with high internal validity, they are to some extent 'black-box' in nature. They can attribute changes in outcomes to a program, however often, these methods do not shed enough light on why the program had the reported impact. Qualitative methods can offer insight on the process of the program and the experience of participants, which is useful for program design and improvement. For example, if a robust quantitative assessment found that a job training program had no positive effect on labour market outcomes for participants, qualitative research can help to uncover why this was the case (e.g. participants may have felt intimidated by the program instructor or felt that the materials were not suitable for their needs). This is valuable additional information for policy analysts and those that design programs and policy interventions.

In terms of internal validity, relying on qualitative approaches on their own is problematic for a number of reasons. First, they tend to use small sample sizes meaning that it is impossible to conduct statistical checks on the data and it is hard to generalise the findings. Second, the findings/results themselves are prone to a range of biases. This is for a number of reasons including:

- Evidence shows that when asked a question about impact over time people (program participants or third party experts) will tend use a before-after comparison to judge the effectiveness of a program. That is, they will assess what their outcomes were like before the program and attribute all changes in outcomes to the program without due consideration of the counterfactual. Thus, qualitative methods tend to inherit the same problems as before-after studies and will typically overstate impact.

- In addition to this, social desirability bias – the desire for survey participants to 'please' the program administrators by providing positive responses – will lead to overstatement of impact. When asked by a program administrator whether the program has had an effect, it is hard for people to say otherwise. Similarly cognitive dissonance can come into play. This is the finding in behavioural science that when there is a contradiction between belief and behaviour people will tend to change their beliefs to not look inconsistent.

Generally, for questions of impact, qualitative research should not be used as a standalone methodology. Qualitative research has a role to play alongside quantitative studies to understand the processes and experiences of program participants to help improve policy design and delivery going forward.

ADDRESSING EXTERNAL VALIDITY

The previous sections have mostly focused on the internal validity of research designs. Once internal validity is assured to a good degree (using the Impact Canada Evidence Scale), external validity will then become an important issue for future roll out and design.

External validity is the extent to which the findings of a study can be generalised to other situations, regions, time periods and individuals. The effects of a policy intervention will often vary given the environment in which the policy is applied, the target population, the timing and implementation details, as well as other context-specific factors. A robust evaluation should be able to demonstrate that the estimated treatment effect is transferrable to other contexts and across sub-populations defined by different levels of some background factor(s) (Cook & Campbell, 1979).

"External validity is the extent to which the findings of a study can be generalised to other situations, regions, time periods and individuals."

Threats to external validity can arise both on the side of data collection and study design, as well on the econometric side. On the data collection side, the main potential threats are lack of generalisability across situations and lack of generalisability across people. The former would occur if the context or details of the experiment make the participants behave otherwise than they would have in a real-world setting. For example, a lab experiment where participants handle fictional money and make consumption decisions that do not involve real goods and services will distort incentives. Participants are likely not to make the same decisions as they would in a real situation. Lack of generalisability across people may arise due to non-representative sampling. For example, the respondents to an online survey are more likely to be young, technologically aware people living in urban areas. They may respond to the program in a different way from the rest of the population.

The results from one study (that has been deemed robust in terms of internal validity) should only be extrapolated to other situations when the context is similar (e.g. similar regional characteristics, similar program participants, similar program in terms of content and duration).

A more technical issue related to external validity is the concept of heterogeneous treatment effects, namely that different individuals in the population react differently to the intervention. The average impact on the population may not be the same as the average impact on those who chose to participate in the program, because the people that signed up are those who would expect to gain a larger benefit from the program. This concept is closely related to, and yet distinct from, selection bias – two individuals can have the same counterfactual outcome in absence of the intervention and yet one can gain more from it than the other. Due to random assignment experimental methods (RCTs) nullify the impact of selection on gains and the impact estimate that they derive is a generalizable impact measure that has high external validity as it represents the impact we would expect on average for the sample used in the study.

The outputs of some methods, however, are even more complex and difficult to interpret, as they capture the impact for only a part of the sample used in the study. For example and as discussed above, encouragement designs estimate the impact only for the sub-population that complies with the encouragement. If compliance is not random, the average impact for the whole population will be different from the average impact for the complier sub-population . RDD estimates also capture a local-type of impact, this time for the sub-population who are close to the cut-off threshold value, which might not be representative for the total target population.

Mathematical methods exist that attempt to neutralize some threats to external validity, provided that certain conditions hold (Pearl & Bareinboim, 2014). Sampling weights are a particular instance of such methods, whereby each unit in the sample is weighted by the ratio of the population group that the unit represents and the size of the respective group in the sample. This makes it easier for the results from the sample in the study to be extrapolated to the general population. Finally, a universally applicable argument to support the external validity of the results of a study is when the results have been replicated by other studies using different contexts or samples. As such, it is important to build up a wide-reaching evidence base of studies using robust methodologies across a broad range of contexts and it is key to conduct comprehensive literature reviews of similar programs and interventions during the impact evaluation as discussed above.

SETTING UP AND CONDUCTING AN IMPACT MEASUREMENT STRATEGY

Here we set out a simple procedure and set of steps for analysts to take when conducting impact evaluations.

1. Determine the program and target group.
2. Conduct a literature review of evaluations of similar programs.
3. Choose the highest-ranking possible method from the evidence scale (given practicality and data constraints).
4. Set out the assumptions of the approach and ensure that you can adhere to them.
5. Conduct the evaluation, wherever possible conducting both quantitative and qualitative research.
6. Set out the results and caveat them wherever necessary. Discuss how robust (internally valid) the results are and compare them to findings form the literature. Note: this could be accomplished with evaluation experts within the Government of Canada and/or in partnership with expert evaluators and research organizations.
7. Finally, discuss generalizability (external validity) of the results.

APPENDIX A - SUMMARY OF METHODS

Method	Description	Assumptions	Advantages	Drawbacks
Randomised Control Trial (RCT)/ Experiment	Splits a sample randomly into treated and control units to ensure there is no selection bias	Treatment is randomly assigned; compliance is perfect; no observer bias; sample size is sufficient.	The best practical approach to yield an unbiased causal effect estimate	Can be costly to conduct, often infeasible
Instrumental Variable (IV) estimation and Encouragement Design	Uses random variation in another variable which drives likelihood of receiving the intervention	The instrument does not affect the outcome directly, but only through the intervention; the instrument is relevant to the intervention	Is able to control for observable and unobservable bias similarly to an RCT, if conducted properly	Valid instruments can be hard to find; encouragement can be ineffective; complier population not representative
Regression Discontinuity Design (RDD)	When the intervention is determined by a cut-off, compares the observations just above and just below the cut-off	Nothing changes abruptly at the cut-off except receiving the intervention	Is able to control for unobservable bias	Requires a specific setup (where a cut-off determines an intervention); may not be generalizable to the population.
Difference-in-Differences (DiD) estimation	Compare the before-after change in the outcome of the treatment group to that of the control group	The outcomes of the two groups would have evolved in parallel in absence of the intervention	Is easy to understand conceptually and allows controlling for some unobservable characteristics	Requires tracking the same individuals over time
Matching	Form pairs of treated and control units with similar background characteristics	All factors relevant for the outcome have been identified and observed; the treatment and control groups are comparable in terms of these factors.	Is valid for any kind of relationship between the outcome and control variables (does not have to be linear)	Requires large data sets; unable to account for omitted variable bias.
Removed Treatment Design	Starts and stops the treatment several times and observed the changes in the outcome variable	The outcome responds immediately to starting/stopping treatment;	A creative approach to overcome the lack of a control group	Is only applicable in very specific situations

Method	Description	Assumptions	Advantages	Drawbacks
Non-Equivalent Dependent Variable (NEDV)	Compares the change in the outcome variable to changes in other, similar variables	The variable used as controls is not affected by the treatment, but is measured on the same scale	A creative approach to overcome the lack of a control group	Is only applicable in very specific situations
Case Control Studies	Separates sample units that have the outcome and those that don't, and compare their exposure to the treatment	All factors relevant for the outcome have been identified and observed	Allows to overcome small sample problems in situations when the outcome is rare	Cannot produce a causal impact of the estimate, but only an odds ratio
Difference of Means	Compare the average outcome values for the treated and control groups	Treatment must be randomly assigned (as in an RCT); otherwise this estimate is mired by selection bias	Very low data requirements and quick to compute; can be used as exploratory analysis	Cannot account for any source of bias due to selection or reverse causality
Before-after comparison	Compare the average outcome values for the treated group before and after the intervention	Nothing except for the intervention changed between the two measurements – this never holds in practice	Very low data requirements and quick to compute; can be used as exploratory analysis	Cannot account for any source of bias due to time/history effects
Benchmarking	Compare the average outcome values for the treated group to a regional or national average	The treated group is the same in terms of all relevant background characteristics as the national average	Very low data requirements and quick to compute; can be used as exploratory analysis	The treated group is likely to be different from the national/ regional average and there is no way to correct for that.

APPENDIX B - MATHEMATICAL PRESENTATION OF IMPACT MEASUREMENT

Let the outcome variable be denoted by Y, and the treatment variable – by D. Our main interest is to find out the causal effect of D on Y. In the simplest setup, D is a binary variable taking the value 1 if the unit is treated and 0 if it is not. The units with D_i=1 make up the treatment group, whereas those with D_i=0 form the control group. For every observed unit i, we define the (hypothetical) values Y_1i and Y_0i to denote the potential treated outcome and potential untreated outcome respectively. These two values refer to the exact same unit at the same point in time under the same circumstances except receiving treatment. The remaining difference, Y_1i-Y_0i is the causal effect of the treatment, or treatment effect.

Then let us denote Y_i the value of the outcome for individual i that is actually observed. It follows that if i is treated, Y_i=Y_1i and Y_0i is unobserved. Conversely, if i is not treated, Y_i=Y_0i and Y_1i is unobserved. The population average treatment effect is E(Y_1i-Y_0i), whereas the naïve difference in outcomes of the treated and untreated group is E(Y_i |D_i=1)- E(Y_i |D_i=0). It can be decomposed as follows:

$$
\begin{aligned}
E(Y_i|D_i=1) - E(Y_i|D_i=0) &= E(Y_{1i}|D_i=1) - E(Y_{0i}|D_i=0) \\
&= E(Y_{1i}|D_i=1) - E(Y_{0i}|D_i=1) + E(Y_{0i}|D_i=1) - E(Y_{0i}|D_i=0) \\
&= E(Y_{1i} - Y_{0i}|D_i=1) + E(Y_{0i}|D_i=1) - E(Y_{0i}|D_i=0)
\end{aligned}
$$

Treatment effect on the treated

Selection bias

As we will see further in this section, only if assignment of the treatment is independent of the potential outcomes, E(Y_0i |D_i=1)= E(Y_0i |D_i=0), does the right-hand term denoting selection bias disappear and the naïve estimator equals the (causal) treatment effect.

WORKS CITED

Angrist, J. D., & Pischke, J.-S. (2015). *Mastering Metrics: The Path From Cause to Effect*. Princeton, New Jersey: Princeton University Press.

Benson, K., & Hartz, A. J. (2000). A Comparison of Observational Studies and Randomized, Controlled Trials. *The New England Journal of Medicine, 342*, 1878-1886.

Bogatz, G., & Ball, S. (1971). *The Second Year of Sesame Street: A Continuing.* Princeton, NJ: Educational Testing Service.

Brown, C. A., & Lilford, R. J. (2006). The stepped wedge trial design: a systematic review. *BMC Medical Research Methodology, 6*(54), 1-9.

Bryman, A., & Teevan, J. J. (2005). *Social Research Methods*. Don Mills, Ontario: Oxford University Press.

Burns, P., Rohrich, R., & Chung, K. (2011). The Levels of Evidence and their role in Evidence-Based Medicine. *Plastic and Reconstructive Surgery, 128*(1), 305–310. doi:10.1097/PRS.0b013e318219c171

Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin, 54*, 297-312.

Concato, J., Shah, N., & Howritz, R. I. (2000). Randomized, Controlled Trials, Observational Studies, and the Hierarchy of Research Designs. *The New England Journal of Medicine*, 342, 1887-1892.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-Experimentation: Design & Analysis Issues for Field Settings.* Chicago: Rand McNally College Publishing Company.

Cook, T. D., & Wong, V. C. (2008, January). Empirical Tests of the Validity of the Regression Discontinuity Design: Implications for its Theory and its Use in Research Practice. *Annales d'Economie et de Statistique, 91*, 127-150.

Cook, T. D., Shadish, W. R., & Wong, V. C. (2008). Three Conditions under Which Experiments and Observational Studies Produce Comparable Causal Estimates: New Findings from Within-Study Comparisons. *Journal of Policy Analysis and Management, 27*(4).

Dunning, T. (2013). *Natural Experiments in the Social Sciences: A Design-Based Approach.* New York, New York: Cambridge University Press.

Farrington, D. P., Gottfredson, D. C., Sherman , L. W., & Welsh, B. C. (2002). Maryland Scientific Methods Scale. *Evidence-Based Crime Prevention*, 13-21.

Gertler, P. J., Martinez, S., Premand, P., Rawlings, L. B., & Vermeersch, C. M. (2016). *Impact Evaluation in Practice - Second Edition*. Washington, D.C.: Inter-American Development Bank and World Bank. Retrieved from http://www.worldbank.org/en/programs/sief-trust-fund/publication/impact-evaluation-in-practice

Heckman, J. J., & Robb, R. (1985). Alternative Methods for Evaluating the Impact of Interventions: An Overview. *Journal of Econometrics*, *30*, 239-267.

HM Treasury. (2011). *The Magenta Book - Guidance for Evaluation*. London.

Jamison, J. C. (May 2017). *The Entry of Randomized Assignment into the Social Sciences*. Policy Research Working Paper, World Bank Group, Development Policy Department.

Moscoe, E., Bor, J., & Barnighausen, T. (2015). Regression discontinuity designs are underutilized in medicine, epidemiology, and public health: a review of current and best practice. *Journal of Clinical Epidemiology*, *68*, 132-143.

OECD. (n.d.). *Evaluation of Development Programs*. Retrieved October 10, 2017, from Outline of Principles of Impact Evaluation: http://www.oecd.org/dac/evaluation/dcdndep/37671602.pdf

Pearl, J., & Bareinboim, E. (2014). External validity: From do-calculus to transportability across populations. *Statistical Science*, 29(4), 579-595.

Reichardt, C. S. (2011). Evaluating Methods for Estimating Program Effects. *American Journal of Evaluation, 32*, 246-272.

Ritchey, F. (2000). *The Statistical Imagination: Elementary Statistics and the Social Sciences*. McGraw-Hill Higher Education.

Rubin, D. (1974). Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. Journal of Educational Psychology, 66(5), 688-701.

Rychetnik, L., Frommer, M., Hawe, P., & Shiell, A. (2002). Criteria for evaluating evidence on public health interventions. *Journal of Epidemiology & Community Health, 56*, 119-127.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference.* Belmont, CA: Wadsworth Cengage Learning.

Sherman, L. W., Gottfredson, D. C., MacKenzie, D. L., Eck, J., Reuter, P., & Bushway, S. D. (1998). Preventing Crime: What Works, What Doesn't, What's Promising. *Research in Brief. National Institute of Justice.*

ACKNOWLEDGEMENTS

The IIU would like to acknowledge the contributions of those who worked to produce Measuring Impact by Design. Its co-authors are:

Craig M. Joyce
Senior Advisor (IIU); co-author and project lead

Daniel Fujiwara
Directot (Simetrica); co-author and reviewer

Iulian Gramatki
Econometrician (Simetrica); co-author and reviewer

The authors would like to acknowledge the contributions of fellow IIU staff for their thoughtful reviews and feedback on various drafts, including Elizabeth Hardy, Senior Lead, Behavioural Insights; David Donovan, Lead, Policy and Innovative Finance; Julie Greene, Lead, Capacity and Partnerships; Victoria Carlan, Lead, Impact Measurement; Haris Khan, Advisor, Behavioural Insights; Alyssa Whalen, Advisor, Behavioural Insights; and IIU Fellows Amanda Desnoyers, Meera Paleja, Lauren Conway, and Guillaume Beaulac.

The authors are also appreciative of the contributions of the IIU's Impact Measurement Technical Working Group, members of which provided excellent feedback, including Sonia Ben Amor, Evaluation Manager, Statistics Canada; Anne-Renee Blaise, Defense Scientist, Department of National Defense; Greg Bridgett, Principal Advisor, Treasury Board Secretariat; Geneviève Boudrias, Senior Evaluation Officer, Canadian Nuclear Safety Commission; Kristina Guiguet, Policy Analyst, Employment and Social Development Canada; Francis Jobin, Senior Analyst, Atlantic Canada Opportunities Agency; Chantal Langevin, Director, Health Canada;

And last but not least, Laurie Bennett, Multimedia Communications Officer, who skillfully produced the final graphic design.