

Methods for Evaluating Impacts of Direct Giving and Cash Transfers

Methods for Evaluating Impacts of Direct Giving and Cash Transfers

Professors Adeline Delavande, Peter Siminski
and Robert Slonim, assisted by Christopher Carter
and Aleksandra Erakhtina

Economics Department
UTS Business School
University of Technology Sydney

Supported by the Paul Ramsay Foundation



About the Authors

Adeline Delavande, Peter Siminski and Robert Slonim are
Professors of Economics at the University of Technology Sydney.
Brief biographies are included in Appendix 1.

Adeline.Delavande@uts.edu.au

Peter.Siminski@uts.edu.au

Robert.Slonim@uts.edu.au

Christopher Carter and Aleksandra Erakhtina are PhD students
in the UTS Economics PhD program and Research Assistants.

June 2023

Contents

Executive Summary	2
1. Introduction	6
2. What is an Impact Evaluation?	8
3. Preparing for an evaluation	10
4. Experimental Methods for Evaluating Cash Transfers	14
5. Quasi-Experimental Methods	22
6. Novel Methods for Evaluation	31
7. Integrating Quantitative and Qualitative Methods	37
8. Australian Data Landscape	38
9. Baby Bonus Natural Experiments	44
10. Coronavirus Supplement Natural Experiment	49
11. Summary Comparison of Techniques	52
12. Recommendations and Conclusion	57
13. References	67
Appendix 1 Biographies of Key Personnel	72
Appendix 2 Designs Used to Create Randomised Variation in Exposure to the Program	73
Appendix 3 International Literature Review: Cash Transfers	76
Appendix 4 Glossary	82

Executive Summary

This report explores the available methods of quantitatively evaluating the impact of social programs; in particular, programs that involve direct giving and cash transfers to vulnerable families with young children. It considers the relative strengths and weaknesses of various options, and offers suggestions about which option(s) to pursue at this time in the Australian context.

Methods for Evaluating Cash Transfers

The report includes five sections which together provide a detailed overview of the evaluation methods that may be relevant to evaluating the impact of cash transfer programs.

Section 2 outlines what is meant by impact evaluation, and describes the ‘fundamental problem of causal inference’ which all impact evaluations must address. Section 3 discusses key considerations that arise when preparing an evaluation, including developing theories of change and conceptualising a results chain, posing appropriate evaluation questions, and choosing outcome variables and performance indicators.

Randomised experiments are the gold standard of causal evaluation methods. Section 4 presents a detailed discussion of randomised experiments. It includes the rationale for random assignment of treatment when the causal impact of the program is being estimated. It discusses the three program elements which can be randomised (access, timing and encouragement), as well as how to choose the level of randomisation, along with key considerations such as spillovers, attrition, compliance, and external versus internal validity.

Quasi-experimental studies can also sometimes provide strong evidence of program impact. Section 5 describes the suite of mainstream quasi-experimental methods, including Regression Discontinuity Design, Difference-in-Differences, Matching and Instrumental Variables.

Section 6 gives an overview of relevant novel and emerging impact evaluation methods. These include adaptive trials, Bayesian adaptive trials, synthetic control, machine learning, event studies, and combining randomised controlled trials (RCTs) with structural modelling.

Section 7 briefly discusses the benefits of combining qualitative techniques with quantitative impact evaluation methods.

Australian Data Landscape

In Section 8, we turn to potential data sources. We focus mainly on linked administrative data sources, especially the NSW Human Services Dataset (HSDS) and the Multi-Agency Data Integration Project (MADIP). Australian governments have invested substantially in linked administrative data in recent times. These may prove particularly useful for retrospective quasi-experimental evaluations, but could also link to new data as part of prospective experimental designs. At present, there is some uncertainty about the ease of access to this data, including cost, request approval time, likelihood of approval being granted, and access to key fields (such as date of birth).

Bank transaction records are an emerging data source with much untapped potential. Bank transactions capture people's expenditure, their saving/investment decisions, income and labour force participation, and government program participation; they can also be analysed to unpack intra-family financial decision-making and income pooling. The Commonwealth Bank of Australia (CBA) has formalised contractual arrangements with some universities (including UTS) which could see detailed transaction data used to explore questions of joint interest. Initial discussions with the CBA suggest that these data may be suitable for studying the short-run effects of the Coronavirus Supplement (an extra \$550 paid fortnightly on top of income support), including identifying the population of greatest interest. Transaction data are also made available by credit bureau illion, and have been used by the e61 Institute to evaluate the 'natural experiment' of the Coronavirus Supplement.

We also discuss well-known representative sample surveys, such as HILDA and LSAC, which have the advantage of comprising very detailed outcome variables, but offer relatively low statistical power.

Baby Bonus Reforms

The Baby Bonus is highly relevant to the present context as it is an unconditional cash payment provided to every family on the birth of a child. Several reforms to the Baby Bonus are well suited to analysis using quasi-experimental techniques:

- An increase in the payment rate for babies born on or after 1 July 2004
- A change from a lump sum to recurring payments (totalling the same value) for babies born on or after 1 January 2009
- A decrease in the payment rate for babies born on or after 1 March 2014.

These reforms could be evaluated using Regression Discontinuity Design or Difference-in-Differences techniques. The 2004 reform has been analysed by several studies, especially de Gendre et al. (2021), and Breunig and Deutscher (2018), who use quasi-experimental techniques to study effects on use of medical care in South Australia, and Year 3 NAPLAN results, respectively. This work could be extended in many ways. In particular, a range of additional outcomes could be studied for children and their parents. The NSW HSDS seems to be the most promising source of data for this purpose. However, children's date of birth (DOB) would be essential for such an evaluation, and it is unclear whether DOBs could be accessed by special request, either directly or possibly through MADIP.

Advantages of Studying Baby Bonus Reforms

- Can be studied retrospectively up to a child reaching age 18 using existing data
- Relatively high statistical power
- DOB almost wholly determines version of payment received, meaning it is unnecessary to link with data on payment records.

Disadvantages

- Relatively small payment
- Can only examine effects of specific reforms rather than questions of most interest
- Limited to outcome variables available in administrative data that also contain DOB.

Coronavirus Supplement Natural Experiment

The Coronavirus Supplement was a recent (2020), large and sustained increase in direct transfers for people receiving government payments. It is therefore highly relevant to the present context. In some ways, the Coronavirus Supplement is less promising as a natural experiment than the Baby Bonus, mainly due to the anticipated difficulty in forming a credible comparison group, or otherwise inferring counterfactual outcomes. On the other hand, initial discussions with the CBA suggest that their data may be well suited to studying the short-run effects of the Supplement in rich detail and for the relevant population. This may be worth exploring further, potentially in partnership with the e61 Institute, who are already working in this area.

Advantages

- Relatively large and sustained increase in income
- Opportunity to learn how (additional) transfers are spent (or saved), using bank transaction data, and to ascertain whether they cushion negative shocks for parents or children (using NSW HSDS).

Disadvantages

- People move in and out of payment types over time (before and after implementation), making it difficult to select a valid comparison group
 - In that context, only large and short-run effects are likely to be detectable
 - Uncertainty about ability to identify the population of interest (disadvantaged parents of young children who received the payment) in any suitable dataset.
-

How each technique could provide evidence for the effectiveness of Direct Transfers

In section 11, we discuss how each of eight impact evaluation techniques (all discussed in more detail elsewhere) could be used to generate relevant evidence in various cash transfer scenarios. The approach is to discuss 'best case' examples of how each technique could be applied. The important distinction between prospective and retrospective evaluation is emphasised throughout. The section concludes with a table summarising the key features of each technique.

Suggestions and Conclusion

Suggestion 1: RCT

RCTs provide the best possible evidence on the causal impact of direct giving and cash transfers. There are many reasons for this, including an RCT's control of the exact conditions in which a cash transfer may occur, its ability to measure a broad range of outcomes, the highly credible and transparent nature of its results, and the fact it has the highest possible external validity. An RCT offers the potential to influence the Australian policy landscape and make a major policy impact, in the event that cash transfers are later rolled out at scale.

Suggestion 2: Baby Bonus

There may be scope to evaluate the three Baby Bonus reforms using quasi-experimental techniques. However, this depends on the availability of access to relevant data, specifically the DOB field in the NSW HSDS database. This access would need to be approved by special arrangement.

Suggestion 3: Coronavirus Supplement

The Coronavirus Supplement natural experiment is to some degree a less promising focus for evaluation, given the difficulty in forming a valid comparison group. However, it remains worthy of further exploration, particularly given the potential of NSW HSDS and emerging sources of financial transaction data, such as those of the CBA.

1. Introduction

1.1 Background

Direct cash transfers are one of many potential programs which may improve the well-being of vulnerable populations in Australia. The effects of such transfers on recipients are hence a policy-relevant question, and likely to be contingent on many contextual factors.

This report explores different methods of quantitatively evaluating the impact of social programs, particularly those involving direct giving and cash transfers. It provides an overview of the spectrum of technical approaches that could be deployed to evaluate the impact of direct cash transfers, and the relative strengths and weaknesses of those approaches.

1.2 Key Questions

General

- What are the research design and statistical analysis options that could be used to generate evidence about the impact of cash transfers on outcomes related to disadvantage?
- What are the relative strengths or weaknesses of these options? Consideration is given to: 1) quality of evidence, including validity and reliability of results, and 2) practical and ethical considerations including cost, burden on participants and timeframe.

Specific to Potential Options

- For options using secondary/administrative data: What historical moments in Australia can be used to approximate a cash transfer intervention? What are the pros and cons of using these events?
 - For options using secondary/administrative data: What data are readily available in Australia for investigating the impact of cash payments?
 - For options related to experimental studies: What are the specific ethical challenges and mitigation strategies?
-

1.3 Hypothetical Scenarios

The report is framed around some potential scenarios for evaluating direct cash transfers in Australia.

Scenario 1: cash transfer

Direct cash transfers provided to a cohort of vulnerable families with newborns. The cash transfers total \$10,000 up to the age of 3 years, and consist of equal monthly payments. Some recipients may also participate in an early childhood wraparound support program that aims to improve developmental outcomes in children.

The families who receive the payment live below the poverty line, and therefore have a higher prevalence of risk factors, such as a history of interaction with the justice system, experience of domestic or family violence, at least one parent experiencing long-term unemployment, living in a single-parent household or having other children currently or formerly in the out-of-home care system.

Scenario 2: secondary administrative data only

A study investigating the efficacy of direct cash transfer(s) using only secondary administrative data. This may look at historical cash transfers that supplemented income in Australia; for example, COVID-19 payments and/or the Baby Bonus.



2. What is

Impact Evaluation

Impact evaluations seek to assess the impact of a program on a set of outcomes.

This causal impact is the difference in outcomes that is caused by the program. For example, we want to evaluate whether cash transfer programs cause better health outcomes in children. The causal impact is also called the **treatment effect**.

Simply observing that a child's health outcome improves after her family receives a cash transfer is not sufficient to establish causality. The child's health might have improved even if her family had not received the cash transfer. It may be that parents who receive a cash transfer are highly motivated to improve their children's well-being and ensure they eat a healthy diet. Furthermore, people who participate in the program may have different unobservable characteristics from people who do not participate. This is what is called **the selection problem**. Such complications may bias naïve estimates of program impact.


In an ideal scenario, in order to precisely evaluate the impact of a program, we would measure a child's health after her family received the cash transfer, and measure the same child's health without her family receiving the transfer. We would then compare the two health outcomes to establish impact, and we would know that any difference between those outcomes had been caused solely by the program. Nothing else in relation to that child would have changed, so nothing else would explain the difference in outcomes.

This is the "fundamental problem of causal inference" (the **counterfactual** problem): we can never observe the same people at the same time, both with and without the program. Because we cannot observe the counterfactual, we need to mimic it by finding a suitable comparison. But finding an appropriate comparison for the child in the above context is challenging because she is unique. Her exact family background and genetic attributes cannot be found in any other child.

Moving from the individual to the group, we can rely on statistical properties to generate two groups of individuals that, if their numbers are large enough, are statistically indistinguishable from each other at the group level. The group that receives the program is called the *treatment group* while the group that does not receive it is called the *comparison (or control) group*. The challenge of an impact evaluation is to identify a treatment group and a comparison group that are statistically identical, on average, in the absence of the program.

We review in this report different empirical methods that have been developed to identify a credible comparable group. Experimental methods (e.g., randomised controlled trials) are considered the gold standard because the random selection of the control and treatment groups eliminates the potential bias inherent in other empirical designs. The results obtained from experimental methods are hence extremely credible. The use of quasi-experimental methods involves finding comparable groups in a context where the program has not been randomly allocated. It requires imposing some assumptions that are typically not testable, which may weaken the credibility of the conclusions.

A fundamental difficulty in empirical research is deciding what assumptions to maintain. Stronger assumptions yield stronger conclusions. However, there is a tension between the strength of assumptions and their credibility. Manski (2007) called this the **Law of Decreasing Credibility**: The credibility of inference decreases with the strength of the assumptions maintained.



Impact evaluations seek to assess the impact of a program on a set of outcomes.

3. Preparing for an Evaluation

The initial steps of an evaluation set-up are essential when preparing for an evaluation. They involve building up a theory of change (how to achieve the project's expected results); constructing a results chain; indicating the evaluation questions; and determining the performance-assessing indicators.

Preparation also involves communication with key stakeholders in order to establish a common vision and to reach consensus on the project's objectives and the questions to be posed. The initial steps of an evaluation set-up, as well as this section, draw heavily on (Gertler et al., 2016).

3.1 Step 1: Theory of Change

Briefly, a theory of change implies providing an overview on how to achieve the intervention's expected result(s). In describing the main logic of how to reach a particular target, one needs to: define a sequence of events and the outcome(s) it will lead to; state the underlying conditions and assumptions that are obligatory for an outcome to be realised; and map the program intervention according to logical cause-effect relationships. These steps will help to differentiate the meaning inputs and activities, the possible delivered outputs and the consequential outcomes, especially for programs aimed at behavioural change. When designing the program, it is extremely important to review and reference the literature on similar programs, as well as the theory behind the causal relationships. Figure 1 presents an example of a Theory of Change diagram for Cash Transfer.

3.2 Step 2: Results Chain

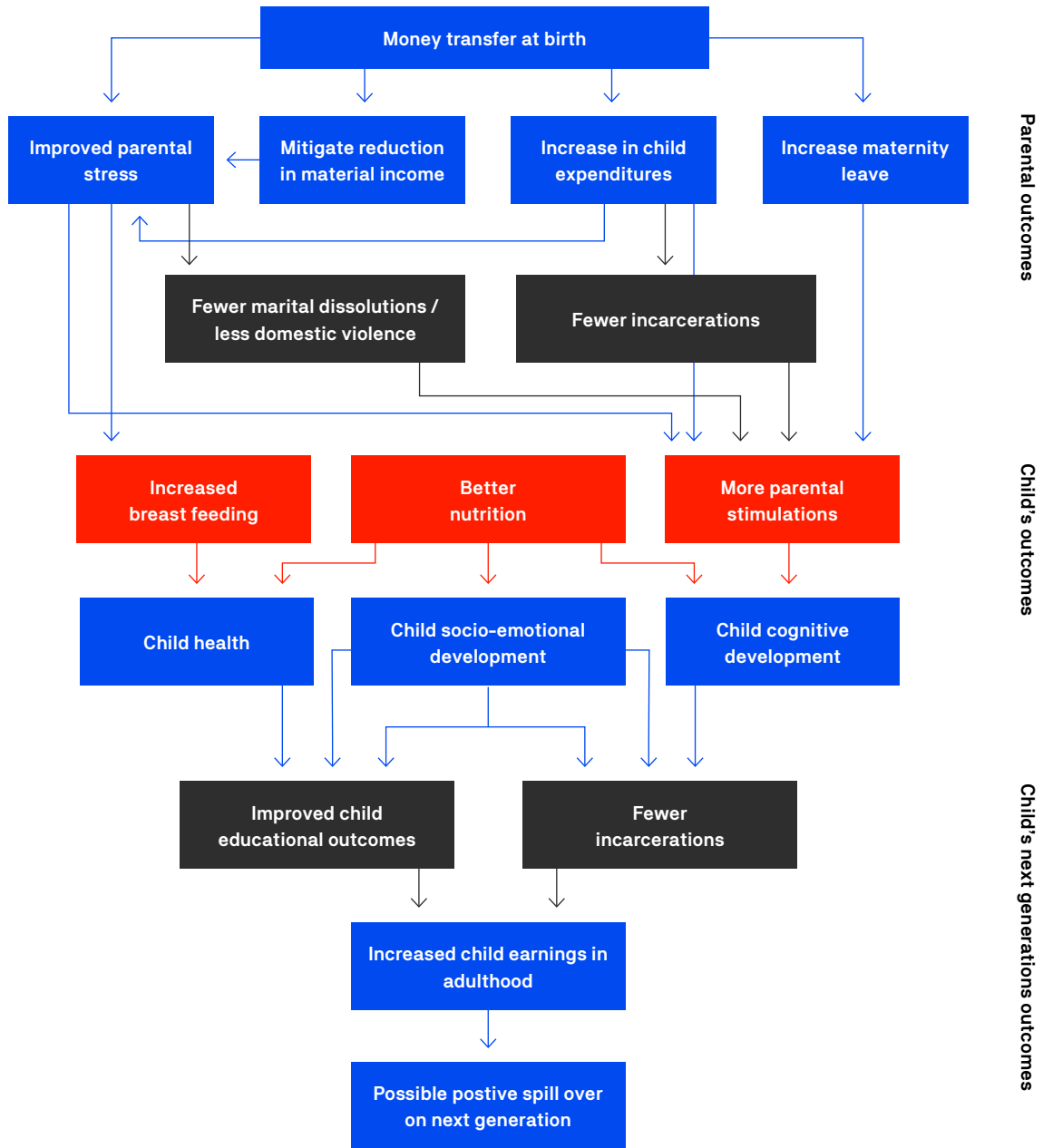
The simplest and clearest model outlining the theory of chain, the results chain helps us to visualise the interaction of inputs, activities and outputs, with behaviour determining the ways to achieve the impacts (as well as assumptions and risk). A results chain maps:

- **Inputs¹:** available resources and budget.
- **Activities¹:** actions converting inputs into outputs.
- **Outputs¹:** goods and services produced by project activities.
- **Outcomes²:** likely results that follow the outputs being used by the population receiving them (not controlled directly; short-term).
- **Final Outcomes²:** results displaying whether the project's goals are achieved (long-term).

1. Implementation (supply side).

2. Results (supply and demand sides)

Figure 1: Theory of Change Diagram



3.3 Step 3: Evaluation Questions

Formulating an evaluation question is an important step for making sure that the research reflects the purpose of the policy (intervention). A basic evaluation question could be: 'What is the impact (causal effect) of a program on an outcome of interest?'

An evaluation question should focus on the impact of the program on the final outcomes for the treatment group (beneficiary population), or it should compare the program modalities for their cost- (or outcome impact) effectiveness. The differences thus should be quantifiable.

It can also be a testable hypothesis, which can then be accepted or rejected according to the evidence provided by the impact evaluation.

3.4 Step 4: Outcome and Performance Indicators

Outcome measures are used to assess the effectiveness of a program in terms of initially stated objectives. Selecting the key outcome indicators enables the setting of clear objectives for the program's success, in terms of intended effect sizes (changes) for each chosen indicator. For example, the intended effect size might be a specific change in school test scores, or in the take-up rate of a new insurance policy, etc.

Along with program effectiveness, outcome indicators form the basis of power calculations. If the sample size appears too small to detect the consequent changes, the impact evaluation may be 'underpowered' and may fail. Hence it is crucial to specify the program's success criteria; i.e., the minimum expected effect sizes. It is also helpful to conduct ex ante simulations with available data, comparing different outcome scenarios and related expected effect sizes, or comparing preliminary measures of the cost-effectiveness of alternative interventions in relation to previously chosen outcomes.

Indicators must be selected for both the implementation and evaluation stages, and should be SMART: specific, measurable, attributable, realistic and targeted. Such indicators enable us to track the causal logic of outcomes, and to check whether the intervention has been carried out according to plan (Kusek & Rist, 2004).

When selecting the indicators, it is useful to determine the source of the data, the frequency of their collection (timeline); the key responsibilities for data collection, analysis and reporting; necessary resources to produce the data; data documentation; and any possible risks.

This report explores the available methods of quantitatively evaluating the impact of social programs; in particular, programs that involve direct giving and cash transfers to vulnerable families with young children.

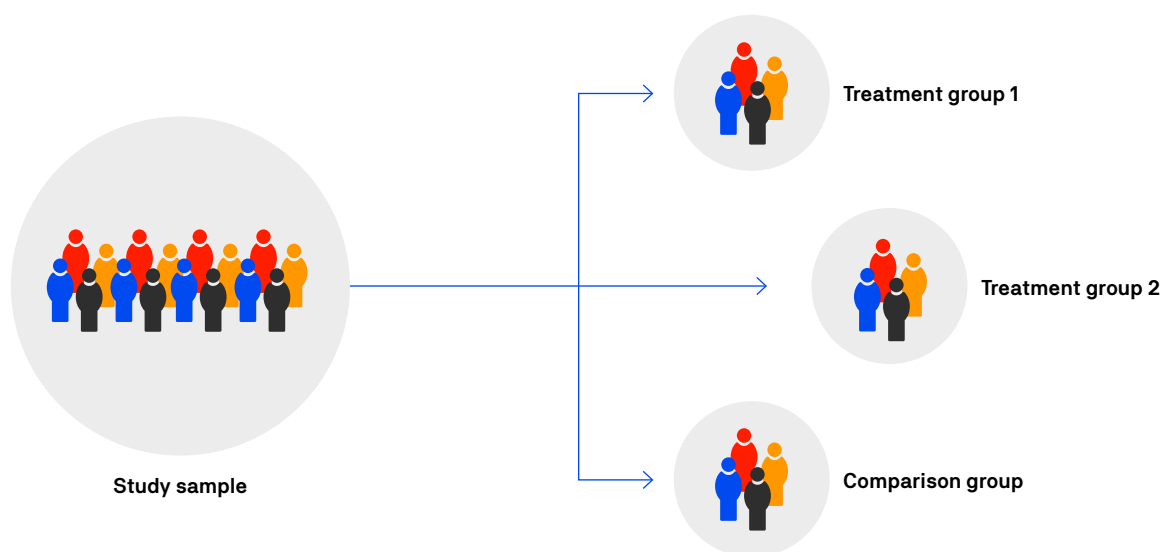
4. Experimental Methods for Evaluating Cash Transfers

4.1 How can the Random Assignment to a Program help?

Randomised experiments are considered the gold standard of impact evaluation. Random assignment to a program means that participants and non-participants are chosen at random. Chance alone is responsible for them being selected for the program. With a large enough number of individuals, the randomised assignment process will produce groups with statistically equivalent averages for all their (observed and unobserved) characteristics. They would, on average, achieve the same outcomes. This assumption that the treatment and control groups are statistically identical, with no significant differences in observed characteristics, is easy to check empirically. If it is confirmed to be the case, the after-program differences between the outcomes of the treatment and control groups can be fully attributed to the program itself. It is the causal impact of the program (or *treatment effect*).

We illustrate the procedure of random assignment in Figure 2 below, in a context where there are two treatment groups; i.e., two groups that receive different versions of the program, and a comparison group.

Figure 2: Randomisation creates groups with similar characteristics



4.2 Estimating the Causal Impact of the Program

To estimate the causal impact of the program, one can simply take the difference between the mean outcomes of those randomly assigned to receive the program (treatment group) and the mean outcomes of those randomly assigned to the control group. The estimates give us an understanding of what happens to an average person (unit of observation) if given access to the program. The estimated impact constitutes a **credible estimate of the true impact of the program**, since all observed and unobserved factors that might otherwise plausibly explain the difference in outcomes are identical for the treatment and control groups.

The impact of the program is often obtained using regression analysis. Including covariates in the estimating regression can deliver a more precise estimate of the impact of a program.

The program's impact may vary for different sub-groups of the eligible population studied. In such a case, the treatment effects are *heterogeneous*. Testing the program's effectiveness for different sub-groups with common characteristics may help the program to be better targeted in future, and to understand the transmission mechanisms.

Sometimes, one may also observe the outcomes of interest at baseline before the start of the program. It is possible to either add the baseline outcome as a covariate to the regression, or to calculate the changes in outcome as a difference between the baseline and endline measures, then estimate the impact of the program on the change in outcome. This process implies making specific assumptions about the relationship between baseline and outcomes.

A well-known paper by Gertler (2004) analyses the impact of the Mexican conditional cash transfer program Progresa on child health. This anti-poverty program provides cash transfers to low-income households conditional on them engaging in a set of behaviours designed to improve health, nutrition and education. Every two months, eligible families receive a cash transfer typically worth about 20% to 30% of household income, providing the conditions are met. The empirical analysis leverages the randomised design implemented by the Mexican government. Due to budgetary and logistical constraints, the government randomly chose 320 treatment and 185 control villages. Eligible households in treatment villages started receiving the transfers in August 1998, while transfers for eligible households in control villages were deferred for two years. The Progresa program had a positive effect on child health. Children born in treatment villages during these two years experienced an illness rate in the first six months of life that was 25% lower than that of control children. Treatment children were also 25% less likely to be anaemic, and grew about 1 centimetre more during the first year of the program.

When we compare the average outcomes of people randomly assigned to receive a program (the treatment group) with those assigned to the control group, we get what's called the **Intention-to-Treat** (ITT) impact. This measures the impact of having access to the program, regardless of whether or not people actually use it. However, sometimes not everyone who is eligible actually uses the program. In cases like cash transfers for low-income families, there can be stigma that lowers the number of people who take it up (Moffitt, 1983).

In these situations, we might also want to know the average effect of the program on those who do use it (called the “compliers”). We can use a method called the Wald estimator for this. It assumes that the difference in outcomes between the treatment and control groups is only due to the extra people who use the program in the treatment group. If this assumption holds, we can convert the ITT estimate into an estimate of the impact on those who actually use the program by dividing it by the difference in the take-up rates between the treatment and control groups. This is referred to as the average treatment effect on compliers.

Another way to estimate the effect on program users is through instrumental variable regression. In this method, the “instrument” is the random assignment of the program itself (see section 5.4., Instrumental Variables).

4.3 What can be randomised?

There are three basic elements of a program which can be randomised:

- **Access:** we can choose which people are offered access to a program
- **Timing:** we can choose when people are offered access to a program
- **Encouragement:** we can choose which people are given *encouragement to participate* in a program. The encouragement can be relatively minor, such as a letter or phone call reminding people of their eligibility and detailing the steps they can take to enrol in the program.

We describe in detail in Appendix 2 the various designs that can be used to create randomised variation in exposure to the program.

In the context of cash transfer programs, randomising access to the transfers is an obvious option. One group would not receive any transfers, while possibly two or three other groups would receive different versions of the program (e.g., with different timings for the onset of transfers, or different installment, etc). With an early childhood wraparound support program, an encouragement design is a natural option, with participants randomly provided encouragement to participate.

4.4 Level of randomisation

Program evaluators need to decide whether to randomise individuals/households or clusters (communities, schools, other units). The larger the number of units randomised, the higher the statistical power, as the outcomes of people in the same unit are sometimes interdependent. Statistical power is the ability to detect an effect of a given size.

However, the level of randomisation should be chosen not only for its statistical power, but also based on other factors that can affect the implementation and evaluation stages of the program. With a cash transfer program, one could randomise at the household or community level. Randomising at the household level may lead to spillover effects if treatment and control households interact (see Threats to the integrity of an experiment, below). It may also lead to differential attrition between the treatment and control groups, if the households in the control group feel they are missing out on the transfers.

4.5 Threats to the integrity of the experiment

It is rare that an RCT goes entirely according to plan. Here, we discuss common threats to the integrity of experiments. These threats are important to consider because they imply that the control group may no longer be a good counterfactual for the treatment group(s).

Spillovers

Programs can have effects that go beyond the immediate participants, and these effects are known as spillovers, or externalities. Spillovers can be either positive or negative and typically occur through different channels:

- **Physical spillovers:** An example of physical spillovers is seen in immunization programs. When a certain population is immunized, it not only benefits the individuals directly receiving the immunization but also reduces the overall transmission of diseases in the community. In this case, the positive impact of the program extends beyond the immediate participants to protect others as well.
 - **Behavioral spillovers:** Behavioral spillovers occur when the behavior of individuals in the treated group influences the behavior of others in the population. For instance, if a program promotes healthy eating habits or encourages energy conservation, some individuals outside the program may observe and imitate those behaviors. This leads to a broader impact beyond the program participants, as their behavior serves as a model for others to follow.
 - **Informational spillovers:** Informational spillovers happen when a program provides valuable knowledge or information that spreads to the wider population. For example, a program focused on promoting effective parenting techniques or sustainable farming practices can share valuable insights and techniques. As this information circulates, it benefits individuals who were not directly part of the program but learn from its findings and recommendations.
 - **Marketwide spillovers:** Marketwide spillovers occur when a program creates market imbalances that can affect individuals who do not qualify for the program. For instance, if a program provides job training or education subsidies to a specific group, it may create a competitive advantage for the participants in the job or education market. This advantage can disadvantage those who do not qualify for the program, leading to marketwide spillover effects.
-



Understanding these different channels of spillovers helps in comprehending how programs can impact not only the individuals directly involved but also the broader community and systems in which they operate. With spillover, the difference in outcomes between treatment and control groups no longer represent the effects of the program itself.

To avoid spillovers, it may be possible to choose an appropriate level of randomisation. Ideally, treatment and control groups should not interact, nor have anything in common that can act as a transmission channel. The unit of randomisation must be chosen such that interactions occur only within groups, not across them. For example, randomising an information program at the classroom level is likely to have high spillover effects, as children will share the information with their school friends in other classes. In this example, randomisation at the school level might be preferable for reducing transmission channels.

Sometimes, rather than seeking to avoid spillovers, it is possible to estimate spillover effects, by observing the different outcomes of the 'spillover' and 'no spillover' groups. These measures help us to choose the optimal number of individuals to be treated in order to achieve the desired outcomes. Since the extent of spillover depends on the proportion of treated individuals in the vicinity of an untreated individual (treatment density), this extent can be varied for the purposes of the evaluation. The best option would be a two-level randomisation, group and individual; first the outcomes of the treated and control groups would be compared, then the difference between the individual outcomes for 'spillover' (within the treatment group) and 'no spillover' (within the control group) modifications would be measured.

Attrition

Attrition occurs when it is impossible to measure the outcomes of some participants in the program. In other words, data are missing, because some participants have either dropped out of the program, or refused to be interviewed, or cannot be tracked after the program. Comparability of the treatment and control groups can be compromised if their attrition levels or types are different. Attrition is more often an issue with control groups, as people not receiving a treatment (especially when it involves a clear benefit, such as a cash transfer, which they see others receiving) are more likely to decline to participate in a survey or experiments measuring the outcomes.

There are ways to lower the level of attrition (Glennester & Takavarasha, 2013):

- 1. Provide access to the program for everyone, but over time.**
- 2. Change the level of randomisation.**
- 3. Improve the processes of data collection.**

With the first approach, the knowledge that the program will become accessible in due course will reduce a person's unwillingness to participate in the study. With the second, a higher level of randomisation will ensure that those who are not treated do not closely observe those who are, thus reducing resentment. With the third, data collection may be improved so that participants can be located/tracked, or provided with incentives to complete the survey.

In any event, data collection procedures should be designed to minimise the level of attrition. For example, surveys should be neither too complex nor too long, so as not to demotivate participants. The administration of surveys is also important, with particular care needing to be taken with sensitive questions. If some respondents are absent during the survey, it is preferable to run a survey with several rounds, ensuring the highest response levels. Moreover, time gaps between surveys should not be too long, to minimise the risk of losing some respondents due to them moving, changing school, etc. Initiatives emphasising the importance of feedback are important, and survey participation can also be induced through minor compensation for time spent.

Compliance

The ideal experiment implies adherence to the treatment/no treatment assigned. However, this is sometimes violated by either program staff or participants. In either case, it may lead to biased estimates of program impact. Implementation staff may depart from the protocol by deviating from the random assignment and providing treatment to people who they think are more in need, or based on a personal relationship. One possible solution is to ensure that program staff are not faced with such decisions; for example, by randomising at the staff level, so that each staff member is responsible for only one version of the program.

The situation becomes more complicated when compliance is violated by participants. There are several scenarios in which this can occur. First, treatment group members may choose not to comply with the rules of the treatment, and may stop participating in the program – or, if the participants are children, their parents may not consent to an intervention. The extent of such partial compliance can be measured by checking how many participants have not complied with the treatment. Second, control group members can end up receiving the treatment, either via one of the transmission channels described in the ‘spillover’ sub-section above, or because they receive a similar or identical treatment (such as a healthcare service, subsidy, scholarship or training program, etc.) through a different avenue. Third, providing access to the program can sometimes have a counter-intuitive effect on take-up; in other words, take-up rates decrease when a treatment is assigned, and increase when it is not assigned. People acting in this manner are called defiers, since they behave counter to what is intended in the experiment (Glennester & Takavarasha, 2013).

Whereas partial compliance can reduce comparability between treatment and control groups, defiers can actually obscure true estimates of the program’s effect. When encouragement design has the effect of discouraging people from taking up the program, it is almost impossible to evaluate the program’s impact. However, both problems can be limited via respective design modifications:

- **Easy take-up of the program** (clear and transparent application and implementation procedures).
- **Provision of incentives** (encouraging participation through small incentives that will not affect the program outcomes).
- **Distribution and simplification of field tasks** (random assignment of staff members to different designs of the program and provision of training to staff).
- **Randomising at a higher level** (minimising spillover effects between the treatment and control groups).
- **Provision of a basic as well as an advanced program** (provide basic service to all groups including control, so that everyone receives some service). (This makes it impossible to measure ‘no effect’ outcomes, but it is possible to evaluate the difference between basic and advanced programs).

Another way of limiting these problems may be to identify levels of compliance and identify defiers among the participants. The former can be measured by adding questions on take-up to the endline survey, documenting the level of compliance during the implementation stage for the treatment group, and monitoring the control group’s potential access to treatment (this can sometimes encourage take-up, so should be performed carefully). The latter can be accomplished by first identifying possible ways in which encouragement can affect participation in the program, and then adding subsequent indicative questions to the survey in order to determine whether a participant is a defier. Once completed, it is possible to subtract the impact of the program on defiers from the overall impact and deduce the program’s true impact.

4.6 External and Internal Validity

With an RCT, both internal and external validity are mostly ensured, but not completely. Internal validity means that the estimated impact of the program is unbiased. It is solely due to the treatment itself. Since the randomised assignment provides true counterfactual estimates in the absence of treatment, the estimated impact is the effect of the program, and internal validity is met. External validity is achieved when the conclusions from a particular study are valid for another context. RCTs can be designed so as to be tailored to specific contexts.

Randomisation is used both to select the sample from the eligible population, and to assign the treatment within the sample. If both randomisation techniques are used, the impact evaluation produces internally valid estimates of the program, and these estimates are generalisable to the eligible population. At times, impact evaluation can lack external validity, and while internal validity is necessary for external validity, it is not sufficient (Duflo et al., 2008).

First, it is important to mention that randomised evaluations are sometimes not able to pick up general equilibrium effects. For example, when a voucher program is being evaluated in a specific area, and the outcomes for people who were given a voucher and those who applied for but did not receive one are being compared, the impact of securing a voucher can be identified, given that the voucher system was introduced. However, the estimated effect measured is only partial (localised), and does not constitute the overall effect on the education system in that area (Duflo et al., 2008).

Second, the evaluation can sometimes have a significant impact on the behaviour of people in the treatment (Hawthorne effect) or control groups (John Henry effect). In the former, treated individuals may change their behaviour during the observation period (for example, they may strive to succeed). In the latter, control group members may also behave differently during the observation period (for example, they may 'compete' with the treatment group) (Duflo et al., 2008).

It is sometimes questionable to what extent even RCT results can be extrapolated or replicated. It depends on the complexity of the implementation process and the specificity of the sample chosen. Mostly, it is difficult to predict whether a slightly different program or a slightly different target population would achieve the same results. Expanding a program may also reduce the quality of its implementation. Moreover, randomised evaluations are often conducted in relatively small and 'convenient' regions, meaning that external validity can be limited. Some evidence suggests that programs tested in different environments may produce quite similar results. Such replications also point to the particular importance of the conditionality factor for program effectiveness (Schady & Araujo, 2006).

5. Quasi-Experimental

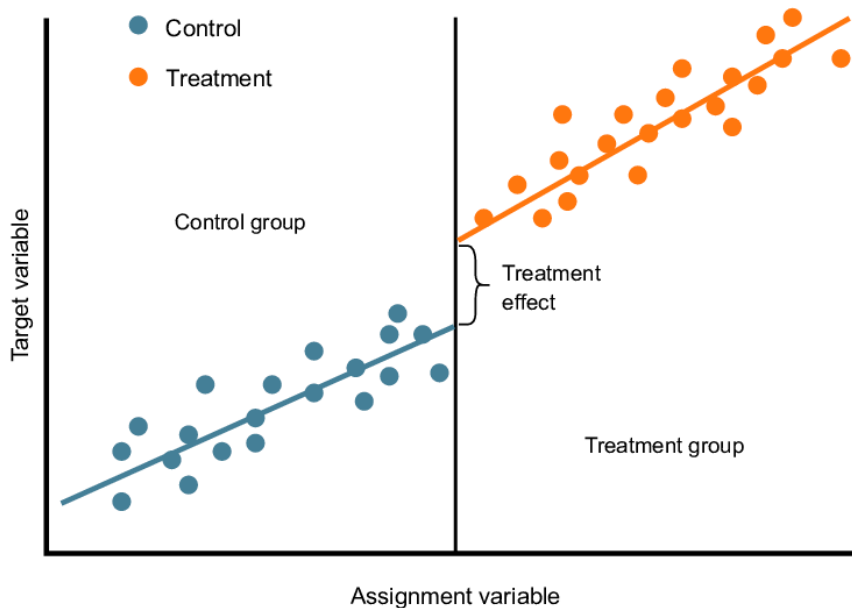
Methods

The use of quasi-experimental methods to evaluate natural experiments has been influential in the identification of important causal relationships in economics. The achievements of the 2021 Nobel Prize winners in economics (Joshua Angrist, David Card and Guido Imbens) demonstrate just how revolutionary these techniques have been.

5.1 Regression Discontinuity Design (RDD)

Regression Discontinuity Design (RDD) has been a popular method in economics for analysing the causal effects of a policy by exploiting the quasi-natural experiment created by a policy threshold or cut-off. A policy threshold establishes who is eligible for treatment and who is not, based on which side of the threshold they lie. RDD analysis then estimates the causal impact of the policy by comparing the average outcome of those who only just qualified to receive the treatment with those who only just missed out. The intuition is that individuals slightly above or below the cut-off are not inherently different, and thus make valid comparison groups, thereby mimicking a randomised experiment around this cut-off. First pioneered by Thistlewaite and Campbell (1960), if correctly applied to a valid setting, the RDD analysis is a powerful tool for estimating causal impacts.

Figure 3: The Regression Discontinuity Design Model



Source: Liu et al. (2022)

Equation 1: A Linear RDD Regression Model

$$Y_i = \alpha + \gamma X_i + \beta D_i + \delta X_i D_i + \varepsilon_i$$

X_i is the eligibility variable or assignment variable

For example, the influential contribution of Ludwig and Miller (2007) identifies the causal effect of Head Start funding for lower-income families on employment and health outcomes in the 300 poorest counties in the US. The funding gave families access to health and social services, employment opportunities and nutrition for children. The funding from the Head Start program was allocated according to whether or not a county was above a poverty rate cut-off of 59%. By comparing counties that were just above the poverty rate cut-off with those who were just below, the authors are able to recover a treatment effect of the policy, since the two sets of counties should differ only in that one set received the funding and one did not. The authors need to consider multiple factors for this analysis. Firstly, how close to the threshold should counties be to be useful for the analysis – and will there be a sufficient number of counties both sides of the cut-off? One feature of this study is that counties cannot manipulate their poverty rates into receiving the funding: usually a key concern with an RDD approach. The authors conduct the RDD analysis, including significant checks for robustness, and find that Head Start substantially reduced child mortality rates as well as having positive effects on educational attainment.

For RDD to be valid, the following assumptions must be satisfied. Firstly, the eligibility variable should be close to continuous, meaning that there are many values of this variable which can be ordered appropriately (for example, income is continuous while marital status is not). Secondly, the policy cut-off must be clearly defined as a single point which determines eligibility, and it must also be unique to the policy of interest. For example, if individuals earning under \$45,000 per year, qualify for both income assistance and additional healthcare benefits, it is not possible to use RDD to separately estimate the effect of income assistance. Finally, the eligibility of an individual cannot be precisely manipulated either by themselves or another party. Without this constraint, people close to either side of the cut-off would no longer be there randomly and RDD could not be used. For example, there is the risk of people misreporting their income and thus manipulating their position in relation to the cut-off. However, there are statistical tests to detect this (Calonico et al., 2014)

RDD is a highly useful way of exploiting an existing policy threshold to capture treatment effects. It mimics some aspects of a randomised controlled trial (RCT), but treatment is not withheld from eligible individuals in order to create a valid control. However, the causal effect identified by an RCT is the average effect for the entire population of interest. In an RDD, by contrast, where the assignment to treatment is assured provided a person is eligible, the causal effect identified is 'conditional', in the sense that it is only true for those exactly at the threshold. This can be a positive or a negative aspect; for example, for learning about those at the margin of a desired threshold (such as a cut-off related to poverty status), RDD is effective. Otherwise, to understand the causal relationship for the entire population of interest, other methods may be more desirable.

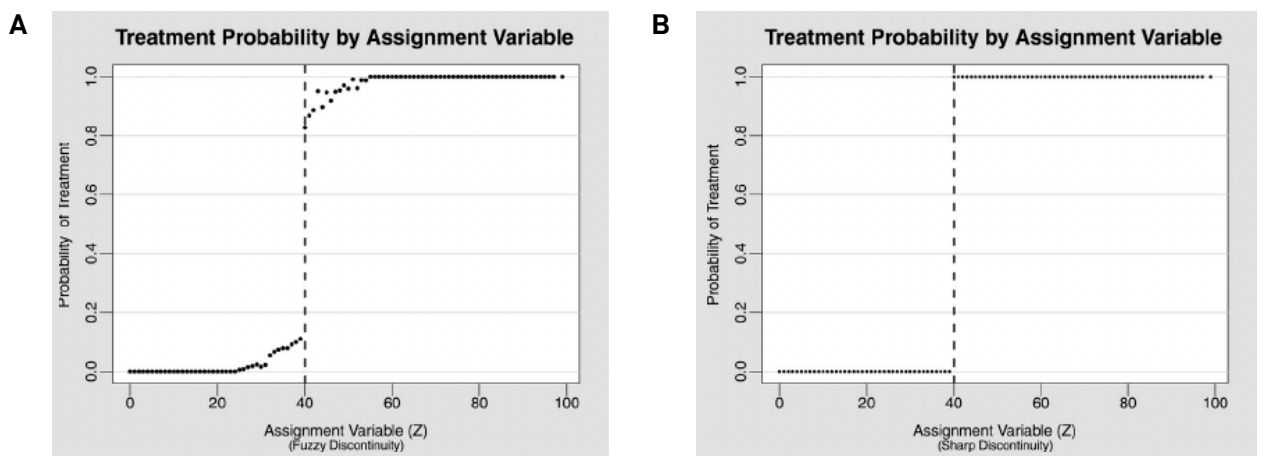
One drawback of the RDD approach is that it is generally opportunistically applied only in retrospective contexts where discontinuities are known to exist. That means researchers usually do not have control over the exact question to be addressed using this method. While it would be possible to implement RDD in a prospective research design, conventional prospective strategies (such as randomised trials) may be preferable.

Another important aspect to consider is how many individuals or observations are located close to the threshold. If these are small, the analysis can also draw on individuals further away from the threshold. Sensitivity tests are usually conducted to determine if the results are sensitive to the ‘bandwidth’ of data used in the analysis. Overall, there is still debate on how to choose between model specification in RDD (Kettlewell & Siminski, 2022).

Also key is the difference between “Sharp” and “Fuzzy” RDD methods. Sharp RDD is applied when ineligible individuals never receive the treatment and eligible individuals always receive the treatment. Fuzzy RDD is applied where it is possible for some ineligible individuals to receive the treatment and/or some eligible individuals to be untreated. The referenced study by Ludwig and Miller (2007) is an example of a Sharp RDD, since once a county is eligible, it does not have the option of declining access to the opportunities presented by the Head Start funding. If the threshold instead consisted of a subsidy to access a specific program, then the design would be Fuzzy. Some people receiving the subsidy might decide against accessing the program, and some unsubsidised individuals may still access the program.

In the case of a Fuzzy RDD, additional statistical techniques are needed (referred to as Instrumental Variable Regression). The causal impact of the policy can still be recovered. But this estimate is the average effect only for the subset of individuals at the threshold whose treatment was determined by which side of the threshold they lie. For example, the analysis cannot provide information on individuals who (hypothetically) would never take the treatment, or who would always take it regardless of the subsidy. These causal impacts are referred to as Local Average Treatment effects. The extent to which the estimates of the causal impacts lose generality when compared with a sharp RDD depends on the take-up of the treatment on either side of the cut-off. For further, detailed reading, see Cattaneo and Titiunik (2022).

Figure 4: Fuzzy versus Sharp RDD



Source: Moscoe et al., 2015

5.2 Difference-in-Differences (DD)

Without random assignment to treatment, finding a valid control group to compare with treated individuals can be difficult. Difference-in-Differences (DD) is an approach that seeks to address this issue, by comparing the outcome trends of the treated group with the outcome trends of a comparison group that does not necessarily have to be similar to the treated group. To establish causal impact, it is usually insufficient to compare outcomes for the treated group before and after treatment, as there may be other unobserved factors that have also changed over time. Similarly, it may not be sufficient to compare a treated group post-treatment with a group that did not receive the treatment, as there may be unobserved differences between the two groups which affect the outcome. DD combines these two comparisons, taking the difference between the treated group before and after the treatment and subtracting from it the difference between the control group (or the group that does not receive the treatment) before and after the treatment. This effectively compares the treated group's trend in outcome with the trend in outcome that would have happened had that group not received the treatment.

A famous paper by Card and Krueger (1994) employs this DD technique in an effort to estimate the impact of a rise in the minimum wage on unemployment in New Jersey. Traditional economic theory suggests that a higher minimum wage will increase unemployment for lower-skilled workers, as a price floor above the equilibrium wage induces an excess of labour supply. The authors investigate this empirically, by exploiting an increase in the minimum wage in New Jersey in 1992. This policy change was not mirrored in the neighbouring state of Pennsylvania, offering a potentially valid comparison group for a DD analysis.

Because the treatment is not randomly assigned, it is problematic to simply compare the two states. However, if employment trends in the two states had been the same in the absence of the policy change, then the DD approach could be used to identify the causal effect. Card and Krueger do exactly this, taking the difference between the trends in the two states and thereby eliminating any potential differences that could affect the outcome, leaving only the effect of the policy. They find that increasing the minimum wage actually increased employment.

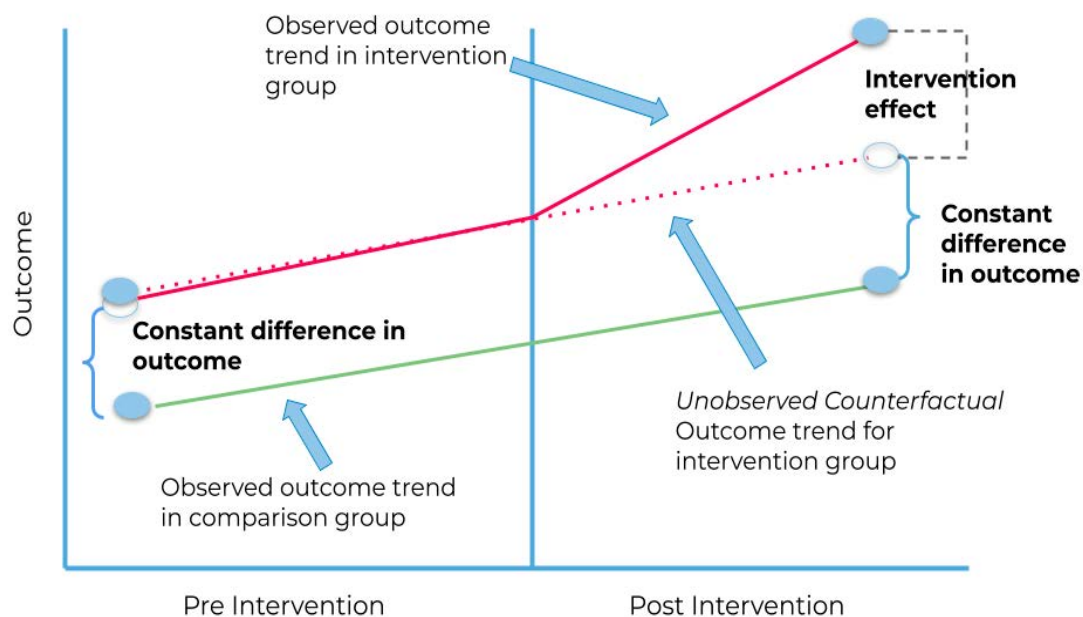
The DD estimate of the causal impact of a policy is as follows, where Y is the outcome of interest:

Equation 2: The Difference-In-Differences Estimate

$$\text{Causal Impact} = (Y_{Treated}^{After} - Y_{Treated}^{Before}) - (Y_{Control}^{After} - Y_{Control}^{Before})$$

To correctly identify the true causal impact, the DD approach requires a restrictive assumption. That is, the control group's trend constitutes what the trend of the treated group would have been without the treatment. Another way of expressing this assumption is that the unobserved characteristics associated with the outcome for both groups are constant over time. Therefore, with this approach, it is of less relevance that the control and treated groups have unobserved differences that may influence the outcome of interest, but it must be the case that such characteristics do not change over time. While the assumption is not testable, as it is not possible to observe what would have happened in the treatment group in the absence of the policy, there are ways to check the validity of this approach. Checking the trends in the outcome variable across the groups before the policy is implemented is one way. If the trends move in tandem, then it may be fair to assume that they would have continued to do so were it not for the intervention. Other methods include using a placebo treatment or treatment group and finding no effect supporting the validity of the DD analysis. Finding a significant result in the absence of an actual policy or treated group suggests there may be some underlying unobserved trend driving the outcome of interest and not the policy intervention.

Figure 5: The Diff-in-Diff estimator



Source: Population Health Methods

The DD approach can be a very useful alternative when there is no potential for a randomised treatment to create valid control groups, nor a policy threshold which randomly assigns treatment. It is particularly appealing because the comparison group does not have to be entirely similar in terms of observed differences. However, in order to identify the causal impact without any bias, the DD approach invokes the very restrictive and possibly unrealistic assumption that the unobserved differences between the groups are constant over time, and that the trends in the outcome are equal across groups apart from the trends due to the treatment. While it is possible to perform many additional exercises to enhance the validity and check the robustness of the DD design, there are multiple potential factors which can create bias or invalidate the method, and which may not be possible to account for. Specifically, if an event takes place at the same time as the policy intervention that has a different average impact on the groups, this will invalidate the results and prevent identification of the causal impact of interest. For further reading on DD, see Imbens and Wooldridge (2009), section 6.5.

5.3 Matching

All the methods discussed seek to solve the ‘missing counterfactual’ problem. Matching methods address this by selecting a control group which has the same observed characteristics as the treated group, on average. In most versions of matching estimators, each treated individual is matched with one or more untreated individuals. The match is made on observed characteristics such as age, income, gender and education.

The simplest version of matching is ‘exact matching’, by which each individual treated unit is matched with an untreated unit who is similar on all dimensions. In practice, it is usually difficult to find every treated unit with untreated units with the same combination of characteristics, and thus to find appropriate matches. This is known as the ‘curse of dimensionality’.

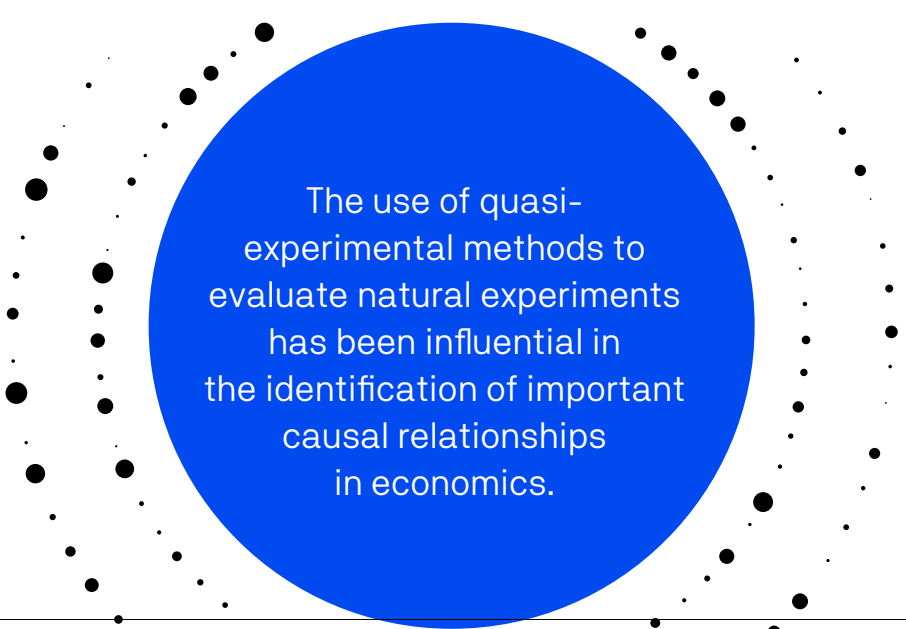
An alternative, and most common, approach is to avoid this issue, by instead estimating the ‘propensity score’, which is simply an estimate of the probability that an individual will receive a treatment, based on observed pre-treatment characteristics. This allows for the inclusion of as many relevant variables as is desired without risking the curse of dimensionality. The matching is then conducted by pairing every treated unit with an untreated unit according to the closest propensity score. However, in practice, certain problems can arise with this approach. For example, it can be problematic if a control unit is a good match for more than one treated unit, or if there exists no match for a treated unit.

Some conditions are required for implementation of the matching method. First, there needs to be substantial overlap across the propensity scores of treated and non-treated individuals. This means that, for at least some range of values for the treated group, there exists a similar range of values for the control group. The best-case scenario is that the range of values between treated and control completely overlap and are more or less equally numerous for each value. This will maximise the chances of finding a close match for every treated unit. However, this is rare. By definition, treated individuals will have higher average propensity scores, making it difficult to find a good match for a treated individual with an extremely high propensity score. The same applies to a treated individual matched with a non-treated individual with an extremely low propensity score. This restricts the range of values for which a treatment effect can be estimated; and if there is no overlap, then matching cannot be conducted.

An example relating to this method is de Brauw and Hoddinot (2011), who evaluate the effect of conditional cash transfers on school enrolment in Mexico. The authors exploit the fact that some families did not receive the necessary forms enabling them to receive the intervention, and they use this group as their control group. To effectively control for the differences between treated and non-treated families in this setting, the authors use a propensity score matching, estimated on characteristics such as a child's age and gender, the number of household members, total household expenditure, and literacy and indigenous status of household members. They find that the transfers significantly increased school enrolment, largely due to them being conditional. However, the authors also note that conditional cash transfers must be carefully designed to achieve the desired outcome.

The second assumption is far more restrictive, and constitutes a major drawback of the matching approach. In order for the estimated treatment effect to be valid, there must be no 'selection on unobservables'. Since the matching process can only be performed on observable characteristics, it is not possible to account for unobserved differences. If there are unobserved factors that are systematically different in treated and non-treated groups and which can affect the outcomes of interest, then the matching estimator is biased. This is a restrictive assumption, and it is also not testable. However, just how untenable selection on unobservables is varies from setting to setting, and usually requires substantial justification.

Matching is often a feasible method which only requires data on treated and untreated units at one point in time. It tends to be more convincing when a rich set of characteristics are observed in the data. However it cannot account for unobserved differences between groups, which usually leaves substantial doubt as to potential bias in the estimated treated effects. For further reading on matching and the associated theory, see Imbens and Wooldridge (2009), section 5.5.



The use of quasi-experimental methods to evaluate natural experiments has been influential in the identification of important causal relationships in economics.

5.4 Instrumental Variable (IV) Regression

A instrumental variable (IV) is a variable which influences the probability of “treatment” but has no direct relationship with the outcome variable. In other words, a valid IV may influence the outcome variable only because it effects the probability of treatment. IV methods are useful with experimental data when there is imperfect compliance. They can also be useful with non-experimental data, when there is some exogenous determinant of treatment. IV is usually implemented with regression, especially two-stage least-squares (2SLS) regressions. A simple version of this is summarised in the two equations below:

Equations 3: The Instrumental Variable Regression Equations

$$Y_i = X_i' \alpha + \beta D_i + \varepsilon_i$$

$$D_i = X_i' \gamma + \pi Z_i + \eta_i$$

- Y_i represents the outcome of interest for individual i
- X_i' are control variables
- D_i is the treatment of interest, usually equal to one if the treatment is taken
- Z_i is the instrumental variable which has a direct impact on D_i but no direct effect on Y_i
- β is the main parameter of interest, representing the treatment effect
- ε_i, η_i are error terms associated with each regression model
- The first equation is sometimes referred to as the structural equation, while the second is called the first stage relationship

Instrumental variables are particularly relevant in experimental settings where the random assignment does not correspond 1:1 with treatment (known as imperfect compliance). Consider Angrist et al. (2002), where the authors study a program in which vouchers for private school scholarships were allocated, through a lottery, to families in Colombia. Not all families who received a voucher used it (although overall take-up was high), and some children who did not win the lottery obtained scholarships from other sources. A direct comparison of lottery winners and losers would not identify the effect of the scholarship. Rather, it would estimate the self-explanatory ‘Intention-to-Treat’ (ITT). To recover the treatment effect, the authors use the randomly-assigned voucher as instrument for scholarship receipt. Adopting this approach, they estimated the effect of scholarships on a range of outcomes and behaviours.

Some important assumptions are required when considering an Instrumental Variable approach. First, the instrumental variable must have an effect on the treatment. This is crucial, since ‘weak’ instruments yield biased results and invalidate the analysis (Baker et al., 1996). Second, the instrumental variable must not directly impact the outcome variable, or indirectly impact the outcome through other variables. While the first assumption is testable, as the first-stage relationship is observed in the data, the second assumption is not testable, and for a valid IV analysis requires a strong argument about why it would hold. For further reading on Instrumental Variable Regression, see Mogstad and Torgovitsky (2018).

Adaptive designs
have the potential to
be very informative
for trialling cash
transfers.

6. Novel Methods for Evaluation

6.1 Adaptive Trials

Regular experiments typically seek to estimate a singular parameter (such as the average treatment effect). However, policy-makers may want to establish the best policy option out of several different options, or ‘arms’ – put simply, the policy which will make the greatest impact on the targeted outcome. This is where adaptive trials or designs can be more useful than a single trial. The main intuition of this approach is that each phase of the adaptive trial is used to modify the treatment in order to most efficiently assess which treatment is most appropriate. Typically associated with clinical and medical applications, adaptive trials are still relatively novel in other disciplines, such as economics (Kasy & Sautmann, 2021).

A crucial element of these designs is that any adaptations must be made according to a pre-planned outline and procedure. For example, one feature of the plan could be a focus on treatments that yield the largest overall impact. This is beneficial both to participants, as they receive progressively ‘better’ treatments (in relation to the outcome of interest), and to researchers, since time and resources are progressively allocated towards more effective treatments.

The manner in which the potentially many treatments within an adaptive trial are sequenced and explored is also key. This is sometimes expressed as the ‘exploitation-exploration trade-off’: researchers need to balance ‘exploiting’ (i.e. sticking with) treatment arms that appear to be doing well on existing information, with the potential benefit of ‘exploring’ (trailing) other versions of the treatment about which less is known. Algorithms can optimise this procedure and create an effective adaptive design (Kasy & Sautmann, 2021). Some pre-trial rules and effectiveness measures to consider include (but are not limited to): abandoning non-effective treatments, changing the allocation of participants to various treatments, and identifying the characteristics of participants most likely to benefit, then recruiting more of these types of individuals.

Adaptive designs have the potential to be very informative for trialling cash transfers. For example, there are a number of design options: size of transfer (e.g. \$5,000 compared with \$10,000), type of transfer (one-off lump payment or recurring payments over a specified period), and conditional/unconditional payments (for example, payments can be made conditional on school attendance requirements). There are also many potential eligibility criteria. The use of different types of cash transfers is explored in Roll et al. (2022).

Adaptive design has many significant advantages, for participants and for the organisation funding a trial. Firstly, it limits the possibility of over-funding treatment options with low returns, based on pre-specified criteria for treatment effectiveness. This means that participants are less likely to be assigned to an ineffective treatment option, and the funding is more efficiently allocated. This approach is well suited to delivering advice on the relative impact of each option in a menu of policy decisions, with high accuracy and equal levels of validity (since all options are evaluated in the same design and context).

However, the efficacy of adaptive design rests on several assumptions. First, the trial designer must be aware of the timeframe for obtaining results from each treatment arm. In some contexts this may not be restrictive, such as in clinical trials where short-term results are readily available (Pallman et al., 2018). In other context, key outcomes may be observed a lot later. Consider the effect of a cash transfer on a child's educational attainment. If such educational outcomes are measured several years later, this may undermine the usefulness of an adaptive design, due to the lag before obtaining feedback from each treatment option. Consider an evaluation of cash transfers to families with newborn children. By the time the first stage is evaluated, the families no longer have newborns, meaning they may not be as relevant to the evaluation question if their treatment status were to be changed.

Bayesian Adaptive Trials

Adaptive designs can be applied with a Bayesian approach instead of a classical 'frequentist' approach. A Bayesian approach more explicitly models a rational decision making process. It begins with a 'prior distribution' which summarises what is already known before the evaluation - for instance an assumed probability density for the possible effect size. Often the prior is 'flat', meaning that no prior beliefs are imposed. Either way, a Bayesian approach uses results as they emerge to update these priors, forming 'posterior' estimates of the true effect of each trialled option after each stage of the experiment. These posteriors can be used to adapt the trial as discussed in the previous section above.

With adaptive trials, the Bayesian approach can be an efficient way to use preliminary results to adapt the trial. For example, Broglio et al. (2022) show that the optimal trial length for obtaining the same conclusion is achieved more often using Bayesian designs.

To our knowledge, Bayesian Adaptive Trials have not been applied in the economics literature, or in related fields. This may be due to the often long lag time between treatment and outcome measurement, which limits the scope to adapt the trial according to interim results. However, it may also be due to lack of awareness of this technique. If planning a large scale experiment, incorporating of a Bayesian adaptive mechanism is worth considering.

6.2 Synthetic Control

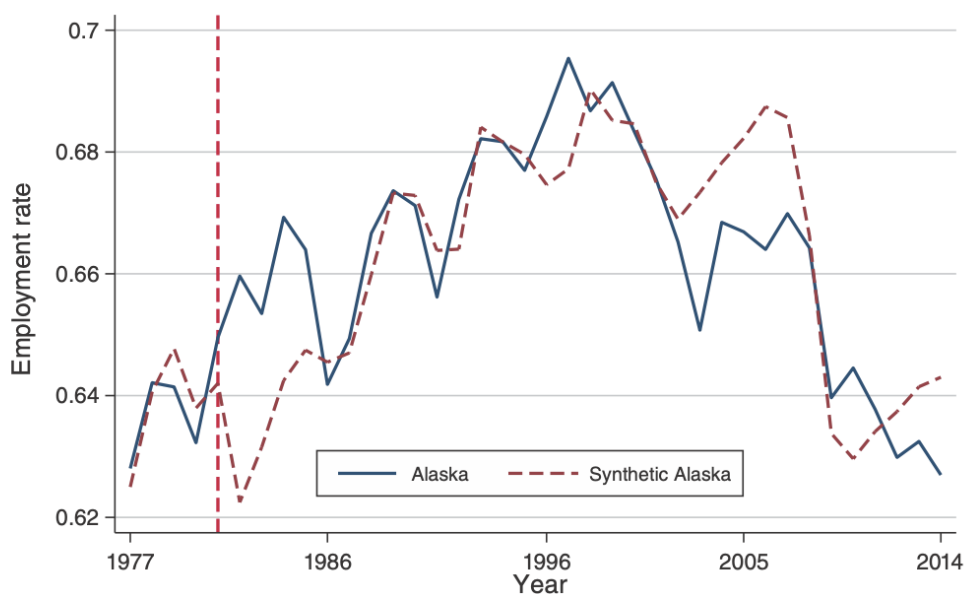
As discussed in the section on Difference-in-Differences (DD) (see Quasi-Experimental Methods), it is often difficult to find a comparison group which provides valid estimates of counterfactual outcomes. DD deals with this by assuming that outcomes from a non-treated group follow a similar trend to that of the treated group. However, this assumption may not be plausible in many contexts.

Abadie et al. (2010) introduced a new approach for constructing a valid control group without imposing strong restrictions on its relationship with the treated group, and to directly apply the framework to the kind of setting described above. This method is known as synthetic control. Instead of trying to find one single valid control group, the method creates a ‘synthetic’ comparison group based on an optimally weighted combination of multiple potential control units. The intuition is that it is difficult to find a perfect comparison in the absence of a randomised controlled trial, and while there might be many potentially good candidates for this comparison group, they might all have certain flaws that compromise them. Synthetic control combines these units, assigning each a specific weighting.

As in DD, Synthetic Control requires both pre-treatment and post-treatment observations of the outcome variable(s) of interest, for both the treated and untreated ‘control’ units. The optimal combination of candidate control units is determined through a data-driven process that finds a weighted average of candidates which minimises the observed difference in characteristics between the mix of control units and the treated unit. This weighting scheme is then used to generate counterfactual post-treatment outcomes for the treated group.

Abadie et al. (2010) use this framework to evaluate the impact of anti-smoking legislation passed in California in 1988 on per capita smoking rates. They constructed a comparison group synthetically using a set of candidate control states. Their synthetic mix included Utah (33%), Nevada (23%), Montana (19%) and Colorado (16%).

Figure 6: Effect of a universal cash transfer on employment in Alaska



Source: Jones and Marninescu (2022)

A recent paper by Jones and Marinescu (2022) evaluates the effect of unconditional and universal cash transfers to Alaskan residents, starting in 1982, on various labour market indicators including employment, labour force participation and part-time employment. As this transfer was universal, no immediate comparison group existed that could function as an acceptable control. The authors therefore implemented the synthetic control method in order to construct a valid comparison group. Their ‘Synthetic Alaska’ is largely made up of Utah (43%), Wyoming (34%) and Washington (9%). This study was motivated by the concern that a universal income would discourage recipients from supplying more labour – since, in theory, individuals receiving extra income can afford to work less. The authors find the opposite (see Figure 4): the universal and unconditional cash transfer had no significant long-term effect on employment, providing evidence in support of a Universal Basic Income scheme (Jones & Marinescu 2022).

6.3 Machine Learning

Machine Learning (ML) techniques are mainly used for prediction/categorisation, rather than addressing causal questions of program impact. Nevertheless, ML can be useful for addressing causal/evaluation questions, when combined with standard causal inference techniques. This is discussed by Athey and Imbens (2017, 2019) and by Mullainathan and Spiess (2017). Where a researcher adopts an observational data analysis technique on the assumption of ‘selection on observables’, and where there are many potential control variables, ML can assist with covariate selection (Athey & Imbens, 2017, 2019). ML can also be useful when analysing the heterogeneity of treatment effects, as it can help to identify appropriate sub-groups by which to stratify the sample (Athey & Imbens, 2017, 2019). Lastly, ML can assist with Instrumental Variable analysis, where there are many instrumental variables to potentially include. The ‘first stage’ of a 2SLS (Two-Stage Least Squares) Instrumental Variable analysis can be seen as a prediction exercise, thereby prompting the standard justification for employing ML (Mullainathan & Spiess, 2017).

However, the emergence of Machine Learning is not regarded as a major development in the technology of impact evaluation methods. In particular, ML does not resolve the major challenges of impact evaluation with non-experimental data, such as the problem of ‘selection on unobservables’ and associated bias.

6.4 Event Studies

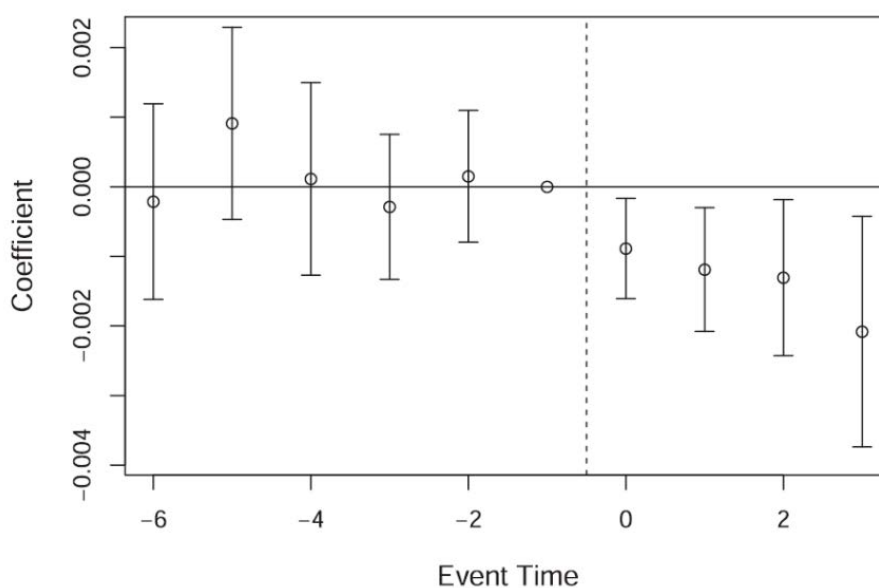
As previously mentioned, the parallel trends assumption is critically important for the Difference-in-Differences (DD) design (see Quasi-Experimental Methods). Event Study plots can be used to gauge the validity of this assumption, and they are especially useful if the treatment is introduced to different units with staggered timing.

Event study plots show whether pre-treatment outcomes follow the same trend for treated and untreated groups. These are typically shown as estimated ‘lead effects’ – that is, as estimate effects of the treatment in the time periods prior to implementation (clearly these estimates should be close to zero, if the groups are indeed following common pre-treatment trends). They also show estimated effects for several periods after implementation – thereby visualising any potential dynamic effects of the treatment over time.

If the treatment is implemented at only one point in time, an event study plot simply includes time on the horizontal axis. However, if the timing of the treatment is staggered, the horizontal axis can be presented as time relative to implementation of treatment.

For example, Miller et al. (2021) use event studies to evaluate the impact of Medicaid enrolment on mortality. First, the authors demonstrated the effects of the Medicaid expansion on eligibility and coverage, using event study plots. The key plot is an event-study showing the estimated treatment effects in several periods before and after implementation. This plot is reproduced below. It shows no significant effects in periods before treatment, providing support for the parallel trends assumption. In contrast it shows significant negative effects in all post-intervention periods, with a suggestion of larger effects over time.

Figure 7: Example Event Study Plot, Effect of Medicaid Expansion on Mortality



Source: Miller et al. (2021)

6.5 Combining RCT and structural estimations

So far, this report has discussed methods for estimating the impacts of existing programs or policies. This is sometimes referred to as ‘reduced-form’ evaluation. A second evaluation approach, often termed structural, entails a fully specified behavioural model. Structural models are often used to evaluate existing policies and perform counterfactual policy experiments, such as the evaluation of new hypothetical policies. A recent literature seeks to find “the best of both worlds” by combining RCTs with structural modelling. This combination of structural modelling and RCT can increase the credibility of inference in various ways. For example, the RCT can be leveraged for model validation and selection of the structural modelling, using either the treatment or the control group as a ‘holdout’ sample for performing out-of-sample model fit tests. As another example, researchers can use the variation induced by the treatment as an additional source of variation for identifying and estimating model parameters and improving precision. The structural models can then be used to estimate the impacts of policies that are different from the ones implemented and evaluated through the RCT. For further reading, see Todd and Wolpin (2002).



7. Integrating Quantitative and Qualitative Methods

Qualitative methods are sometimes used as a valuable complement to quantitative impact evaluation. Mixed methods, combining quantitative and qualitative data, can be used to form hypotheses and validate results, both during the preparation stage of the program and during/after the implementation stage. Qualitative methods are a large, separate field. Approaches include focus groups and extended interviews with selected informants, life histories, case studies and observational assessments (Rao & Woolcock, 2003). Quantitative and qualitative methods sometimes overlap in the sense that the former can include some numerical data, and the latter can include some open-ended questions.

Qualitative approaches are not intended to be statistically representative, nor are they generalisable. However, they are often used to provide context to the quantitative results, and to generate deeper understandings of particular themes and experiences. The various mixed method approaches include:

- **Convergent parallel:** obtaining early results on a program's implementation from both quantitative and qualitative data collected.
- **Explanatory sequential:** explaining outliers or interesting patterns in the quantitative data, with a deep dive into case-by-case qualitative data analysis of these particular results.
- **Exploratory sequential:** interviewing key beneficiaries in order to form hypotheses, correct the design alternatives, specify research questions or choose the survey and sample design.

A mixed method approach helps to verify the validity and reliability of quantitative data. It can also help with comparing and collecting data sources, and with formulating the theory underlying the impact evaluation (Bloomquist, 2003).

While quantitative methods address the 'what' and 'where' questions, qualitative methods may help with the 'why' and 'how'. Determining which interventions were apparently successful is crucial, but it is also important to know why or how the success or failure occurred (Prowse, 2007).

8. Australian Data Landscape

8.1 NSW Human Services Dataset

8.1.1 Overview

The NSW Human Services Dataset (NSW HSDS) was created in the context of the 2020 'Their Futures Matter' reform, a government initiative aimed at improving outcomes for vulnerable children and their families (Audit Office of NSW, 2020). The NSW government sought detailed and highly informative data, enabling it to identify and target the most vulnerable groups in the state, as well as to assess effectiveness of programs and interventions.

The NSW HSDS is sourced from the administrative records of a variety of state government departments and ministries including the Department of Communities and Justice (DCJ), the Ministry of Health and the Department of Education. The resulting dataset covers individuals born after 1 January 1990 until 2017, and contains over 7 million records on the primary cohort of children (DCJ, 2021).

An important element of the NSW HSDS is the sensitivity of the information within the data. Records originating in government agencies and departments such as the NSW Registry of Births, Deaths and Marriages (RBDM) contain individual information which in its raw form would violate the Privacy and Personal Information Protection Act 1998. Hence, the DCJ (the data custodian) is obliged to ensure the data has been completely de-identified. This has implications for data access.

8.1.2 Identifiers

Key identifiers required for the analysis considered in this report are contained within the RBDM sub-dataset component of the HSDS. This includes date of birth and registration date, socio-economic status, gender, geographical area, family composition, marital status of parents, and Indigenous status.

8.1.3 Child Outcomes

As the focus of the NSW HSDS was originally on delivering better outcomes for disadvantaged children, it contains many indicators of child well-being. Sub-datasets such as Out-of-Home Care Placements and Child Protection Reports contain indicators of out-of-home placements, including the reasons for such placements, and reports of abuse, including abuse type (for example, alcohol/drug or domestic abuse), as well as risk level assessments. These may be useful outcomes for investigating the impacts of a cash transfer for vulnerable children.

Similarly, sub-datasets from agencies such as the NSW Bureau of Crime Statistics and Research which can provide other types of child well-being indicators. They contain criminal charge reports, victim impact reports and juvenile offence reports. It is possible to ascertain whether a child's family member was involved in a criminal incident, or whether a child was a victim of or witness to a crime. Information on the type of charge a child has faced, its severity and the child's involvement in an incident may be relevant. These sub-datasets also contain information on offence type and history, and family circumstances.

Other outcomes found in data sources pertain to early childhood development. The Perinatal Data Collection contains data on the health of the mother and child during and just after pregnancy, while the Best Start Kindergarten Assessment has information relating to early childhood development domains, such as writing, numeracy, comprehension and speech.

The Student Details dataset includes data on various student and family characteristics, such as remoteness, school type, parents' education level and language spoken at home. This dataset also includes diverse educational information and outcomes, including school mobility, frequency of moving school, NAPLAN results and performance in the Higher School Certificate, including course enrolment, and grades in each course.

8.1.4 Parental Indicators

Whilst the focus of the HSDS is to provide insights on vulnerable children, it also includes extensive data on parental indicators of well-being, specifically in relation to health. The Admitted Patient Data Collection contains data on hospital admissions; similar data are available for emergency department visits. There are also data on drug and alcohol treatment services, and enrolments into programs associated with these services. A cash transfer program might be hypothesised to reduce the likelihood of parents needing to use such services.

Other administrative sources include the Vulnerable data project from the NSW Office of State Revenue (now Revenue NSW), which focuses on the enforcement of fines, and the state Department of Industry's data on training programs funded by NSW. It should be noted that although an overview provided by the DCJ (NSW Department of Communities and Justice, 2021) refers to a link between the NSW HSDS and Commonwealth welfare payments, income and taxation, no reference to these appears in the data item list (FACSIAR, 2021). This issue and its implications are discussed in greater detail in the next section on the MADIP dataset.

8.1.5 Barriers to Accessing the NSW HSDS

Due to the highly personal and sensitive nature of the data in the NSW HSDS, access is tightly guarded by the data custodian, the DCJ. The guidelines for accessing the dataset (FACSIAR, 2021) include requirements to demonstrate technical ability in data analysis, sign legally binding privacy and confidentiality agreements, supply both a national police check and a working with children check, and be willing and able to undertake training in use of the data and in privacy, confidentiality and security practices. These requirements reflect the seriousness with which the data custodians view any security or privacy breach; in terms of accessing this data, the necessary time investment must be a consideration.

The approval process outlined in FACSIAR (2021) is extensive, and likely to take a significant time. Applicants must first send a detailed proposal to the HSDS governance team to review. The governance team then provides feedback, and may request additional information. If this stage is successfully completed, the Human Services Dataset Governance Advisory Committee next considers whether the proposal aligns with HSDS values, whether the data are suitable for the proposal, the potential risks associated with the proposal, and the technical feasibility. Finally, the data custodian can approve access, although further multiple checks will be made, including of the project's output.

In conclusion, while the HSDS data provide very detailed information which links individuals across many different variables and agencies, there is a significant time cost associated with gaining access to it. This may lead to significant lags between the time of data collection and time of data analysis. There may also be issues of potential sample loss if children move interstate. Such issues are not unusual when seeking to access administrative datasets.

8.2 Linked Administrative Datasets in Other State and Territory Jurisdictions

Linked data projects are also found in other states:

The Centre for Victoria Data Linkage is a key data asset containing health and human services data, combined with other key administrative and clinical sources (Victorian Department of Health, 2022).

The South Australian and Northern Territory dataset SA-NT DataLink is similar, with access to major administrative sources, similar to those covered by the HSDS (SA-NT DataLink, 2022). BEBOLD is a platform for academic and research partnerships which is compiled through a collaboration with SA-NT DataLink.³

Western Australian has the Development Pathways Project (DPP) which also enables linkage of de-identified data from Western Australian government departments and agencies or relevance to developmental outcomes for children and youth.⁴

We have not explored the feasibility or logistics of accessing data from these sources.

3. See <https://health.adelaide.edu.au/betterstart/bebold>

4. See <https://www.telethonkids.org.au/projects/developmental-pathways-project/>

8.3 Multi-Agency Data Integration Project (MADIP)

MADIP was developed with the goal of providing a holistic, population-level data asset capable of providing advanced insights and facilitating research agendas (Australian Bureau of Statistics (ABS), 2022). With rich data at the individual level spanning many years, MADIP's key areas include health, education, government payments, income and taxation, employment, and population demographics (Wright, 2021). As it was designed specifically for both program evaluation and addressing policy questions, MADIP is a potentially useful resource for some of the project ideas discussed in this report. Since it contains private and sensitive information, the data have been de-identified (ABS, 2022).

A number of government agencies contribute to the pooled data asset at the federal level. They include the ABS, Australian Taxation Office (ATO), Department of Health, Department of Social Services and Department of Education. Data from these agencies, and from others that contribute to MADIP, are integrated in such a way that individuals in each dataset can be indirectly linked across different outcome variables. This structure is known as the 'Person linkage spine'. The total number of individuals observed in MADIP is very high, comprising a very large proportion of the population: over 35 million individuals (Wright, 2021).

8.3.1 Key Datasets

The largest component of the MADIP core databases is the Medicare Consumer Directory, which has detailed information on all individuals who were enrolled in Medicare between 2006 to 2020. This is the largest administrative component of the dataset, and it allows for accurate linking of individuals across the entire data asset. Other important datasets are the Australian Census Longitudinal Dataset (ACL) and the Census of Population and Housing, which enable the classification of families and individuals into associated strata of interest. Key characteristics covered in these datasets include household composition, educational attainment, geographical location (at LGA level), Indigenous status, marital status and household income.

8.3.2 Health

In terms of indicators of health and well-being for both parents and children, the MADIP datasets are extensive. The National Health Survey provides information on a broad range of health and well-being characteristics, such as medical conditions, health and lifestyle risk factors, mental health and use of health services. However, it is limited to a randomly selected representative sample of the population. The Pharmaceutical Benefits Scheme (PBS) provides data on individuals accessing services and medications covered by the PBS.

8.3.3 Education and Childhood indicators

The Australian Early Development Census (AEDC) outcome measures include physical health and well-being, social competence, emotional maturity, language and cognitive skills, and communication skills (Australian Early Development Census, 2022). Other important educational data items include government training programs and higher education outcomes, including school enrolment and completion. The AEDC has been conducted nationwide every three years since 2009.

8.3.4 Income and Social Security

Other major components of the MADIP data asset are the DOMINO (Data Over Multiple Individual Occurrences) CAD (Centrelink Administrative Data) from the federal Department of Social Services, and the tax data collated from personal income tax returns. The personal income tax dataset contains comprehensive data on taxpayers' income from each sources, including salaries and wages. The DOMINO database is also potentially helpful, as it identifies the recipients of various social security payments and transfers (including the Coronavirus Supplement).

8.3.5 Barriers to Accessing MADIP

As with the NSW HSDS, access to MADIP data is subject to approval, following a lengthy process. Access is facilitated through the ABS DataLab, which enables users to view and analyse detailed microdata (such as MADIP). The ABS DataLab also checks any analytical output before it is permitted to leave the secure ABS environment. Use of the ABS DataLab is subject to an annual fee starting at \$2,200; that grants standard virtual machine access for up to 5 users.

The request process for MADIP starts with a discussion with the ABS about the desired project; the ABS provides feedback on data suitability, as well as anticipated costs and any training required. The project proposal is then submitted, and formally reviewed and quoted by the ABS. If the proposal is accepted, the ABS approves access and organises researcher onboarding, including ABS DataLab training. Outputs from the research project may be vetted and audited in case of deviations from the original proposal. The ABS aims to approve or deny access within one month of receiving a completed project proposal, but notes that this depends on the complexity of the proposal (ABS, 2022).

8.3.6 Survey Data

The NSW HSDS and MADIP are administrative datasets sourced from government agencies. In addition, data are available from many surveys carried out among representative samples of the population. However, it is important to note that administrative datasets have multiple advantages over sample surveys. The sheer number of observations in administrative datasets typically implies much greater statistical power than sample surveys, with implications for the precision of estimates. Furthermore, survey data are more likely than large administrative datasets to be affected by issues such as attrition and non-response. On the other hand, sample surveys typically include a broader set of outcome variables.

8.3.7 Household Income and Labour Dynamics in Australia (HILDA)

The HILDA household panel survey is run by the Melbourne Institute of Applied Economic and Social Research. Its first wave was implemented in 2001. The aim of the survey design was to collect a significant random sample of the Australian population and follow those people each year, indefinitely, to observe dynamics in the domains of household and family relationships, income and employment, and health and education. On average, each wave observes around 17,000 individuals and 7,000 households. HILDA includes a very broad range of relevant outcome variables. One of its limitations is that it does not capture the precise timing of receipt of a government payment, which is relevant to a potential evaluation of the Coronavirus Supplement (see Section 10). There is also concern about

lower interview rates and observations for certain groups: young people, individuals born in a non-English-speaking country, Aboriginal and Torres Strait Islanders, unemployed people and low-skilled occupation workers (Summerfield et al., 2021).

8.3.8 Longitudinal Study of Australian Children (LSAC)

Implemented through the federal Department of Social Services and the Australian Institute of Family Studies, the LSAC follows cohorts of children over time and tracks their development and life course trajectory. The aim of LSAC is to uncover and analyse opportunities for policy interventions during early childhood development in order to improve outcomes for the children as they develop. The survey, which started in 2004, collects data from two cohorts: children initially aged 0-1 and those aged 4-5. Data collection is conducted every two years, with an initial sample size of 5,000 (Department of Social Services, 2022). This survey collects an extensive set of early childhood development indicators, such as general physical development, emotional and behavioural development, and social capital accumulation. However, there are some drawbacks when it comes to utilising these data. The relatively small sample limits statistical power. It is also open to question whether these samples are the appropriate cohorts for studying the impact of specific reforms.

8.4 Bank Transaction Data

The credit bureau illion provides access to the illion dataset, which contains billions of bank transactions from millions of individual bank accounts. This is the result of illion consensually collecting and organising data from credit applications (Elias, 2022). From this data, it is possible to observe an individual's gender, home state, income and social security status, with a high degree of accuracy. The key feature of the data is the expenditure patterns of individuals, showing what people are spending their money on, along with broad estimates of how much is being spent in each product category, and frequency of spending. A drawback of this dataset is that individual characteristics that can be inferred are quite limited (i.e., family composition, measures for disadvantage).

As previously mentioned, the CBA has signed agreements with some university partners, including UTS, to collaborate on research projects of interest to researchers and the bank. It seems that there is no publicly available documentation on the CBA's database, or the conditions of access. However the bank has recently briefed the UTS project team, and this database appears to be a very promising source in the present context. It is probable that the data can reliably identify households in the target population, and that it contains information about receipt of the Coronavirus Supplement (and other payments), as well as a rich set of variables capturing types of expenditure and saving, and collective family decision-making. A limitation of these data is that they may only be accessible for 2-3 years after collection.

9. Baby Bonus

Natural Experiments

9.1 Information and Background

Previously referred to as the Maternity Payment (which replaced the long-standing Maternity Allowance), the Australian government payment to new parents was reformed in 2004 and renamed the 'Baby Bonus', with the aim of boosting fertility rates. This policy reform constituted a natural experiment where eligibility depends on a child's precise date of birth. Importantly, the Baby Bonus was universally accessible and unconditional. Parents of a baby born within the eligible timeframe could access the payment and use it however they pleased. As a result, the Baby Bonus offers a potentially fruitful environment for evaluating the impact of an unconditional cash transfer.

9.2 Policy Changes

The main changes made in relation to the Baby Bonus are summarised in Table 1. Australia has a long-standing history of maternity-based payments (Daniels, 2009), dating back to the Maternity Allowance Act of 1912. In 1978, the Maternity Allowance was abolished, but then re-introduced in 1996. This saw the parents of newborns receive \$840 (all amounts are expressed in dollar terms of the relevant time period). The payment was given to families who met the criteria of an income and assets test, and later was also partially conditional on a child being immunised.

Table 1: Summary of key policy changes. Sources: Deutscher and Breunig (2018), Daniels (2009)

Date	Amount of Payment	Frequency of Payment	Conditions for Payment
1 February 1996	\$840	Lump Sum	Family Income and assets must be lower than a certain threshold (Daniels, 2009)
1 July 2004	\$3,000	Lump Sum	Unconditional
1 July 2006	\$4,000	Lump Sum	Unconditional
1 January 2009	\$5,000	Fortnightly Payments	Conditional on income being less than \$75,000
1 March 2014	\$500 + \$1,500	\$500 lump sum payment, fortnightly payments of \$230	Conditional on being eligible for Family Tax Benefit A

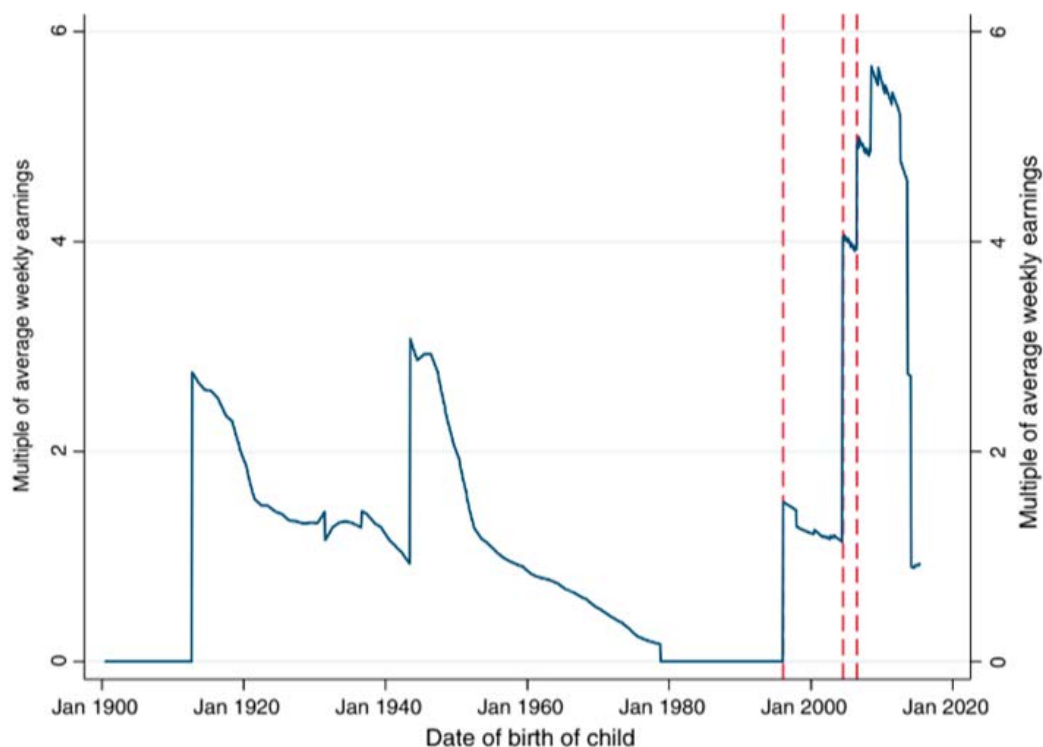
As these payments were indexed to the cost of living, the Maternity Allowance increased gradually between 1996 and 2004 (Department of Social Services, 2021). On the eve of the Baby Bonus being introduced, an eligible family receiving the Maternity Allowance was being paid \$1,054 (as of 20 March 2004). The Baby Bonus lump sum was significantly higher, at \$3,000. Later changes comprised the transition from a lump sum to a fortnightly payment in 2009, then a steep reduction in payment when the Baby Bonus was abolished in 2014 and replaced by the Newborn Supplement.

9.3 Earlier Work on the Baby Bonus

One aspect of the natural experiment presented by the introduction of the Baby Bonus was evaluated in Deutscher and Breunig (2018). Exploiting the threshold that was created by the design of the policy, the authors compared children born after 1 July 2004 (the treated group) with valid comparison groups. They were interested in the effect of the additional, unconditional \$2,000 cash transfer on the early educational outcomes of the treated children.

This additional one-off cash transfer is equivalent to 4 times the average weekly earnings in 2004 (See Figure 6). Notably, the payment was of a similar size, coverage and accessibility to potential transfers considered in this report.

Figure 8: Historic size of cash transfers for childbirth. Source: Deutscher and Breunig (2018)



There are several potential approaches to analysing this natural experiment. Given the nature of the threshold or cut-off for the policy, one option is to conduct an RDD analysis. However, Gans and Leigh (2009) highlight complications of applying such a methodology to precisely this context. They show that prospective parents significantly altered the timing of their child's birth in order to be eligible for this payment. They estimate that around 1,000 births were delayed in order to receive the payment. This poses a significant threat to the validity of an RDD approach, since individuals can change their treatment status and hence the assignment to treatment is not longer locally random.

One possible solution is to apply a 'Donut RDD', where observations are excluded from the analysis from the DOB region where they may have manipulated their treatment status, as per Barreca et al. (2011).

Deutscher and Breunig instead implement a Difference-in-Differences (DD) design. This design compares the difference in outcomes for children born in 2004 in the months before versus after implementation of the policy, to the corresponding difference in outcomes for children born in the same months, but in adjacent years (2003 or 2005). This approach accounts for the fact that children born in certain times of the year may be different to children born at other times of the year. It accounts for this under the assumption that those differences are the same in adjacent years.

The outcome variable considered by Breunig and Deutscher is the NAPLAN scores for children in Year 3, for which the authors find zero effect of the Baby Bonus payment. They also evaluate the effect for families with lower socio-economic status (SES), and find suggestive evidence that disadvantaged families benefited more from the policy. While this analysis does not find evidence that the Baby Bonus had an impact, the authors only consider NAPLAN scores as an outcome variable, there are other outcomes to consider.

In ongoing work, de Gendre et al. (2021) also study the effect of the Baby Bonus, using RDD analysis around the 2004 reform. They draw on linked administrative data from South Australia, and focus on interactions with the medical system. They find that the Baby Bonus "reduced emergency department presentations and inpatient services utilization, mostly for respiratory problems in the first two years of life", with effects concentrated amongst disadvantaged families. The mechanisms for this effect are not yet fully understood.

9.4 Extension to a Broader Set of Outcomes – the 2004 Reform

Building on the contributions of de Gendre et al. (2021), and Deutscher and Breunig (2018), we suggest considering an extension of their analysis to a broader set of outcomes. This recommendation is largely due to the advantages associated with the Baby Bonus natural experiment; namely, that eligibility is dependent on date of birth, and access to the policy is universal. This has two important consequences for any potential analysis. Firstly, date of birth is included in most administrative datasets (eliminating the need to match outcomes databases to payment receipt records). Secondly, since all families with an eligible newborn have access to the policy, the results can be reflective of the broader population, and they also allow for sub-group analysis (for example, low-SES families, Indigenous Australians).

Such an analysis may consider outcomes for parents and children, and may consider short-term or long-term impact. The parental outcomes included in the administrative datasets considered by this report (see Australian Data Landscape) include health and well-being indicators, employment status, social security payments and income, incidents related to drug and alcohol abuse, and domestic violence. Potential effects on parental outcomes may, of course, affect short- and long-term child outcomes via a variety of mechanisms, which are worthy of further exploration. The de Gendre et al. (2021) analysis could also be extended to other states and territories, as well as later periods.

Some influential studies suggest that the benefits of some interventions are not realised immediately, in early childhood, but have significant effects later in life, as found by the Perry Preschool Project (Heckman & Karapakula, 2019). One positive aspect of the Baby Bonus natural experiment is that children born near the July 2004 threshold are now approaching their 19th birthdays. Administrative data includes observations on these individuals over the course of their whole pre-adult lives. It may also be possible to run a survey or experiment with individuals born either side of the threshold, and examine factors not observed in great detail in the data, such as non-cognitive outcomes.

9.5 Lump Sum vs Recurring Payments – the 2009 Reform

Another noteworthy policy change in relation to the Baby Bonus occurred on 1 January 2009. The payment was changed from a lump sum to a recurring fortnightly payment over 13 fortnights. The total amount paid was unchanged at \$5,000, having increased slightly six months earlier (1 July 2008). The varying effects of recurring payment schemes as opposed to lump sum transfers have been studied in Roll et al. (2022). The authors find that recipients of lump sum transfers are more likely to improve their household balance sheet (by paying off household or student debt or saving for emergencies and retirement, for instance) than recipients of recurring payments.

The mechanisms by which a cash transfer affects outcomes of interest may differ depending on whether the transfer is a lump sum or recurring payments. Therefore, a potential study on the impact of the 2009 change to the Baby Bonus could be illuminating.

Another, important change made at the same time was to means-test the payment according to income. Only families with a newborn baby and earning under \$75,000 a year were eligible for the payment. This adds a complication to the analysis, since it limits the population of interest, and must be taken into account when designing the study. For a valid analysis of this policy change, the estimation sample (including comparison groups) has to be restricted to lower-income families.

Given the timing of this policy change, the children of targeted families are of course younger than those born around the original 2004 Baby Bonus eligibility threshold. It is therefore more appropriate to evaluate early development outcomes, for which there are multiple options; see Section 8 (Australian Data Landscape).

The data required is available within the MADIP administrative dataset (see Australian Landscape).

9.6 Reduction in Payment – The 2014 Reform

On 1 March 2014, the Baby Bonus was abolished and replaced with the Newborn Supplement and Newborn Upfront Payment. The total payment was reduced from \$5,000 to about \$2,000 for a first-born child. RDD analysis around this date of birth threshold would reveal the impact of reducing the cash transfer and thus demonstrate how valuable the initial transfer was in terms of the various outcomes of interest. It is important to note that, on each side of this threshold (those receiving the original Baby Bonus and those now receiving the Newborn Supplement), both payments are means-tested and involve similar complications to those mentioned above, in relation to sample restriction.

The Newborn Supplement is administered through the Family Tax Benefit (FTB), which has implications for who can access the payment, compared with those who were able to access the Baby Bonus. For a study on this policy change, an investigation of other coinciding social security payment changes would be required.

The value of implementing such a study also needs to be weighed against the question of whether any data sources contain accurate date of birth data, as well as the need to restrict the sample to the sub-population of interest. Key data sources to be considered are the NSW HSDS dataset and the MADIP dataset, discussed in Section 8.

10. Coronavirus Supplement

Natural Experiment

10.1 Information and Background

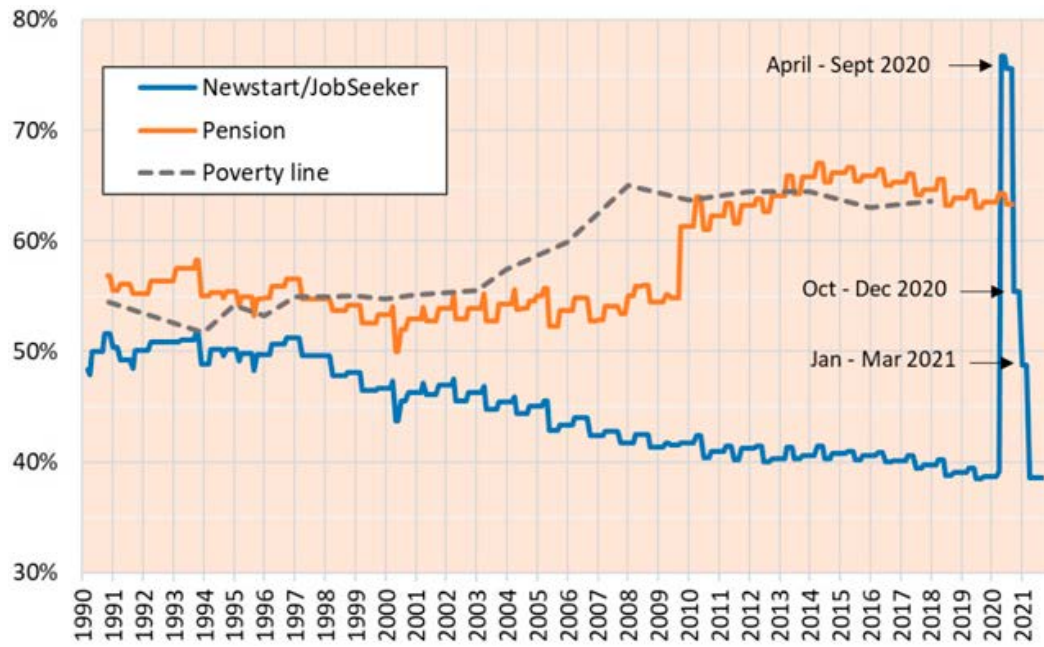
Part of an array of economic stimulus and assistance policies designed to mitigate the impact of the COVID-19 pandemic, the Coronavirus Supplement was announced on 22 March 2020. The fortnightly payment was administered through the social security system and was available to recipients of JobSeeker, Youth allowance, Austudy Parenting Payment, amongst other payments.

Date	Fortnightly Amount	Accessibility dates
22 March 2020	\$550	27 April 2020 – 24 September 2020
6 October 2020	\$250	25 September 2020 – 31 December 2020
10 November 2020	\$150	1 January 2021 – 31 March 2021

The first wave of Coronavirus Supplement payments saw eligible Australians receive \$550 per fortnight. Figure 6, from an inquiry into the Social Services Amendment Bill by Bradbury and Whiteford, shows how the Supplement bolstered existing payments to the extent of almost matching the minimum wage (Ferlitsch, 2022).

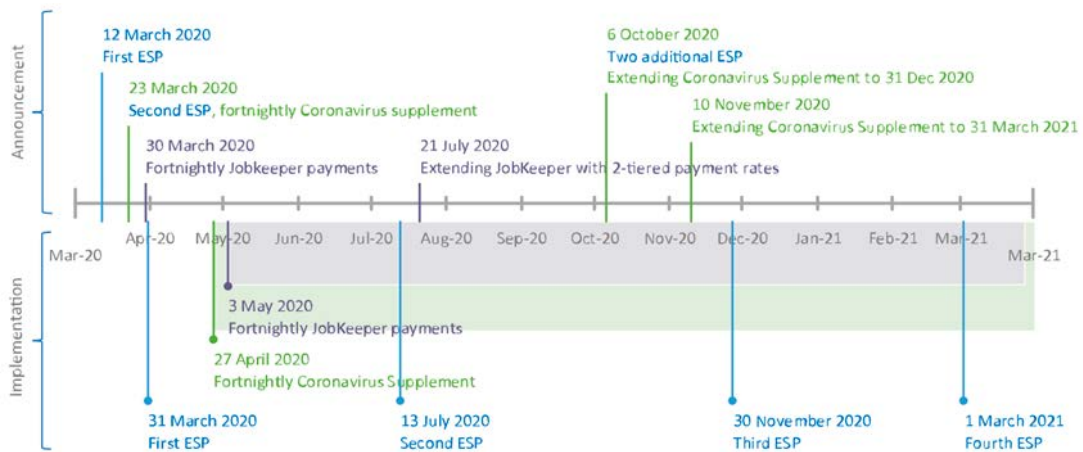
As mentioned above, other stimulatory and supportive policies ran concurrently with the Coronavirus Supplement. Identifying the impact of one intervention requires accounting for other interventions that may confound the estimate (see section on Quasi-Experimental Methods). Figure 7, taken from the RBA's analysis of COVID-19 payments, illustrates other interventions (JobSeeker and Economic Support Payment) which ran in tandem with the Coronavirus Supplement.

Figure 9: Social Security Payments as a proportion of the minimum wage.



Source: Ferlitsch (2022)

Figure 10: Timeline of Economic Payments in Response to the Pandemic.



Source: RBA 2021

10.2 Potential Framework for Analysis

In the case of the Coronavirus Supplement, it is difficult to identify an untreated group who could serve as a credible comparison group from which to infer counterfactual outcomes for the treated group. However, if the hypothesised effect is likely to be quite large, this challenge may prove less prohibitive.

With this challenge of finding a valid comparison group, and with the dynamic nature of the policy, it may be that the effects most feasible to identify are those that are immediate and large effects. To confidently identify smaller or longer-term effects would be tricky, because of the issue of the valid comparison group.

One might consider adopting a synthetic control group approach. The use of synthetic control techniques with disaggregated data is discussed by Abadie and L'Hour (2021). These techniques may not solve the problem of people entering and exiting treatment status (payment receipt) over the period of interest. Nevertheless, they may assist in generating credible counterfactual outcomes.

A set of outcomes of interest that might satisfy these requirements are expenditure and saving patterns. Effects on expenditure and saving can be immediate, and can give rich insights into how targeted families are using additional funds. Such an analysis would facilitate a better understanding of the short-term responses of households to a short-term cash transfer, and of how different households prioritise their financial issues.

For this sort of analysis, potential datasets would need an identifier for receipt of the Coronavirus Supplement (or qualifying payments such as JobSeeker). This is a more stringent requirement than the Baby Bonus analysis, which needs only access to an individual's date of birth. Data with such identifiers include financial transaction data, which can be accessed through the large databases of credit bureaus such as illion. These data have been used to estimate consumer spending patterns, as in Elias (2022), as well as the Covid-era policy of Superannuation withdrawal (Hamilton et al., 2023). A serious challenge for using such data in the present context is the difficulty in identifying families, as transactions are recorded at the individual level. Identifying families, let alone disadvantaged families with young children, would be difficult.

Other administrative datasets such as HSDS and MADIP face similar challenges, although they may be surmountable. The HSDS contains many detailed potential outcomes. While it appears that it is not currently feasible to identify individuals who received relevant payments, this would be possible if HSDS could be linked to MADIP over the period of interest. Without that link, MADIP would be of limited value. While MADIP would be able to identify those who received the payment, the set of potential outcome variables in that database is rather limited. See Section 8 (Australian Data Landscape) for further detail.

Overall, there are considerable challenges associated with conducting an informative analysis of the impacts of the Coronavirus Supplement, due to the nature of the reform and the strong data requirements. However, this may be worthy of further investigation, potentially in partnership with organisations which hold and analyse financial transaction data, such as the CBA, the e61 Institute, and the TTPI at ANU.

11. Summary Comparison of Techniques

We now discuss ‘best cases’ of how each technique could be used to evaluate the potential effects of unconditional cash transfers on the outcomes of vulnerable children and their families. There is a key distinction between prospective and retrospective evaluations – which we highlight repeatedly below. A prospective evaluation would be appropriate for Scenario 1. A retrospective evaluation would be appropriate for Scenario 2. See Section 1.3 for a discussion of these Scenarios.

11.1 Summary

The Table below summarises key characteristics of the impact evaluation techniques discussed below.

Methods	Typical use	Assumptions required to infer treatment effects in the population	Data requirements for typical use
Randomised Controlled Trials	Prospective	Minimal	Survey and/or administrative data linked with program assignment
Adaptive Randomised Controlled Trials	Prospective	Minimal	Same as RCTs with enrolment and data collection for successive cohorts
Regression Discontinuity Design	Retrospective	RDD estimates apply to people who are far from the threshold (untestable)	Need large number of observations (surveys or administrative data) in a context where program eligibility creates discontinuities (e.g., income eligibility threshold)
Difference-in-Differences	Retrospective	The treated group follows the same trend as the control group over time (untestable)	Panel data (survey or administrative data) for treated and comparison groups before and after program implementation
Matching	Retrospective	There is no difference in unobserved characteristics between the treatment and the control group (untestable)	Survey and/or administrative data linked with program take-up
Instrumental Variables	Prospective (within an RCT) / retrospective	The instrumental variable must have an effect on the treatment take-up (testable) and cannot directly impact the outcome of interest (untestable)	Depends on use
Synthetic Control	Retrospective	The treated group follows the same trend as the control group over time (untestable)	Panel data (survey or administrative data) for treated and comparison groups before and after program implementation
Machine learning	Not suitable for impact evaluation on its own		Large dataset

11.2 Randomised Controlled Trials (RCTs)

The RCT is a **prospective** evaluation design. RCTs are not feasible to employ in retrospective evaluations.

Out of all the options for prospective evaluation, RCTs are always a first choice unless there are compelling reasons to the contrary. RCTs offer complete control over all aspects of the study, and are limited only by budget, time, ethical and feasibility constraints. Controllable features include the design features of the transfer (e.g., size, timing, regularity, plus combinations of those aspects), the population (and sub-populations) to be studied, the outcome variables to be observed, and the timing of observation.

11.3 Regression Discontinuity Design (RDD)

While RDD could in principle be used for a prospective evaluation, it is predominantly (perhaps exclusively) used for retrospective evaluations. Perhaps the best prospect of using RDD is in relation to the Baby Bonus natural experiments, where eligibility for the various versions of the payments is defined by precise date of birth thresholds. Deutscher and Breunig (2018) ultimately preferred a Difference-in-Differences approach, due to concerns about the birth-timing effects induced by the 2004 policy change, which make RDD problematic. We believe there is, nevertheless, scope to use RDD in this context, in a number of ways and for a number of reasons:

1. In order to navigate birth-timing effects, one can adopt a:
 - a. ‘donut hole’ technique, as per Barecca et al. (2011), or
 - b. ‘bounds’ technique, as per Rosenman, Rajkumar, Gauriot and Slonim (2021).
2. Later policy changes may have been less prone to DOB manipulation, and
3. DOB manipulation may be less problematic for some outcome variables.

Ultimately, the best approach is likely to be to use RDD **in combination with** a Difference-in-Differences analysis strategy of this natural experiment. The greatest limitation of this approach is that it can only address very specific questions. As described in Section 9, one can use RDD to examine (i) the effect of a small cash transfer received as a lump sum on the birth of child, and (ii) the effect of receiving a small lump sum versus regular payments equalling the same value. This is likely to be informative, but may differ from any specific transfer that is implemented in the future. Strictly speaking, these RDD effects are estimated only for children born on precise dates, although it seems reasonable to assume that such effects would be similar for children with other DOBs.

Using RDD in a prospective design would avoid the need to randomise, but would have severe limitations. It would require withholding access to the program for some people, and doing so based on a threshold (for example, only households below a given income threshold are eligible). An important limitation is that the RDD estimates would apply only to households at that precise threshold, rather than all eligible households. For example, it would not be informative about the impact of the program for households with very low income, well below the threshold. Because of the focus on specific thresholds, RDD is also ‘data hungry’ – in the sense that very large samples are usually required for adequate statistical power, compared with RCTs.

11.4 Difference-in-Differences (DD)

DD can be conceptualised in several ways, which we consider in turn below:

DD can be thought of as an **analytical technique** rather than a research design. It can be used, for example, as a way of analysing data from an RCT, simply by choosing an outcome variable that is defined as the change in a given variable (See, e.g., Stock & Watson, 2019, Section 13.3). An example of this in the cash transfer context is to define the outcome variable as: (a) the difference between maternal stress measured after and before the transfer (the first difference), and (b) the extent to which this first difference would differ between people who received the transfer and those who did not. This would be a DD analysis of experimental data.

Much more commonly, DD refers to the application of such techniques for **retrospective** evaluation, usually leveraging some sort of policy change. Typically, such studies consider the change in outcome (before and after) for a treated group, and compare this to the change for an untreated comparison group. There are many variations to this basic approach. For example, Deutscher and Breunig (2018) compare the difference in outcome between children born in the few months after 1 July 2004, and children born in the few months before (the first difference), and then compare this to similar differences for adjacent years of birth (the second difference). As discussed elsewhere, we regard this as a promising avenue for further exploration, using additional outcome variables and other DOB thresholds. DD is less 'data hungry' than RDD, but requires stronger assumptions – in the example above, it requires an assumption that differences in the characteristics of children across months of birth are the same for those born in 2004 as for those born in 2005. This is an untestable assumption. As per the RDD example above, DD can only be used to evaluate precise policy changes that have already been implemented.

Similarly, DD-style techniques could be used to evaluate the effects of the Coronavirus Supplement. One complication of such an analysis is that selection of a valid untreated comparison group would be difficult. Many people moved in and out of receiving the payment over time, while those who did not receive it may not serve as a good comparison group. Nevertheless, we believe this is worth investigating further, especially to study short-run effects on types of expenditure, saving and debt reduction.

Using DD within a non-randomised prospective evaluation is rarely performed. While it would avoid the need to randomise, it would provide lower-quality evidence than a randomised experiment, due to the assumptions it requires, and it would also come with essentially the same costs. The key assumption is that the treated group would follow the same trend as the untreated group over time. This is untestable. By contrast, this is assured by design where there is random allocation to treatment.

The same discussion applies to **Event Studies**.

11.5 Matching

Like DD, matching can be conceptualised as an analytical technique, and can be combined with other techniques. It is usually discussed in the context of **retrospective** evaluations, as a partial solution to the problem of ‘selection bias’ in the absence of random assignment. It is difficult to think of any contexts where a simple matching analysis would provide credible evidence on the impact of an unconditional cash transfer program in Australia. This is because matching can only account for differences in the observed characteristics of treated and untreated groups. It invokes the major assumption that there are no important unobserved differences in characteristics. However, matching can be used in combination with DD, where it can help in the selection of a credible comparison group based on pre-treatment characteristics. See Doiron (2004) for an example in the context of welfare reform in Australia. One could consider using matching in combination with DD to analyse the Coronavirus Supplement natural experiment. As discussed above, though, this analysis may be hindered by people moving in and out of payment receipt over time.

It is challenging to conceive of a role for matching in the context of prospective evaluation design. Treatment allocation is within the control of a prospective research design. It would make little sense to allocate treatment in a way in which matching techniques would be useful.

11.6 Instrumental Variables (IV)

Like many of the other techniques discussed, IV regression is an analytical technique, and it has broad application.

IV is commonly used for both **prospective** and **retrospective** evaluations. In the case of **prospective evaluation**, it is applied in two main scenarios. The first is as a solution to the problem of partial compliance in an RCT – that is, a situation where not everyone who was assigned to the treatment group actually received the treatment, and/or where some people in the control group actually received the treatment. The second is by deliberate design. Rather than randomly assigning a treatment, some experiments randomly assign eligibility for treatment (this is often the case with conditional cash transfer experiments), or they may randomly nudge or encourage people to take a given treatment. In both cases, partial compliance is built into the design of an experiment, for which Instrumental Variable Regression is the appropriate analytical tool. There may be reason to consider an experiment with such characteristics in the present context. However, it seems likely that a simple RCT would be the better option.

In the context of **retrospective evaluation**, IV can also be used to solve similar complications of partial compliance. For example, if eligibility for the Baby Bonus or Coronavirus Supplement did not always translate into take-up of the payment (and if eligibility is observed), one could use eligibility as an instrumental variable for take-up – either in the context of RDD or DD. However, we do not think this scenario is likely for either payment.

11.7 Adaptive Trials (including Bayesian Adaptive Trials)

Adaptive Trials are a specific type of experimental design and hence a type of **prospective evaluation** design. As discussed in Section 6, this type of experiment involves pre-specification of decision rules which allow the experiment to be flexible, according to early results. This necessitates the ability to alter the nature of the treatment (e.g., size or timing), or the scope of who receives it, or to cease the experiment, depending on early results. This is a complex set of factors that needs to be considered in the context of a specific experimental design. It is worthy of close consideration if the decision is made to proceed with experimental design.

11.8 Synthetic Control (SC)

Synthetic Control can be seen as an extension of DD. It has been applied (perhaps exclusively) for retrospective evaluation. While DD generally assigns equal weight to all comparison entities, SC instead assigns unequal weights which are chosen in a way that ensures 'pre-treatment' trends (and observed characteristics) of the treated and untreated groups are similar. SC studies are often characterised by pre-treatment trends which appear remarkably similar in the treatment and control groups. But this can be misleading, because the similarity in pre-treatment trends is precisely what the weights are chosen to optimise. Nevertheless, SC does not avoid the untestable assumption (shared with DD) that this 'common trend' would have continued into the post-treatment period. SC can therefore be seen as a refinement of the DD approach, with arguably stronger validity, but not as a silver bullet for the evaluation problem. It could be considered as a component of a Coronavirus Supplement evaluation, discussed above.

It is possible to design a prospective evaluation, with a plan to adopt SC. For example, a decision could be made to provide cash transfers to a particular sub-population (e.g., people in a given state or demographic group). The costs of such a design would be similar to an RCT. However, the strength of resulting evidence would be lower, as it would rely on untestable assumptions, and it is a less transparent technique. Compared with an RCT, a prospective SC would be less likely to have broader policy impact, and appropriately so.

11.9 Machine Learning (ML)

ML has made an impact in many realms. ML techniques have been designed for the purpose of prediction/categorisation tasks, leveraging the strengths of 'big data' with many observations and many variables. However, as discussed in Section 5.4, the emergence of ML is not regarded as a major development in the technology of impact evaluation methods, neither for prospective nor retrospective evaluation. In particular, ML does not solve the major challenges of impact evaluation with non-experimental data, such as the problem of 'selection on unobservables' and associated bias. At best, ML may be useful in the present discussion in assisting with selecting appropriate groups for sub-group analysis (either for prospective or retrospective evaluation methods), or with refining the selection of covariates to use in an observational data analysis.

12. Recommendations & Conclusions

Conducting an RCT to identify the best implementation choices for cash transfer programs for vulnerable Australian families is crucial, considering the current lack of confident knowledge on optimal approaches. While existing literature suggests that unconditional cash transfers can enhance family well-being, determining the most effective way to implement such programs remains uncertain. In light of this, the potential benefits of conducting an RCT far outweigh the relatively small marginal costs.

12.1 Recommendation to conduct an RCT

Randomised Controlled Trials (RCTs) are widely regarded as the gold standard for providing empirical evidence establishing causal relationships in the social sciences and more broadly. The current context – how to better understand the effects of cash transfers on the well-being of recipients – is no exception. An RCT provides numerous insights that go well beyond the evidence in the extant literature, or that could be established using alternate methods. Furthermore, there are very few drawbacks to an RCT. Below we briefly summarise the main advantages and disadvantages of conducting an RCT in order to understand the effects of direct cash transfers on a targeted population's well-being, and the channels through which these impacts occur.

Primary Advantages

External Validity: External validity is the single most important reason for running an RCT. A study's external validity refers to the applicability of its conclusions to inform the possible outcomes of a similar policy in a different context. While existing empirical evidence and other alternative research methods can provide insights into the effectiveness of unconditional cash transfers, there will always remain reasonable questions as to whether the same results (i.e., impact) would hold for a new/different cohort of people receiving cash transfers. Conducting an RCT to test different implementation strategies and options with a sample of people from the target cohort would eliminate virtually all doubt about its external validity; in other words, if a policy was implemented that had a set of effects within a sample of people in an RCT, one could be extremely confident that it would have the same effects if implemented across the entire cohort of people from which the sample was drawn.

Testing Implementation Choices: There are many ways one can choose to implement/deliver an unconditional cash transfer, and neither theory nor existing evidence can pinpoint which one(s) are the most beneficial. Yet the option(s) chosen are likely to be critical to the effectiveness of the transfers. For instance, we can intuitively (and theoretically) conjecture that effectiveness will depend on how and when the transfers are distributed to families with newborn children. However, it is theoretically ambiguous

whether the optimal time is immediately after a child is born, or at some point before the birth (e.g., one, two or three months prior), and it is also unclear whether the transfer should take the form of a lump sum or a series of payments over time (e.g., monthly or quarterly).⁵ Thus, an RCT in which the initial timing and form/frequency of payment are varied would provide valuable evidence on the implementation approach(es) that are most effective.

Note that there are multiple methods of conducting an RCT.

- One method is a **'regular' RCT with multiple treatment arms**, in which participants are randomly allocated to a control group (i.e., no cash transfer) or to different treatment arms (e.g., Treatment 1, allocates \$10,000 in a lump sum payment at some point during pregnancy; Treatment 2, allocates a total of \$10,000 in equal monthly instalments beginning at some time during pregnancy). **This method is suitable for conducting an RCT within an existing longitudinal or cohort study.** The number of treatment arms that can be considered depends on various factors, including anticipated effect size, sample size and levels of randomisation.
- A second method is to conduct **'adaptive' RCTs** (see Novel Methods, Section 6). Specifically, this would involve first running an RCT with a cohort 1, then using initial results from cohort 1 to refine and improve the program design to conduct a second RCT with a cohort 2, then using longer-term results from cohort 1 and initial results from cohort 2 to further refine and improve the program for a third RCT with a cohort 3, and so on. Note that the time period between cohorts depends on which short-term measures are of most interest. Using pre-specified criteria, the treatments with the most promising results from previous cohorts can be identified and adapted to develop and test a narrower set of treatments. Such an approach is powerful in terms of optimising implementation choices, but requires **multiple birth cohorts** over time. An adaptive RCT approach can be either 'Bayesian' or 'frequentist':
 - To use the Bayesian approach, it is helpful to impose 'priors' – a pre-trial probability distribution – on the potential size of the treatment effect based on existing knowledge. This can be derived from a thorough literature review, although our understanding is that the empirical evidence on cash transfers comes primarily from low-income countries (LICs), and it is unclear whether the treatment effects identified in LICs can serve as a credible prior for the treatment effect of cash transfers for vulnerable families in Australia.
 - Nonetheless, even with 'flat' priors, a Bayesian approach or a frequentist approach can be used to iterate program implementation choices towards the most successful unconditional cash transfer parameters over time.

5. For instance, from a rational economic decision-making perspective, providing an unconditional cash transfer prior to a child being born could theoretically alleviate financial constraints and improve the mother's and the (unborn) child's health, by enabling the mother to either seek increased pre-natal medical care and medicine or to leave the workforce at a more optimal time for the child's birth. On the other hand, people do not always make optimal choices, and they often spend cash immediately rather than spreading their expenditure over time and using the money more effectively (known as hyperbolic discounting in the literature); thus a lump sum transfer before birth could lead to money being spent unnecessarily early, rather than later, when it could be more beneficial.

Multiple Outcomes and Underlying Mechanisms: An RCT on an unconditional cash transfer can deliver evidence on a broad range of outcomes, providing a better understanding of not only its effectiveness across many well-being outcomes, but also of the channels through which these outcomes are affected. As discussed earlier, unconditional cash transfers can have an impact on multiple outcomes affecting the recipient's well-being (i.e., on both the child/children and parents). These outcomes could include, but are not limited to, improvements in health, educational achievement, employment and income. The data collection associated with an RCT can also include measures for many additional (intermediate) outcomes that are both cognitive and non-cognitive, and which result in better understanding of the channels through which the cash transfers may be effective, (and consequently help improve the design of future cash transfers). These intermediate measures can be collected through a combination of surveys, time-use diaries, and physiological, biological and other measures, starting from before the RCT begins to a few months or years after.⁶ Understanding the mechanisms – especially if the number of treatment arms is limited – is key to providing credible recommendations on how to improve the program design when rolling it out at scale. In particular, detailed measures offer the possibility of combining RCTs with structural modelling (see Section 6).

The need for few assumptions

RCTs allow precise control over what is randomised and thus what is evaluated. Moreover, because an RCT creates random variation in exposure to the cash transfer program, all else equal, the evaluation does not require as many assumptions in relation to causal statements as are required with other methods. In an RCT, the random assignment of recipients to different treatment arms minimises the likelihood that any eventual differences in outcomes between the treatment arms are due to differences in recipients, while maximising the likelihood that any differences in outcomes are caused by the variations in treatment conditions. In other words, an RCT greatly increases the credibility of the inference that can be made and the lessons that can be learned.

Transparency of findings

The results of a randomised evaluation are relatively straightforward to interpret, which gives them transparency. For example, in a simple RCT with a treatment and a control group, the treatment effect can be identified by comparing the mean of the relevant outcomes in the treatment and control groups. This transparency of findings can be useful for convincing policy-makers to expand or roll out a program.

6. Note: we are not suggesting that every RCT participant would be required to provide multiple measures, as that could be overwhelming, but rather that sub-sets of participants could be asked to comply with one or two of these additional measurements.

Potential to influence the Australian policy landscape

RCTs produce evidence that is more likely to influence policy due to several key factors. Firstly, RCTs are widely recognized as providing the highest-quality evidence among all impact evaluation techniques. This high quality is derived from the rigorous design and control they employ, which enables researchers to confidently attribute observed outcomes to the intervention being evaluated. This credibility and reliability make RCT findings more compelling to policymakers, public servants, politicians, and even the general public.

Secondly, the simplicity and intuitive nature of RCTs make them more accessible and easily understood by a broader audience. Unlike alternative quasi-experimental evaluation methods, which may be complex and require specialized knowledge to comprehend, RCTs offer a straightforward approach that can be communicated effectively to non-experts. This ease of comprehension facilitates the communication of RCT findings, enhancing their potential to garner support and sway decision-makers towards evidence-based policies.

This is the strong view of the Hon Dr Andrew Leigh, an assistant minister in the federal government, the author of *Randomistas* (Leigh, 2018), and driver of new Australian Centre for Evaluation.

Considering these factors, conducting an RCT can yield substantial benefits. RCTs possess the ability to generate compelling evidence that resonates with policymakers and the public, potentially leading to significant policy changes based on the robust empirical findings they provide.

Potential Disadvantages:

There are two potential disadvantages to conducting an RCT: (1) the time required, and (2) the cost of running it. Below we provide different scenarios to assist with the weighing up of time and financial costs.

Time spent waiting for results

Some of the primary outcomes of interest include children's educational attainment, their employment as an adult, incarceration, etc. These outcomes can only be observed and measured decades after the program implementation.

However, many shorter-term outcomes that can be measured in an RCT are good indicators of outcomes later in life, and they can be used to evaluate the success of the program much more quickly. A number of studies show that early cognitive ability and non-cognitive behaviour have significant, long-term effects on education, earnings, health, longevity and other long-run measures of well-being (for example, see Almond & Currie, 2010; Heckman et al., 2010). Preferences, non-cognitive skills and executive function measured in early childhood also predict outcomes later in life (e.g., Castillo et al., 2020; Shoda et al., 1990).

Cost of running an RCT study

It is useful to consider various scenarios in order to establish the cost of running an RCT.

Scenario 1: An agency has committed to implementing a cash transfer program

In this scenario, we assume that an agency will run a cash transfer program, and the decision it now faces is whether to embed an RCT to the roll-out of the program. So for this discussion, we will consider the alternative to running an RCT will be to instead move to directly to implementing a conditional cash transfer without running an RCT.

To illustrate why running an RCT in this scenario would not be especially costly in time or money, let us consider a hypothetical unconditional cash transfer program that would be implemented from 1 January 2024. Suppose the following choices have been made for this illustrative exercise:

1. Eligibility for the first cohort will be any family with an income below a certain threshold and a child born in 2024. (There may be additional conditions, TBD). Eligibility for the second cohort will be the same, for children born in 2025, and this may continue into the third, fourth and further cohorts in subsequent years.
2. Each family meeting the eligibility requirements will receive three lump sum payments of \$10,000, over three years.
3. The first lump sum will be paid one month after the child is born, then the two other payments will follow one year and two years later.
4. Families will be required to complete regular surveys examining various outcomes, and to give the funding organisation permission to access information from other sources, such as Medicare and the ATO.
5. The program will be accessible to a large number of families. It will be either exhaustive (i.e., accessible to all families that meet the eligibility criteria), or will be offered to a sub-set of families that meet the criteria.

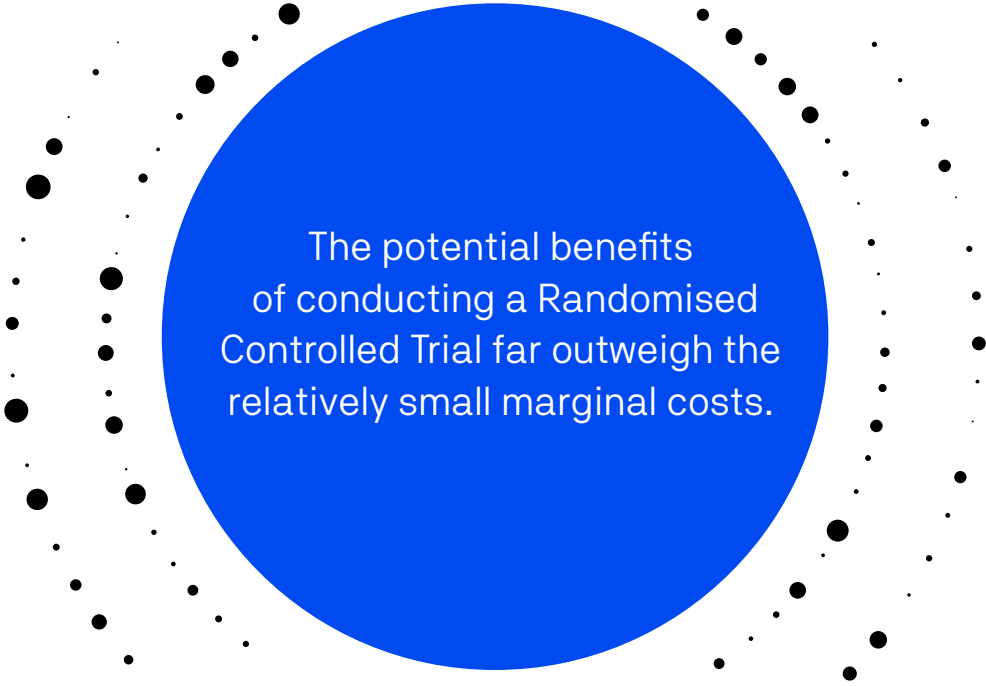
If we are confident that a program run according to these principles will deliver the largest possible impact on the well-being of recipient families, then an RCT is unnecessary.

However, it is unlikely that we can have that level of confidence based on the existing evidence. In terms of running an RCT, the main advantages are articulated above. In relation to the disadvantages, let us focus on the **marginal** (i.e., additional) **costs** of implementing an RCT, over and above the costs of implementing the program without such an evaluation.

Implementation cost: One straightforward way to run an RCT would be to explore one of the design options. For instance, in order to determine whether the timing of the first payment just after birth is optimal, a sub-set of eligible participants could be randomly assigned to receive their first payment three months before the child is born. This would be trivial to implement in terms of time and financial cost, and would not require any additional cash transfers.

Survey costs: Data collection often represents the bulk of the financial cost of an RCT. This cost will vary greatly, depending on the particular data needs of a study. Data collection costs can be reduced by obtaining consent from participants for linkages with administrative data. Evidence from other Australian studies suggests that the consent rate for such data linkages is high. For example, the Longitudinal Study of Australian Children obtained a 93% consent rate for linkages with the Medicare Benefits Schedule in wave 1, 90 to 95% consent for linkages with NAPLAN, and 90% for linkages with the Australian Early Development Census.⁷ While using existing data greatly reduces the cost of data collection, it means that only outcomes for which data have already been collected can be examined. Combining surveys with administrative data can be a good compromise. Survey costs may be low if tracking is put in place for implementation of a cash transfer program without an RCT (although tracking of the control group would need to be added to the original cost), or if the RCT is embedded in an existing longitudinal study.

Respondents' burden: Participation in an RCT can be quite demanding, and can represent a significant burden. The burden can be reduced by: (i) using well-tested questions and visual aids (Delavande & Rohwedder, 2008); (ii) using administrative data linkages; (iii) measuring outcomes that do not require respondents to answer questions (e.g., persistent stress level can be measured in children and adults through hair samples (Bates et al., 2017); parental stimulation can be measured via a child-safe recorder that a child wears for a day (Zimmerman et al., 2009)). Participants can also be given a small financial incentive to as a thank you for their time devoted to data collection.



The potential benefits
of conducting a Randomised
Controlled Trial far outweigh the
relatively small marginal costs.

7. Source: <https://growingupinaustralia.gov.au/data-and-documentation/lsac-data-linkages>

Scenario 2: An agency has not committed to implementing a cash transfer program

In this scenario, we assume that an agency does not intend to run a cash transfer program, and the decision is whether to conduct an RCT to evaluate the effectiveness of cash transfer programs to families below the poverty line in Australia, or to do nothing. Again, the main advantages are articulated above. For the disadvantages, we focus on the **total** costs of implementing an RCT.

Implementation cost: The main implementation cost is the cash transfers given to the RCT participants allocated to the treatment groups. The actual cost will depend on the size of the transfer, and the sample size of the treatment groups.

Survey costs: The costs are similar to those described in Scenario 1. In Scenario 2, however, total survey costs need to be calculated, as no other tracking would be put in place by default. One exception would be if the RCT was embedded into an existing longitudinal study, which could considerably reduce the survey costs.

Respondents' burden: As in Scenario 1.

Note that these costs do not account for any potential improvement in well-being of families who receive the cash transfers. These benefits are hard to quantify prior to running a proper evaluation.

Scenario 3: An agency allocates cash transfers to participants instead of supporting other programs

Scenario 2 was extreme in that it allocated all of the implementation costs to the cost of running an RCT. An alternative way to consider these costs is in terms of them displacing other programs supported by the same budget.

Implementation cost: The marginal implementation cost of conducting an RCT is the difference between the cost of the cash transfers and the cost of alternative programs that are not going ahead as a result of implementation of the cash transfer program. This cost may well be zero. Another important consideration is the net change in participants' well-being as a result of the changed program focus. Again, this is very difficult to assess, because (i) the cash transfer program has not been properly evaluated; (ii) it is not clear whether the impacts of the alternative programs have been evaluated.

Survey costs: As in Scenario 2 (or lower, if other programs that are displaced had a tracking system that is shut down).

Respondents' burden: As in Scenario 1.

Ethical considerations of conducting an RCT

Ethical guidelines for evaluation and research include four key principles:

Respect for persons: participants should be informed of risks and given a choice about participation (Informed consent).

Beneficence: the risks of research should be carefully weighed against the benefits. Risks should be minimised. Researchers should avoid knowingly doing harm – for example, by encouraging take-up of a program that they have reason to believe is harmful.

Justice: The allocation of risks and benefits between different groups of people should be fair. Minority groups should be explicitly included in trials.

Respect for law and public interest: It extends the principle of Beneficence beyond specific research participants to include all relevant stakeholders.

More details can be found in the National Statement on Ethical Conduct in Human Research (2007).

Informed consent

The respect-for-persons principle requires that we tell those who we would like to take part in a study what the research is about and ask their permission to make them part of the study. This is usually done as part of the baseline survey. The experiment is explained, and participants are asked for their permission to continue. This should be straightforward to achieve in the context of a cash transfer RCT involving a baseline survey.

Evaluating questions of relevance to the population being studied

Usually, impact evaluations focus on the relevant population. Running an RCT with families below the poverty line to study the impact of a cash transfer at birth on children's outcomes for families below the poverty line enables one to comply with the justice principle.

Do randomised evaluations deny access to a program to some?

A concern related to the beneficence principle is whether randomised evaluations do harm because they lead to people being denied access to a program from which they would otherwise have benefited. This concern invokes the assumption that we know that a program is beneficial. In most cases, however, we do not know the impact of the program – that is why we are evaluating it. The argument that randomised evaluations are ethical stems from the position that, if we do not know whether a program is beneficial (or which implementation choice is optimal), society benefits if its impacts are tested before it is scaled up to more people. There may be some risk to those who experience the program first, or some potential loss of well-being to those denied the program first, but these need to be balanced against the potential benefits of a better understanding of the program's impacts. There are many examples of cases where an intervention that was assumed to be helpful was later found to be ineffective or harmful.

Furthermore, denying access to a program to some people effectively already occurs in many scenarios without RCTs: (i) Interventions are often piloted, which is equivalent to excluding a group of people who could benefit. (ii) Interventions are often scaled up, meaning it takes time for everyone to receive the intervention. An RCT might be considered more ethical than either of those scenarios, as it is a similarly phased introduction, but one with a mechanism to determine its effectiveness and improve future outcomes.

Finally, conducting a randomised evaluation usually neither increases nor decreases the number of people likely to receive a program. It may instead shift the geographical location of participants. For example, instead of all individuals in one suburb being given access to a program, half of individuals across two suburbs may be given access.

Weighing the risks and benefits

It can be ethical to perform a randomised evaluation even if doing so reduces the number of people receiving a program in the short term. Again, the likely risks need to be weighed against the likely benefits. We have to weigh the potential harm caused to those who would have received the program without an RCT against the potential benefits of having rigorous evidence on the program's impact. The benefits include the creation of credible evidence on the positive impact of the program, which can lead to more funding being raised and the program being rolled out on a wider scale in the long run if it is effective. Another benefit could be the replacement of the program with a more effective iteration if the impact is small. The RCT may also avoid harm, in a scenario where the program has an unintended negative effect, and is not rolled out as a result of the evaluation.

12.2 Suggested Study 2: Baby Bonus Reform Evaluation

It may be worthwhile to evaluate the three Baby Bonus reforms using quasi-experimental techniques (See Section 9). However, this should be conditional on confirming the availability of access to relevant data; in particular, whether the DOB field in the NSW HSDS database can be accessed. This would need to be approved by special arrangement.

12.3 Suggested Study 3: Coronavirus Supplement Evaluation

In some ways, the Coronavirus Supplement is a less promising natural experiment than the Baby Bonus, mainly due to the difficulty in forming a credible comparison group, or otherwise inferring counterfactual outcomes. But this is less of an impediment for effects that are likely to be large and immediate. There is the potential to use transaction data from the CBA to study the short-run effects of the Supplement in rich detail and for the relevant population. This is worth exploring further.

Conducting a Randomised Controlled Trial to identify the best implementation choices for cash transfer programs for vulnerable Australian families is crucial.

References

- Abadie, A. and L'Hour, J. (2021). A Penalized Synthetic Control Estimator for Disaggregated Data. *Journal of the American Statistical Association*, 116(536), pp.1817–1834. doi:10.1080/01621459.2021.1971535.
- Almond, D., and Currie, J. (2010). Human Capital Development before Age Five, Handbook of Labor Economics, vol 4. Ashenfelter, O., and Card, D., editors, Ch 15, 1315-1486.
- Álvarez, C., Devoto, F., & Winters, P. (2008). Why do beneficiaries leave the safety net in Mexico? A study of the effects of conditionality on dropouts. *World Development*, 36(4), 641-658.
- Angrist, J., Bettinger, E., Bloom, E., King, E. and Kremer, M. (2002). Vouchers for Private Schooling in Colombia: Evidence from a Randomised Natural Experiment. *American Economic Review*, 92(5), pp.1535–1558. doi:10.1257/000282802762024629.
- Athey, S. and Imbens, G.W. (2017). The State of Applied Econometrics: Causality and Policy Evaluation. *Journal of Economic Perspectives*, [online] 31(2), pp.3–32. doi:10.1257/jep.31.2.3.
- Athey, S. and Imbens, G.W. (2019). Machine Learning Methods That Economists Should Know About. *Annual Review of Economics*, 11(1), pp.685–725. doi:10.1146/annurev-economics-080217-053433.
- Attanasio, O. P., Oppedisano, V., & Vera-Hernández, M. (2015). Should cash transfers be conditional? Conditionality, preventive care, and health outcomes. *American Economic Journal: Applied Economics*, 7(2), 35-52.
- Attanasio, O., Sosa, L. C., Medina, C., Meghir, C., & Posso-Suárez, C. M. (2021). Long term effects of cash transfer programs in Colombia (No. w29056). National Bureau of Economic Research.
- Audit Office of New South Wales (2020). Their Futures Matter. <https://www.audit.nsw.gov.au/our-work/reports/their-futures-matter>.
- Australian Bureau of Statistics (2020). *MADIP Modular Product (2011 - 2019) Data Item List*. [online] ABS. Available at: <https://www.abs.gov.au/statistics/microdata-tablebuilder/available-microdata-tablebuilder/multi-agency-data-integration-project-madip#data-downloads>.
- Australian Bureau of Statistics (2021a). *Access and services | Australian Bureau of Statistics*. [online] www.abs.gov.au. Available at: <https://www.abs.gov.au/about/data-services/data-integration/access-and-services>.
- Australian Bureau of Statistics (2021b). *Multi-Agency Data Integration Project (MADIP) | Australian Bureau of Statistics*. [online] www.abs.gov.au. Available at: <https://www.abs.gov.au/about/data-services/data-integration/integrated-data/multi-agency-data-integration-project-madip>.
- Australian Early Development Census (2022). *About the AEDC*. [online] www.aedc.gov.au. Available at: <https://www.aedc.gov.au/about-the-aedc>.
- Bahety, G., Bauhoff, S., Patel, D. and Potter, J. (2021). Texts don't nudge: An adaptive trial to prevent the spread of COVID-19 in India. *Journal of Development Economics*, [online] 153, p.102747. doi:10.1016/j.jdeveco.2021.102747.
- Baird, S., McIntosh, C., & Özler, B. (2011). Cash or condition? Evidence from a cash transfer experiment. *The Quarterly Journal of Economics*, 126(4), 1709-1753.
- Banerjee, A. V., Hanna, R., Kreindler, G. E., & Olken, B. A. (2017). Debunking the stereotype of the lazy welfare recipient: Evidence from cash transfer programs. *The World Bank Research Observer*, 32(2), 155-184.
- Barr, A., Eggleston, J. and Smith, A.A. (2022). Investing in Infants: the Lasting Effects of Cash Transfers to New Families. *The Quarterly Journal of Economics*, 137(4). doi:10.1093/qje/qjac023.
- Barreca, A.I., Guldi, M., Lindo, J.M. and Waddell, G.R. (2011). Saving Babies? Revisiting the effect of very low birth weight classification. *The Quarterly Journal of Economics*, 126(4), pp.2117–2123. doi:10.1093/qje/qjr042.
- Bates R, Salsberry P, Ford J. Measuring Stress in Young Children Using Hair Cortisol: The State of the Science. *Biol. Res. Nurs.* 2017 Oct;19(5):499-510. doi: 10.1177/1099800417711583. Epub 2017 Jun 15. PMID: 28617035; PMCID: PMC6775674.

- Benhassine, N., Devoto, F., Duflo, E., Dupas, P., & Pouliquen, V. (2015). Turning a shove into a nudge? A “labeled cash transfer” for education. *American Economic Journal: Economic Policy*, 7(3), 86-125.
- Bertrand, M., Duflo, E. and Mullainathan, S. (2004). How Much Should We Trust Differences-In-Differences Estimates? *The Quarterly Journal of Economics*, 119(1), pp.249–275. doi:10.1162/003355304772839588.
- Blomquist, J. (2003). Impact evaluation of social programs: A policy perspective
- Bound, J., Jaeger, D.A. and Baker, R.M. (1995). Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogenous Explanatory Variable is Weak. *Journal of the American Statistical Association*, 90(430), p.443. doi:10.2307/2291055.
- Broglio, K., Meurer, W.J., Durkalski, V., Pauls, Q., Connor, J., Berry, D., Lewis, R.J., Johnston, K.C. and Barsan, W.G. (2022). Comparison of Bayesian vs Frequentist Adaptive Trial Design in the Stroke Hyperglycemia Insulin Network Effort Trial. *JAMA Network Open*, 5(5), p.e2211616. doi:10.1001/jamanetworkopen.2022.11616.
- Calonico, S., Cattaneo, M.D. and Titiunik, R. (2014). Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs. *Econometrica*, 82(6), pp.2295–2326. doi:10.3982/ecta11757.
- Card, D. and Krueger, A.B. (1994). Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania. *The American Economic Review*, [online] 84(4), pp.772–793. Available at: <https://www.jstor.org/stable/2118030>.
- Castillo, Marco, List, John A, Petrie, Ragan, Samek, Anya. (2020). Detecting Drivers of Behavior at an Early Age: Evidence from a Longitudinal Field Experiment, National Bureau of Economic Research Working Paper Series No. 28288.
- Cattaneo, M.D. and Titiunik, R. (2022). Regression Discontinuity Designs. *Annual Review of Economics*, 14(1), pp.821–851. doi:10.1146/annurev-economics-051520-021409.
- Chan, T. Y., & Hamilton, B. H. (2006). Learning, private information, and the economic evaluation of randomised experiments. *Journal of Political Economy*, 114(6), 997-1040.
- Chen, J. and Langwasser, K. (2021). COVID-19 Stimulus Payments and the Reserve Bank’s Transactional Banking Services. Reserve Bank of Australia.
- Cobb-Clark, D.A. and Siminski, P. (2019). Labor’s idea of an Evaluator General could dramatically cut wasteful spending. *The Conversation*. Available at: <https://theconversation.com/labors-idea-of-an-evaluator-general-could-dramatically-cut-wasteful-spending-115840>
- Columbia University (2022). *Difference-in-Differences Estimation | Columbia University Mailman School of Public Health*. [online] www.publichealth.columbia.edu. Available at: <https://www.publichealth.columbia.edu/research/population-health-methods/difference-difference-estimation>.
- Cook, T.D. (2008). ‘Waiting for Life to Arrive’: A history of the regression-discontinuity design in Psychology, Statistics and Economics. *Journal of Econometrics*, 142(2), pp.636–654. doi:10.1016/j.jeconom.2007.05.002.
- Cunningham, S. (2021). Causal inference. In *Causal Inference*. Yale University Press.
- D, A. and J, C. (2010). Human Capital Development before Age Five. *Handbook of Labor Economics*, vol 4, Ashenfelter, O., and Card, D., editors (Ch 15), pp.1315–1486.
- Dahl, G.B. and Lochner, L. (2012). The Impact of Family Income on Child Achievement: Evidence from the Earned Income Tax Credit. *American Economic Review*, [online] 102(5), pp.1927–1956. doi:10.1257/aer.102.5.1927.
- Daniels, D. (2009). *Social Security payments for people caring for children, 1912-2008: a chronology*. [online] www.aph.gov.au. Available at: https://www.aph.gov.au/About_Parliament/Parliamentary_Departments/Parliamentary_Library/pubs/BN/0809/children.
- de Brauw, A. and Hoddinott, J. (2011). Must conditional cash transfer programs be conditioned to be effective? The impact of conditioning transfers on school enrollment in Mexico. *Journal of Development Economics*, 96(2), pp.359–370. doi:<https://doi.org/10.1016/j.jdeveco.2010.08.014>.

- de Gendre, A., Lynch, J., Meunier, A., Pilkington, R. and Schurer, S. (2021). Child Health and Parental Responses to an Unconditional Cash Transfer at Birth. IZA DP No. 14693 <https://doi.org/10.2139/ssrn.3917308>.
- Delavande, A. and Rohwedder, S. (2008). Eliciting Subjective Probabilities in Internet Surveys. *Public Opinion Quarterly*, Vol. 72, 5, 866–891.
- De Janvry, A., & Sadoulet, E. (2006). Making conditional cash transfer programs more efficient: designing for maximum effect of the conditionality. *The World Bank Economic Review*, 20(1), 1–29.
- Department of Communities and Justice (2021). *About the Human Services Dataset*. [online] Family & Community Services. Available at: <https://www.facs.nsw.gov.au/resources/research/human-services-dataset-hsds/about-the-human-services-dataset>.
- Department of Health (2021). *The Centre for Victorian Data Linkage*. [online] Vic.gov.au. Available at: <https://www.health.vic.gov.au/reporting-planning-data/the-centre-for-victorian-data-linkage>.
- Department of Social Services (2021). *3.6.4 Maternity payments - historical rates | Family Assistance Guide*. [online] guides.dss.gov.au. Available at: <https://guides.dss.gov.au/family-assistance-guide/3/6/4#NoteA> [Accessed 29 Sep. 2022].
- Department of Social Services (2022). *Longitudinal Studies Growing Up in Australia: The Longitudinal Study of Australian Children*. [online] Available at: https://www.dss.gov.au/sites/default/files/documents/08_2022/longitudinal-studies-lsac-factsheet-2022-aug_0.pdf.
- Deutscher, N. and Breunig, R. (2017). Baby Bonuses: Natural Experiments in Cash Transfers, Birth Timing and Child Outcomes. *Economic Record*, 94(304), pp.1–24. doi:10.1111/1475-4932.12382.
- DiNardo, J., McCrary, J., & Sanbonmatsu, L. (2006). Constructive proposals for dealing with attrition: An empirical example. Working paper, University of Michigan.
- Doiron, D. (2004). Welfare Reform and the Labour Supply of Lone Parents in Australia: A Natural Experiment Approach. *The Economic Record*, 80(249), pp.157–176.
- Duflo, E. (2003). Grandmothers and granddaughters: old-age pensions and intrahousehold allocation in South Africa. *The World Bank Economic Review*, 17(1), 1–25.
- Duflo, E., Glennerster, R., & Kremer, M. (2007). Using randomisation in development economics research: A toolkit. *Handbook of Development Economics*, 4, 3895–3962.
- Elias, M. (2022). Constructing real time estimates of Australian consumer spending using bank transactions. e61 Institute. [online] Available at: <https://www.e61.in/e61spendtracker>.
- Engberg, J., Epple, D., Imbrogno, J., Sieg, H., & Zimmer, R. (2014). Evaluating education programs that have lotteried admission and selective attrition. *Journal of Labor Economics*, 32(1), 27–63.
- FACSIAR (2021a). *Data Item List: Guidelines for the access to and use of the Human Services Dataset*. [online] Department of Communities and Justice. Available at: <https://www.facs.nsw.gov.au/download?file=813632>.
- FACSIAR (2021b). *Guidelines for the access to and use of the Human Services Dataset*. [online] Department of Communities and Justice. Available at: <https://www.facs.nsw.gov.au/download?file=813632>.
- Ferlitsch, P. (2022). Changes to Australian income support settings during the COVID-19 pandemic. *TPI - Working Paper*. [online] Available at: <https://www.austaxpolicy.com/news/tpi-working-paper-changes-to-australian-income-support-settings-during-covid-19/>.
- Gans, J.S. and Leigh, A. (2009). Born on the first of July: An (un)natural experiment in birth timing. *Journal of Public Economics*, [online] 93(1-2), pp.246–263. doi:10.1016/j.jpubeco.2008.07.004.
- Garbarino, S., & Holland, J. (2009). Quantitative and qualitative methods in impact evaluation and measuring results.
- Gertler, P.J., Martinez, S., Premand, P., Rawlings, L.B. and Vermeersch, C.M.J. (2016). *Impact Evaluation in Practice*, Second Edition. World Bank. [online] doi:10.1596/978-1-4648-0779-4.

- Giovagnoli, A. (2021). The Bayesian Design of Adaptive Clinical Trials. *International Journal of Environmental Research and Public Health*, 18(2), p.530. doi:10.3390/ijerph18020530.
- Glennester, R., & Takavarasha, K. (2013). Running randomised evaluations. In *Running Randomised Evaluations*. Princeton University Press.
- Hamilton, S., Liu, G. and Sainsbury, T. (2023). Early Pension Withdrawal as Stimulus. TTPI Working Paper 3/2023.
- Heckman, J. and Karapakula, G. (2019). Intergenerational and Intragenerational Externalities of the Perry Preschool Project. *NBER Working Paper Series*, Working Paper 25889. doi:10.3386/w25889.
- Heckman, J., Moon, S., Pinto, P., Savelyev, P., Yavitz, A., et al (2010), 'Analyzing Social Experiments as Implemented: A Reexamination of the Evidence from the High Scope Perry Preschool Program', *Quantitative Economics*, 1:1, 1-46
- Huber, M. (2013). A simple test for the ignorability of non-compliance in experiments. *Economics Letters*, 120(3), 389-391.
- Imbens, G.W. and Wooldridge, J.M. (2009). Recent Developments in the Econometrics of Program Evaluation. *Journal of Economic Literature*, [online] 47(1), pp.5-86. Available at: <https://www.jstor.org/stable/27647134>.
- Jimenez, E., Waddington, H., Goel, N., Prost, A., Pullin, A., White, H., ... & Narain, A. (2018). Mixing and matching: using qualitative methods to improve quantitative impact evaluations (IEs) and systematic reviews (SRs) of development outcomes. *Journal of Development Effectiveness*, 10(4), 400-421.
- Kasy, M. and Sautmann, A. (2021a). *Adaptive experiments for policy research*. [online] voxdev.org. Available at: <https://voxdev.org/topic/methods-measurement/adaptive-experiments-policy-research>.
- Kasy, M. and Sautmann, A. (2021b). Adaptive Treatment Assignment in Experiments for Policy Choice. *Econometrica*, 89(1), pp.113-132. doi:10.3982/ecta17527.
- Kettlewell, N. and Siminski, P. (2022). Optimal Model Selection in RDD and Related Settings Using Placebo Zones. *Life Course Centre Working Paper*, 21. doi:10.2139/ssrn.3690751.
- Klapdor, M. (2013). *Abolishing the Baby Bonus*. [online] www.aph.gov.au. Available at: https://www.aph.gov.au/About_Parliament/Parliamentary_Departments/Parliamentary_Library/pubs/rp/BudgetReview201314/BabyBonus#:~:text=Background.
- Kremer, M. (2003). Randomised evaluations of educational programs in developing countries: Some lessons. *American Economic Review*, 93(2), 102-106.
- Kusek, J. Z., & Rist, R. C. (2004). Ten steps to a results-based monitoring and evaluation system: a handbook for development practitioners. World Bank Publications.
- Leigh, A. (2018). *Randomistas: How Radical Researchers Changed Our World*. La Trobe University Press.
- Liu, Y., Mattos, D., Bosch, J., Olsson, H. and Lantz, J. (2022). *Bayesian causal inference in automotive software engineering and online evaluation*. [online] arXiv:2207.00222. Available at: <https://arxiv.org/abs/2207.00222>.
- Ludwig, J. and Miller, D.L. (2007). Does Head Start Improve Children's Life Chances? Evidence from a Regression Discontinuity Design. *The Quarterly Journal of Economics*, 122(1), pp.159-208. doi:10.1162/qjec.122.1.159.
- Ludwig, J., Kling, J. R., & Mullainathan, S. (2011). Mechanism experiments and policy evaluations. *Journal of Economic Perspectives*, 25(3), 17-38.
- Macours, K., & Molina Millan, T. (2017). Attrition in Randomised Control Trials: Using tracking information to correct bias.
- Manski, C.F. (2007). *Identification for Prediction and Decision*, Cambridge: Harvard University Press.
- Miller, S., Johnson, N. and Wherry, L.R. (2021). Medicaid and Mortality: New Evidence From Linked Survey and Administrative Data. *The Quarterly Journal of Economics*, 136(3), 1783-1829
-

- Mogstad, M. and Torgovitsky, A. (2018). Identification and Extrapolation of Causal Effects with Instrumental Variables. *Annual Review of Economics*, 10(1), pp.577–613. doi:10.1146/annurev-economics-101617-041813.
- Morra-Imas, L. G., Morra, L. G., & Rist, R. C. (2009). The road to results: Designing and conducting effective development evaluations. World Bank Publications.
- Moscoe, E., Bor, J. and Bärnighausen, T. (2015). Regression discontinuity designs are underutilized in medicine, epidemiology, and public health: a review of current and best practice. *Journal of Clinical Epidemiology*, [online] 68(2), pp.122–33. doi:10.1016/j.jclinepi.2014.06.021.
- Moffitt, R. (1983). An Economic Model of Welfare Stigma. *American Economic Review*, ol.73 (5), p.1023-1035.
- Mullainathan, S. and Spiess, J. (2017). Machine Learning: An Applied Econometric Approach. *Journal of Economic Perspectives*, 31(2), pp.87–106. doi:10.1257/jep.31.2.87.
- Pallmann, P., Bedding, A.W., Choodari-Oskoei, B., Dimairo, M., Flight, L., Hampson, L.V., Holmes, J., Mander, A.P., Odondi, L., Sydes, M.R., Villar, S.S., Wason, J.M.S., Weir, C.J., Wheeler, G.M., Yap, C. and Jaki, T. (2018). Adaptive designs in clinical trials: why use them, and how to run and report them. *BMC Medicine*, 16(1). doi:10.1186/s12916-018-1017-7.
- Prowse, M. (2007). Aid effectiveness: the role of qualitative research in impact evaluation. Background Note. London: ODI.
- Rao, V., & Woolcock, M. (2003). Integrating qualitative and quantitative approaches in program evaluation. The impact of economic policies on poverty and income distribution: Evaluation techniques and tools, 165-190.
- Roll, S., Constantino, S., Hamilton, L., Miller, S., Bellisle, D. and Despard, M. (2022). *How Would Americans Respond to Direct Cash Transfers? Results from Two Survey Experiments*. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4136559.
- Rosenman, E., Rajkumar, K., Gauriot, R. and Slonim, R. (2021). Optimized Partial Identification Bounds for Regression Discontinuity Designs with Manipulation. arXiv:1910.02170 [stat]. [online] Available at: <https://arxiv.org/abs/1910.02170>.
- SA-NT DataLink (2022). Available Datasets | SA-NT DataLink | Supporting health, social and economic research, education and policy in South Australia and the Northern Territory. [online] www.santdatalink.org.au. Available at: https://www.santdatalink.org.au/available_datasets.
- Schady, N., Araujo, M. C., Peña, X., & López-Calva, L. F. (2008). Cash transfers, conditions, and school enrollment in Ecuador [with Comments]. *Economía*, 8(2), 43-77.
- Schady, N. R., & Araujo, M. (2006). Cash transfers, conditions, school enrollment, and child work: Evidence from a randomised experiment in Ecuador (Vol. 3). World Bank Publications.
- Shoda, Y., Mischel, W., & Peake, P. K. (1990). Predicting adolescent cognitive and self-regulatory competencies from preschool delay of gratification: Identifying diagnostic conditions. *Developmental Psychology*, 26(6), 978–986. <https://doi.org/10.1037/0012-1649.26.6.978>.
- Stock, J.H. and Watson, M.W. (2019). *Introduction to Econometrics*. 4th Edition. Pearson.
- Summerfield, M., Garrard, B., Hahn, M., Jin, Y., Kamath, R., Macalalad, N., Watson, N., Wilkins, R. and Wooden, M. (2021). *HILDA User Manual -Release 20*. [online] Available at: <https://melbourneinstitute.unimelb.edu.au/hilda/for-data-users/user-manuals>.
- Thistlethwaite, D.L. and Campbell, D.T. (1960). Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational Psychology*, 51(6), pp.309–317. doi:10.1037/h0044319.
- Wright, D. (2021). MADIP -What is it and where is it headed? http://nceph.anu.edu.au/files/0_D%20Wright_The%20MADIP%20Asset.pdf.
- Zimmerman, F. J., Gilkerson, J., Richards, J. A., Christakis, D. A., Xu, D., Gray, S., & Yapanel, U. (2009). Teaching by listening: The importance of adult-child conversations to language development. *Pediatrics*, 724(1), 342-349. doi:10.1542/peds.2008-2267.

Appendix 1 Biographies of Key Personnel

Peter Siminski

Peter Siminski is a Professor and Head of the Economics Department at UTS. His research is in Applied Microeconomics and Microeconometrics, in the fields of inequality and economic mobility, education, health, labour and public economics. Much of his work applies modern impact evaluation techniques to estimate the causal effects of Australian government policies and programs on people's lives. The measurement of inequality and inter-generational economic mobility is a key theme of his work. He has published in leading economics journals such as the *American Economic Review*, *American Economic Journal: Applied Economics*, *Journal of Labor Economics*, and *Review of Economics and Statistics*. He is Associate Editor of the *Economic Record*, and is on the editorial board of *Economics of Education Review*. profiles.uts.edu.au/peter.siminski

Adeline Delavande

Adeline Delavande is an Economics Professor at UTS. She specialises in Applied Economics and Econometrics, including Development Economics, Health Economics, Education Economics and Labour Economics. Her research focuses on understanding how people's subjective beliefs and expectations about future events shape their current decisions in the health, labour market and education spaces. She has made major contributions to survey methodology for the elicitation of such beliefs from individuals, and to economic analysis of the impact of these beliefs on people's behaviour. Adeline has published extensively in top international journals in economics, including the *Review of Economic Studies*, *Journal of Political Economy*, *International Economic Review*, *American Economic Journal: Applied Economics*, *Journal of Economic Behavior & Organization* and *Journal of Applied Econometrics*, as well as in top general-interest journals and top-field journals of other disciplines: *Proceedings of National Academy of Sciences*, *Demography*, *Public Opinion Quarterly*, and *The New England Journal of Medicine*. Adeline has generated about \$13 million in external funding, including highly competitive grants from the UK's Economic and Social Research Council (UK equivalent of ARC), the US's National Institute on Aging, the Higher Education Funding Council for England and other funding agencies. profiles.uts.edu.au/adeline.delavande

Bob Slonim

Bob Slonim is an Economics Professor at UTS. He is recognised as a pioneer in the areas of Experimental and Behavioural Economics. He has published academic research articles on a range of topics including game theory, education, public policy, charitable donations and altruism across several academic disciplines. He has received several internationally competitive grants including multiple National Science Foundation and Australian Research Council grants. He was the co-founding editor of the *Journal of the Economic Science Association* (2015–2020), and currently serves as associate editor at *Management Science*. He was research the Director of Research of the Prime Minister and Cabinet's Behavioural Economics Team of Australia (2016–2017); he serves on multiple government and private sector advisory panels, and has provided expert witness evidence based on his Behavioural Economics expertise for several public sector bodies including the ACCC and the NZ Commerce Commission. profiles.uts.edu.au/robert.slonim

Appendix 2 Designs Used to Create Randomised Variation in Exposure to the Program

Lottery

Units are randomly assigned to treatment and control groups, meaning that only the former can access the program. This design is used when resources are scarce and a clear and fair mechanism is needed to justify the allocation. However, government regulations usually do not permit such an arbitrary mechanism. Another caveat is that attrition levels may be higher than with other designs, as people can drop out if not assigned to a preferred treatment. Ethical issues also arise, as not everyone in need is given access to the program, and it may be harmful for those in need to see people benefiting from the program while they have no access to it. In order to justify using the lottery design, it is important to either randomise at a higher level, so that the issues just described do not arise, or to have a clear vision of the benefits of undertaking this research with scarce resources in order to later use the evidence to expand the program to everyone in the eligible population (Glennerster & Takavarasha, 2013).

Lottery around a cut-off

Programs that select participants based on their characteristics cannot use a regular lottery, so it is necessary to divide potential applicants into sets: those who do not qualify for the program and will not be accepted under any circumstances, those who must be prioritised for acceptance, and those who are not prioritised but can be accepted. People in the third set are randomly assigned to treatment and control groups. Allocation to one of these groups thus depends on a person's characteristics. Such randomisation around the cut-off can be used for programs providing scholarships to students, loans to individuals, welfare programs to those in need, etc. Versions of the lottery around the cut-off include:

1. **Lottery among the marginally ineligible** (program is expanded to include those who previously just missed the cut-off evaluating the impact of the program's eligibility expansion).
2. **Lottery among the marginally eligible and marginally ineligible** (in the case of scarce resources, the program assigns the lottery to those who were previously eligible and those who were previously just below the cut-off evaluating the impact of the program at the margin).
3. **Lottery among the qualified** (program assigns a lottery to eligible individuals, but there is also a group of ineligible people, making it possible to test the impact of the program on the average participant).

Overall, this design evaluates the effect of the program around the cut-off, which is especially important if the question being grappled with is whether to expand the program to those currently ineligible. The range around the cut-off depends on the needs of the program, its statistical power, and ethical and political considerations (Glennerster & Takavarasha, 2013).

Phase-in Design

When all of the eligible population must receive the treatment, it is possible to randomise who will be phased in first and who will receive the treatment later, for now forming a control group. When phased in, participants join the treatment group. This design is used when everyone must eventually get access to treatment, but not everyone can be enrolled at the same time. However, there is a constraint: anticipation of treatment should not change the behaviour of the control group. Hence the effect is measured as an average over different years. This design frequently involves lower levels of attrition, as anticipation of treatment increases people's willingness to cooperate. Nonetheless, anticipation may itself affect the outcomes, meaning that the control group cannot be considered the 'pure' counterfactual for comparison. Furthermore, the evaluation will identify only the effects of the program which are manifested before the last phase-in (Glennerster & Takavarasha, 2013).

Rotation Design

Rotation design is also used when the program cannot be immediately assigned to everyone who is eligible. Participants are divided randomly into two groups, and the groups take turns in receiving the treatment. This design compares the effects of the program during the time that treatment is being received, on the assumption that any effects vanish when treatment is suspended. Rotation design is also useful for measuring seasonal influences, or the effects of a length of exposure to the treatment, since treatment status varies over time (Glennerster & Takavarasha, 2013).

Encouragement Design

Encouragement design is suitable for open-access programs with no potential for randomisation and a low take-up rate. Encouragement, like treatment, is assigned randomly, and the estimated difference between the 'encouragement' and 'no encouragement' groups is created as a higher proportion in the former takes up the treatment. Overall, this design is used when the program is accessible but undersubscribed, and when encouragement does not directly affect the outcomes. In this context, encouragement is essentially an instrumental variable (IV), generating the variation needed to estimate the effects (Glennerster & Takavarasha, 2013).

Stratified and Pairwise Randomisation

As previously mentioned, random assignment should ensure that treatment and control groups are statistically identical, with equivalent average characteristics. Stratified random assignment is a way to meet this condition. First, the eligible population is divided into strata, with each stratum falling under the terms of simple random assignment. For example, there is a pool of high-school students, 100 girls and 60 boys, half of them from a central district and half from a remote area. As a first step, they are divided according by area and gender, creating 4 groups. The second step is to randomly assign half of each group to cell A and the other half to cell B. Finally, cells A and B are randomly assigned to be a treatment/control group. By stratifying, the treatment/control groups are assured to be balanced by area and gender.

Stratified randomisation guarantees this balance, which is important when the sample size is small. It also increases statistical power, since the levels of stratification are chosen as strong predictors of the studied outcomes. These levels should be discrete variables, in order for each group to have a substantial number of observations. The variables should highly correlate with the outcomes studied, or even constitute the outcome variables themselves, but measured at the baseline. Stratified randomisation also enables analysis by sub-groups, making it possible to analyse the effect of the program on different sub-groups of the population.

The number of variables that ideally should be selected as stratification levels is a multiple of the number of randomisation cells, meaning there are no 'leftovers'. Generally, the fewer the variables, the easier it is to achieve a balance on them all. A greater number of variables will lead to a loss of degrees of freedom in the final analysis, and will also render comparisons unviable if there are too few units in the cells. Therefore it is important to choose a smaller number of more key variables.

Pairwise random assignment involves matching two units based on their characteristics and assigning them randomly to treatment and control groups. Here, pairing is possible on a continuous variable, as only two units are needed for each value. This method is an extreme version of stratified randomisation, and is often used when the sample size is small. Attrition is a more significant threat, since the loss of one unit within a pair will mean excluding the whole pair from the analysis (Glennerster & Takavarasha, 2013).





Appendix 3 International Literature Review: Cash Transfers

This Appendix reviews studies on the effects of conditional and unconditional cash transfers programs on different outcomes, related to schooling, enrolment, dropouts and other (for children), and to work behaviour, healthcare and other (for adults). By conditional cash transfers we mean transfers that can be received by an eligible population based on certain characteristics, and also subject to their adherence to certain rules for the duration of the program. By contrast, unconditional cash transfers do not require adherence to any rules by an eligible population, nor do they involve a complicated process of entering the program.

The Appendix is divided into three sections, as follows:

Section	Number of Studies
1: Conditional Treatment	5
2: Unconditional Treatment (+ Labelled)	4
3: Both Conditional and Unconditional Treatment	3

The following colour scheme is used to designate:

-  Blue: RCTs (Randomised Controlled Trials).
-  Red: RDD (Regression Discontinuity Design).
-  Green: DD (Difference in Differences).
-  Grey: Other methods.

Section 1: Conditional Treatment

Citation	Location of Study	Beneficiaries / unit of observation	Conditionality Basis	Value of Cash Transfer, Frequency	Outcomes studied	Methodology	Abstract (Key findings)	Attrition/ Compliance
Álvarez, Devoto & Winters (2008, World Development)	Mexico	Poor rural children / child (conditional on health check-ups and attending health/nutrition lectures, school attendance)	Registration and 85% monthly attendance rate, health check-ups	Education: from \$10.50 to \$66 monthly. Health: healthcare providers to visit. Nutrition: \$15.50 monthly. Overall: 20% of total household expenditures. Payments received every 2 months. Duration: primary school – end of high school.	Dropouts	Discrete duration model	This paper analyzes the characteristics of beneficiaries who drop out of the Mexican conditional cash transfer program Oportunidades to determine if dropping out of the program is a result of self-targeting by the non-poor or the exclusion of the target poor population. Using Oportunidades administrative data and a discrete duration model, the analysis indicates that wealthier beneficiaries have greater odds of dropping out, suggesting that conditionality acts as a screening device. The results also indicate that administrative factors and the provider of health services to beneficiaries also have an important influence on whether beneficiaries remain in or leave the program.	8-27% overall attrition rate/0.5% bimonthly dropout rate under conditionality
Attanasio, Oppedisano & Vera-Hernández (2015, American Economic Journal: Applied Economics)	Colombia	Low-income families with children (mothers) /child (conditional on health and education activities)	Number of healthcare centre visits and school attendance	Education: \$7 a month for primary school children, \$14 a month for secondary school. Health: \$15 a month. Total: approximately 24% of total household expenditure. Payments received every 2 months. Duration: birth – 7 years old.	Child health outcomes, preventive care visits.	RDD, 2SLS (date of birth as IV)	“We study a Conditional Cash Transfer program in which the cash transfers to the mother only depends on the fulfilment of the national preventive visit schedule by her children born before she registered in the program. We estimate that preventive visits of children born after the mother registered in the program are 50% lower because they are excluded from the conditionality requirement. Using the same variation, we also show that attendance to preventive care improves children’s health.”	87% of eligible households registered/0.2% of the beneficiaries were suspended overall
Attanasio, Sosa, Medina, Meghir & Posso-Suárez (2021, NBER)	Colombia	Low-income families conditional on children’s school attendance/ adolescents	Number of healthcare centre visits and school attendance	Education: \$7 a month for primary school children, \$14 a month for secondary school. Health: \$15 a month. Total: approximately 24% of total household expenditure. Payments received every 2 months. Duration: birth – 7 years old.	Crime, teenage pregnancy, high school dropout, tertiary education.	RDD (fuzzy)	“Conditional Cash transfer (CCT) programs have been shown to have positive effects on a variety of outcomes including education, consumption and health visits, amongst others. We estimate the long-run impacts of the urban version of Familias en Acción, the Colombian CCT program on crime, teenage pregnancy, high school dropout and college enrollment using a Regression Discontinuity Design on administrative data. ITT estimates show a reduction on arrest rates of 2.7pp for men and a reduction on teenage pregnancy of 2.3pp for women. High school dropout rates were reduced by 5.8pp and college enrolment was increased by 1.7pp for men. ”	87% of eligible households registered/0.2% of the beneficiaries were suspended overall

Citation	Location of Study	Beneficiaries / unit of observation	Conditionality Basis	Value of Cash Transfer, Frequency	Outcomes studied	Methodology	Abstract (Key findings)	Attrition/ Compliance
De Janvry & Sadoulet (2006, The World Bank Economic Review)	Mexico	Low-income rural families/child	School attendance 85% and set of behaviours designed to improve health and nutrition	\$70-\$255 per year (primary/secondary school) paid monthly over 3 years. Total: over 28% of average monthly per capita expenditure.	Enrolment	Predictive model of the probability of attending school	Conditional cash transfer programs are now used extensively to encourage poor parents to increase investments in their children's human capital. These programs can be large and expensive, motivating a quest for greater efficiency through increased impact of the programs'-imposed conditions on human capital formation. This requires designing the programs' targeting and calibration rules specifically to achieve this result. Using data from the Progresa randomised experiment in Mexico, this article shows that large efficiency gains can be achieved by taking into account how much the probability of a child's enrollment is affected by a conditional transfer. Rules for targeting and calibration can be made easy to implement by selecting indicators that are simple, observable, and verifiable and that cannot be manipulated by beneficiaries. The Mexico case shows that these efficiency gains can be achieved without increasing inequality among poor households.	23.4% attrition rate/12% dropout rate for the analysed sample
Gertler (2004, American Economic Review)	Mexico	Low-income rural families/child	School attendance 85% and set of behaviours designed to improve health and nutrition	\$70-\$255 per year (primary/secondary school) paid monthly over 3 years. Total: over 28% of average monthly per capita expenditure.	Morbidity, height, anemia	RCT, comparing mean outcomes; logit	"I found a significant improvement in the health of children in response to PROGRESA. Specifically, children born during the two-year intervention to families benefiting from the program experienced an illness rate in the first six months of life that was 25.3 percent lower than that of control children. Treatment children aged 0-35 months at baseline experienced a reduction of 39.5 percent in their illness rates after 24 months in the program. Moreover, the effect of the program seems to increase the longer the children stayed on the program, suggesting that program benefits were cumulative. I also found that treatment children were 25.3 percent less likely to be anemic and grew about 1 centimeter more during the first year of the program. While these results suggest that PROGRESA has had a positive effect on child health, they do not indicate which aspects of this complex program really matter. PROGRESA combines large cash transfers with requirements that individuals engage in a number of preventive health and nutrition activities. One cannot tell if the same results could have been achieved with just a large cash transfer and no behavioral requirements. It is also hard to distinguish between the relative effects of compliance with the various requirements. Answers to these questions would facilitate a better package and therefore improve the cost-effectiveness of the intervention."	5.1-5.5% dropout rate for the studied sample/7% attrition rate for survey responses

Section 2: Unconditional Treatment

Citation	Location of Study	Beneficiaries / unit of observation	Value of Cash Transfer, Frequency	Outcomes studied	Method	Abstract (Key findings)	Attrition/ Compliance
Duflo (2003, The World Bank Economic Review)	South Africa	Men and women over 50 in poor rural areas/child	Pension program: twice the median income per capita	Health and nutrition	2SLS (receiving pension by men and women in the HH as IV)	“This article evaluates the impact of a large cash transfer program in South Africa on children’s nutritional status and investigates whether the gender of the recipient affects that impact. In the early 1990s the benefits and coverage of the South African social pension program were expanded for the black population. In 1993 the benefits were about twice the median per capita income in rural areas. More than a quarter of black South African children under age five live with a pension recipient. Estimates suggest that pensions received by women had a large impact on the anthropometric status (weight for height and height for age) of girls but little effect on that of boys. No similar effect is found for pensions received by men. This suggests that the efficiency of public transfer programs may depend on the gender of the recipient. ”	-
Barr, Eggleston & Smith (2022, The Quarterly Journal of Economics)	US	Lower-income families with a newborn/adult	Average tax benefit provided by a child - \$1300 (10% of income)	Tracked young adult outcome in test scores and later earnings	RDD (based on date of birth cut-off)	“We provide new evidence that cash transfers following the birth of a first child can have large and long-lasting effects on that child’s outcomes. We take advantage of the January 1 birthdate cutoff for U.S. child-related tax benefits, which results in families of otherwise similar children receiving substantially different refunds during the first year of life. For the average low-income single-child family in our sample this difference amounts to roughly \$1,300, or 10 percent of income. Using the universe of administrative federal tax data in selected years, we show that this transfer in infancy increases young adult earnings by at least 1 to 2 percent, with larger effects for males. These effects show up at earlier ages in terms of improved math and reading test scores and a higher likelihood of high school graduation. The observed effects on shorter-run parental outcomes suggest that additional liquidity during the critical window following the birth of a first child leads to persistent increases in family income that likely contribute to the downstream effects on children’s outcomes. The longer-term effects on child earnings alone are large enough that the transfer pays for itself through subsequent increases in federal income tax revenue. ”	-
Benhassine, Devoto, Duflo, Dupas & Pouliquen (2015, American Economic Journal: Economic Policy)	Morocco	Fathers in poor communities; parents of primary age children/child	\$8-\$13 per month (depending on age) for 2 years. Total: from 3% to 5% of monthly HH consumption, but compensating the direct costs of schooling.	Variables on schooling and performance at school	RCT	“Conditional Cash Transfers (CCTs) have been shown to increase human capital investments, but their standard features make them expensive. We use a large, randomised experiment in Morocco to estimate an alternative government-run program, a “labeled cash transfer” (LCT): a small cash transfer made to fathers of school-aged children in poor rural communities, not conditional on school attendance but explicitly labeled as an education support program. We document large gains in school participation. Adding conditionality and targeting mothers make almost no difference. The program increased parents’ belief that education was a worthwhile investment, a likely pathway for the results. ”	97% take-up rate

Citation	Location of Study	Beneficiaries / unit of observation	Value of Cash Transfer, Frequency	Outcomes studied	Method	Abstract (Key findings)	Attrition/ Compliance
<u>Baby's First Years</u>	US	Mothers after giving birth at hospitals in four metropolitan areas/child	Mothers receive either (1) \$333 each month (\$4,000 each year), or (2) \$20 each month (\$240 each year), for the first 52 months of the children's lives.	Child's health and development, maternal health, family income, family life	RCT	<p>To understand how poverty reduction affects children's development and family life, quantitative data will be collected on or around the children's first, second, third, and fourth birthdays. Each wave of data collection will capture:</p> <ul style="list-style-type: none"> • Aspects of family life hypothesized to be affected by poverty, including parent stress, family expenditures, family routines, parents' time use and parenting practices, and child care arrangements. • Children's development, as well as their physical health, stress, and behavior. <p>In addition, qualitative semi-structured interviews are conducted with 80 randomly selected mothers in two of the four study sites. There will be four rounds of interviews over the first four years of the focal child's life.</p> <p>The study is designed to produce strong and clear evidence about the magnitude and pathways of causal connections between family income and early childhood development. Beyond its core contributions to science, the study will provide important evidence about the likely effects of tax and income-enhancement policies for young children, such as the Child and Earned Income Tax Credits, and related social policies designed to enhance family economic stability and well-being.</p>	-

Section 3: Both Conditional and Unconditional Treatment

Citation	Location of Study	Beneficiaries / unit of observation	Conditionality Basis	Value of Cash Transfer, Frequency	Outcomes studied	Methodology	Abstract (Key findings)	Attrition/ Compliance
Baird, McIntosh & Özler (2011, The Quarterly Journal of Economics)	Malawi (East Africa)	Adolescent girls	Regular school attendance 80%	Parents: \$4, \$6, \$8 and \$10 per month. Adolescent girls: \$1, \$2, \$3, \$4 or \$5 per month. Total of approximately \$10 is 10% of the average household monthly consumption. School fees paid to one RCT arm. Total duration: 3 years.	Schooling, marriage & fertility	Linear probability model (choice), single difference impact regressions, ind. fixed effects, baseline controls for balance.	“Conditional Cash Transfer programs are “...the world’s favorite new anti-poverty device,” (The Economist, July 29 2010) yet little is known about the specific role of the conditions in driving their success. In this paper, we evaluate a unique cash transfer experiment targeted at adolescent girls in Malawi that featured both a conditional (CCT) and an unconditional (UCT) treatment arm. We find that while there was a modest improvement in school enrollment in the UCT arm in comparison to the control group, this increase is only 43% as large as the CCT arm. The CCT arm also outperformed the UCT arm in tests of English reading comprehension. The schooling condition, however, proved costly for important non-schooling outcomes: teenage pregnancy and marriage rates were substantially higher in the CCT than the UCT arm. Our findings suggest that a CCT program for early adolescents that transitions into a UCT for older teenagers would minimize this trade-off by improving schooling outcomes while avoiding the adverse impacts of conditionality on teenage pregnancy and marriage.”	Over 90% tracking rate
Banerjee, Hanna, Kreindler & Olken (2017, The World Bank Research Observer)	Different developing countries	Low-income households/men and women within the households	Different conditions for different programs	Transfer amounts and frequency vary for different programs	Work behaviour	RCT (18) + pooling with Bayesian hierarchical model (comparing programs)	“Targeted transfer programs for poor citizens have become increasingly common in the developing world. Yet, a common concern among policy makers and citizens is that such programs tend to discourage work. We re-analyze the data from 7 randomised controlled trials of government-run cash transfer programs in six developing countries throughout the world, and find no systematic evidence that cash transfer programs discourage work. ”	-
Schady, Araujo, Peña & López-Calva (2008, Economía)	Ecuador	Families with school-aged children/child	School enrolment (but actually not enforced)	\$15 monthly transfers (2 years). Total: 9% of the pre-transfer expenditures of the median household in the study sample.	Enrolment, change in enrolment	RCT + change in enrolment b/w baseline and follow-up. Bias-adjusted matching estimator (Abadie & Imbens); reweighting scheme for the data, propensity score, DD.	“We compare the impact of the program among conditioned households (that is, those who told survey enumerators that school enrollment was a BDH requirement) and unconditioned households (those who told enumerators that there was no enrollment requirement attached to transfers). Our estimates show that program effects on enrollment are only significant among conditioned households. Because exposure to the information campaign was not assigned randomly, these comparisons are not experimental. However, the effects we estimate are insensitive to adding a large number of controls, trimming the data, and sweeping out fixed differences between conditioned and unconditioned households. We therefore argue that the larger program effect among conditioned households most likely has a causal interpretation. These results complement evidence from a variety of structural and microsimulation models for Mexico and Brazil, all of which conclude that conditions attached to transfers explain the bulk of the effect of conditional cash transfer programs on school enrollment. ”	Take-up amount of 78% (lack of information, the cost of travelling to a bank, and stigma)/ 42% of control group received treatment/94% reinterviewed

Appendix 4 Glossary

Attrition

The phenomenon of participants dropping out or withdrawing from a study or program over time. It can introduce bias if the attrition is related to the outcomes being measured, and it is important to consider and address attrition in the analysis.

Causal evaluation methods

Techniques used to estimate causal effects of a program or intervention on an outcome or outcomes. They aim to isolate causal links, as distinguished from merely identifying associations.

Compliance

The degree to which participants adhere to or follow the assigned treatment or program. It is a crucial factor in randomised experiments as it can affect the validity and interpretation of the results.

External validity

External validity relates to the generalizability or applicability of the findings from a study to a broader population or real-world settings beyond the specific study context.

Internal validity

Internal validity refers to the extent to which a study is able to establish a causal relationship between the program or treatment and its observed effects, ruling out alternative explanations or confounding factors within the study design.

Level of randomisation

The unit of analysis where randomisation is applied within an experiment. For example, randomising at the level of the individual, family, class, school, or some other unit.

Random assignment of treatment

Randomly assigning units of the study sample to different treatment conditions or groups. In its simplest form, such randomisation ensures that each unit has an equal chance of being assigned to any given group, helping to eliminate potential biases or confounding factors that could influence the results.

Randomised experiments

Also known as randomised control trials, these are studies where participants or subjects are randomly assigned to different groups or conditions, typically a treatment group(s) and a control group. This random assignment helps minimize biases and allows for causal inferences to be made about the impact of a particular program or intervention.

Spillovers

In the context of evaluation studies, spillovers (also known as externalities) are the (sometimes unintended) effects or impacts of the program that extend beyond the treated individuals or groups, affecting others who were not directly part of the program. These effects can be either positive or negative.

Statistical power

The probability of correctly detecting an effect or relationship in a statistical test. It represents the ability of a study to detect a true effect, given a specific sample size, effect size, and significance level. A study with high statistical power has a greater chance of detecting a real effect if it exists.

Statistical power calculation

Power calculation is a statistical technique used to determine the required sample size for a study in order to achieve a desired level of statistical power. It involves considering factors such as the expected effect size, desired level of significance, and anticipated variability in the data. By conducting a power calculation, researchers can estimate the sample size needed to have a reasonable chance of detecting the effect they are interested in.

About UTS Business School

The world is in a period of unprecedented change, and that includes changing expectations of business and industry to take more of a central role in addressing the critical social, political and economic issues of which they are a part.

At UTS Business School, we work closely and collaboratively with businesses, policymakers and public institutions, and our community to produce socially responsible and economically fair outcomes, and use education and research as a pathway to individual mobility, social diversity and economic equality.

We are internationally recognised for our innovative research, with our academics taking a comprehensive, interdisciplinary approach – bringing together skills and knowledge from diverse fields – to develop high-impact research that is not just characterised by scholarly excellence but also makes a meaningful contribution to the public good.

About Paul Ramsay Foundation

The late Paul Ramsay AO established the Paul Ramsay Foundation (PRF) in his name in 2006 and, after his death in 2014, left most of his estate to continue his philanthropy for generations to come.

At PRF, we believe in a world where all people can live their best lives. Our purpose is to help end cycles of disadvantage in Australia by enabling equitable opportunity for people and communities to thrive.

As one of the largest philanthropic foundations in Australia, we take our social responsibility seriously and aim to make a lasting contribution to positive change.

For more information

UTS Business School

Web: business.uts.edu.au

Email: Adeline.Delavande@uts.edu.au
Peter.Siminski@uts.edu.au
Robert.Slonim@uts.edu.au