# Complements and Alternatives to RCTs in Evaluation in Development

Orazio Attanasio[1], Elisa Cavatorta[2]

August 2017




1 Department of Economics, University College London, Institute for Fiscal Studies, NBER and CEPR
2 Department of Political Economy, King's College London and Institute for Fiscal Studies

# Colophon

Suggested Citation: Attanasio O, Cavatorta E, 2017 Complements and Alternatives to RCTs in Evaluation in Development. CEDIL Pre-Inception Paper: London

About CEDIL: The Centre of Excellence for Development Impact and Learning (CEDIL) is an academic consortium initiative supported by UKAID through DFID. The objective of the centre is to commission and implement impact evaluations, promote the uptake and use of evidence from impact evaluations, and develop and demonstrate new and innovative methodologies for impact evaluation and evidence accumulation. Centre for the Evaluation of Development Policies, Institute for Fiscal Studies (EDePo) is one of a number of CEDIL members that form its intellectual leadership team.

Corresponding Author: Prof Orazio Attanasio, email: o.attanasio@ucl.ac.uk

# Table of Contents

# Abstract

Randomised control trials (RCTs) have attracted increasing attention in evaluation of public policy. They have become some sort of gold standard to determine “what works” in international development. However, in some cases, the key feature of RCTs – the random allocation of treatment(s) – may be unethical, politically infeasible, or simply impossible. This paper discusses the main barriers to answer causal questions of key interest in policy evaluation. It then reviews approaches complement to RCTs which can be used to answer evaluation questions in conjunction with RCTs and alternatives useful whenever RCTs are not applicable. The paper discusses issues in drawing causal inference from these alternative approaches and provides some examples from specific sectors where standard RCTs are less likely to be applicable.

Section 1

# Introduction

Over the last 25 years, the identification of causal relationships has received considerable attention in economics. Many economists have warned against the dangers of using observational data of various nature, without a proper consideration of what can explain an observed correlation between two or more variables. The fact that such correlations might have nothing to do with causation has been stressed many times, and rightly so.

The extensive debate on these issues, however, indicates that the profession has not reached a consensus. On the one hand, there are scholars who think that theoretical models can be useful to put some structure on a given problem and to construct a logical framework that can be used to exploit certain sources of variation to identify causal links. These links, however, depend, obviously on the validity of the models used to identify them. On the other hand, there are economists who have pushed the almost exclusive use of randomised controlled trials to identify causal links, especially when evaluating policy interventions. In some situations, one can appeal to features of the data that generate a random allocation of a certain 'treatment'. One well-known example, for instance, is the paper by Angrist (1990) who studies the effect of military service on earnings. Or one can appeal to variation induced by what have been called 'natural experiments'.

The issues are not simple, both to justify the use of 'natural' experiments (see, for instance, the discussion in Rosenzweig and Wolpin, 2000) and structural theoretical models. In addition, much of the debate has considered contributions to this debate in other disciplines, from medicine, to epidemiology, psychology and philosophy, only marginally.

Randomised control trials (RCTs from now on) have played a big role in this debate, as they constitute an attractive tool that can be implemented to identify in a conceptually rigorous fashion the causal impact of certain phenomena, in particular, policy interventions. RCTs have attracted considerable attention in international development, partly because it is relatively easy and cheap to run one in a developmental context, partly because some of the questions addressed in development might be better suited to such an instrument. They are often considered as the conveyors of accurate, objective and data-driven knowledge of "what works" in international development.

This impetus of using RCTs in development has certainly been a positive development, as it forces empirical researchers to think hard about the variation that can be used to identify legitimately causal effects. However, the indiscriminate and exclusive use of RCTs can present several problems. First of all, the type of questions answered by RCTs can be very narrow and may be of limited usefulness for the design of effective policies. For the latter, understanding the mechanisms behind an observed impact is very important and a pure comparison of means, or even of distributions might not be

enough. Second, what is observed in a given specific context does not necessarily apply to different situations or even to slightly different interventions. This issue, which is usually referred to as that of the limited external validity of RCTs, can be very relevant in many contexts. Third, there might be situations where it is simply not possible to run an RCT; and yet it is possible to provide an informed analysis of existing and new data sources that, together with the appropriate conceptual framework can be useful for the design of policies.

These remarks indicate that there is more in policy evaluation than the mere identification and estimation of impacts of an intervention or new policy on certain outcomes. In this paper, we discuss these issues and refer to some of the existing literature. It should be stressed, however, that this is not an exhaustive survey paper. It rather conveys our views on the set of issues we have been discussing, with a particular emphasis on the developmental context.

Section 2

# Barriers to Establishing Causal Claims

## 2.1 The Presence of Confounders

In its minimum form, the first causal claim one may wish to make is "intervention X causes outcome Y". This consists of solving an attribution problem: one would like to establish that the observed changes on a given subject are due, exclusively so, to the intervention. To do so, ideally one would like to observe the same subject targeted by a given intervention under two scenarios: one scenario in which the subject is exposed to the intervention and one scenario in which the subject is not exposed to the intervention. The scenario in which the subject is not exposed to the intervention represents what would have happened to the subject had she not received the intervention (i.e. the counterfactual scenario to receiving the intervention).[1] The difference in outcomes on the given subject would be the impact of the intervention on a given subject. This ideal setting has one fundamental problem: it is impossible to observe the same subject in two different scenarios – we either observe a subject receiving a given intervention or not receiving it.

The apparently simple solution of inferring causal impact by comparing the outcomes of a given subject before and after the intervention delivers the impact of the intervention only under a strong assumption: nothing else that is relevant to the outcome changes over time except the intervention. This assumption is extremely unlikely to hold in any circumstance as many events can happen as time passes and affect the outcome of

[1] If the subject does not actually receive the intervention, the scenario in which the same subject is exposed to the intervention represents what would have happened to the subject had she received the intervention, that is the counterfactual scenario to not-receiving the intervention.

interest. In light of this problem, it is virtually impossible to identify and estimate the causal impact of an intervention on a given subject. This issue may not be so disappointing because knowing the outcome on a given subject is not particularly useful from a policy-making perspective.

Policy measures cannot be tailored to every individual's needs, they may be more effective for some than for others or even effective for some and not for others. Once the heterogeneity of impacts is recognised, it is clear that some information on the distribution of the impacts of an intervention is a more useful piece of information for policy-makers. One natural summary measure of the distribution of impact is the average impact (i.e., which in the literature is referred to as average treatment effect, ATE), but the median or the mode, as well as the minimum or the maximum impact or the impact within specific subgroups, may be useful in specific circumstances. It may also be important to know the degree of variation of the impact: this requires measures of the spread of the impact distribution, such as the variance of the impact distribution or the interquartile range of the variation.

To estimate the distribution of impacts of a given intervention one needs to compare the distribution of outcomes for the people receiving the intervention to what that distribution would have looked like had the same group not received the intervention (*counterfactual distribution*). If this was possible, one could estimate the average treatment effect by comparing the average outcome of the treated group and the average outcome of the same group when not-treated. The problem is that a group is either treated or not so that only one of these averages is actually observable.

The key challenge in policy evaluation is to find a suitable comparison group such that their outcome would be the same as that of the treated group in absence of the intervention. If such comparison group is found, then it is natural to estimate the average treatment effect by taking the difference between the average outcome $Y_i$ for the group receiving the intervention and the average outcome for the group not receiving the intervention. Let the treatment be $D_i$, let D=1 indicate the group that receives the treatment and D=0 indicate the group who does not receive the treatment. Formally we can write the average treatment effect as

$$ATE = E[Y_i | D=1] - E[Y_i | D=0] \quad (1)$$

where E[.] indicates the average operator and the conditioning is the group over which the average is computed.

Whether the difference can be interpreted as the causal treatment effect depends on the extent to which the assignment of the treatment is related to the outcome of interest. In other words, whether the comparison group provides a good representation of the counterfactual distribution of the outcome of interest for the treatment group. In most cases, this is difficult because the assignment of an intervention depends on characteristics that are related to the outcome or depends on the outcome variable itself. For example, take a job training programme and its effect on earnings. Those who partake in a job training programme are more likely to be more motivated than those who do not partake. There is a correlation between ability and motivation and the likelihood of receiving the treatment. This is usually referred to in the literature as a

selection bias. Ability and motivation are related to the earnings capacity. Had the participants not received the job training, they were likely to do better than those not participating anyhow. Thus, a simple comparison of outcomes between the training's participants and the non-participants would not yield a causal estimate of the job training programme. There is an attribution problem: do any difference in outcomes come from the job training intervention or the higher motivation of those taking part in it? Another example is when interventions are targeted to the most in need. Take a nutrition intervention targeted to the poorest areas. Comparing the poorest areas with the less poor areas not receiving the intervention would not be an accurate comparison because poor and less poor areas are likely to differ in many other characteristics that do matter for the outcome of the intervention.

A random allocation of the treatment among eligible people would solve the attribution problem by eliminating the selection bias. The reason is that a random assignment makes the receipt of the treatment due to only chance and hence that is independent of (observable and unobservable) characteristics that influence the outcome. In sufficiently large samples, the random allocation of the treatment guarantees that those receiving the treatment and those who don't receive it have the same distribution of underlying characteristics influencing the outcome. Their influence will average out in large samples and the only systematic difference between the two groups is the exposure to the intervention.

In spite of this useful feature, random allocation is not always feasible: random allocation may be unethical, politically infeasible, or simply impossible. In the next paragraphs, we briefly summarise the main arguments against the applicability of random assignment. The discussion aims to highlight the conditions under which other methods than RCTs can be applicable to answer causal questions. For an extensive discussion about the barriers to RCTs, we refer the reader to Bonell et al. (2009).

Randomisation is generally infeasible when it would generate unethical circumstances. Protecting study participants' integrity and freedom of choice should always be a primary concern. Ethical implications may also occur when an intervention has side-effects that are not known. This is a common concern in evaluations of health interventions but can also be a concern in social policy interventions (Thomson et al., 2004). Ethical concerns would also warrant against the possibility of creating negative externalities in the context in which the intervention takes place. For example, forms of interpersonal conflict in the local environment as a result of the delivery of some services to some but not others in the same community. Although there are design practices that can alleviate some of these concerns (e.g. cluster randomisation), violations of these ethical standards would make a control group unethical. For an extensive discussion on the ethics of social experimentation, see Humphreys (2015).

Besides the limits imposed by ethics, randomisation can be politically infeasible. In some cases, random allocation of interventions is often opposed on the principle that beneficial interventions should not be denied to anyone in need of it. This renders evaluation difficult in certain contexts, such as in the context of humanitarian interventions.

Finally, in certain situations, random allocation is simply impossible. This is the case in which the aim of the evaluation is a "systemic" phenomenon (e.g. a civil war) or a policy affecting the entire region or country (e.g., institutional reforms, tariff liberalisation, counter-terrorism policies) which raise important questions for social scientists, but are inherently non-random and not suitable for manipulation. Take the example of the impact of violent conflict on people's behaviour. While there are important causal, policy-relevant, questions about the impact of exposure to violent conflict, the inherently non-random phenomenon calls for alternative methods.

## 2.2 Knowledge of the Causal Pathways of Impact

The ambition to isolate and quantify causal pathways of the impact of an intervention, that is the mechanisms through which an intervention delivers the observed impacts, is an important step to equip researchers and policy-makers with knowledge about how a policy intervention works. Being able to answer the (causal) evaluation question "Why does the treatment work?" requires identifying the mechanisms through which the treatment operates to generate the outcome we observed in an RCT. In other words, it requires identifying how the treatment transmits its effects through "mediator" variables. Knowledge of this opens avenues to devise more efficient ways to achieve the same treatment effect. It also offers an important piece of knowledge towards the generalisability of the treatment to other contexts.

Take the example of a psychosocial stimulation programme for young children. Evidence from many of these programmes shows that early interventions of this kind generate improvements in children's cognitive and socio-emotional skills (e.g. see evidence in Jamaica, Gertler et al., 2014; Grantham-McGregor et al., 1991; in Colombia, Attanasio et al., 2014). There are various mechanisms through which the psychosocial stimulation intervention may be effective: the intervention may work by improving the mother's psychological well-being as she feels more supported, or by shifting the parent's beliefs about the importance of interacting with their children, or by improving the quality and time of parents' activities with their kids. All of these are examples of possible mediator factors that might be influenced by the intervention and exert an influence on the outcome, children's skills. Overall, the intervention has an effect in two ways: a *direct* way, in which the treatment (D) influences the outcome directly, $D \rightarrow Y$, and an *indirect* way in which the treatment (D) influences the mediator (M) and the mediator changes the outcome, $D \rightarrow M \rightarrow Y$. Decomposing the impact and understanding how a certain programme works has important implications for the design of policies that might want to scale up a certain intervention, taking into account limited resources.

A proper causal analysis of mediating factors would aim to identify separately these two effects. There are two different casual questions underpinning the direct and indirect effect. Identifying the direct effect of the intervention is equivalent to answering the causal question: "What would happen to the outcome if we were to change the intervention such that the level of the mediator variable remains constant at a particular level?" Identifying the indirect effect requires knowledge of how much the intervention changes the mediator variable and what the impact of the mediator on the outcome is.

Knowing how much the intervention affects the mediator (i.e., the ATE of the treatment on the mediator, D→M) can be recovered with some confidence if the intervention is randomised: the mediator can be seen just as an outcome of the treatment in a typical regression framework. However, the effect of the mediator on the outcome is more difficult to detect. Identifying the impact of the mediator on the outcome is equivalent to answering the question: "What would happen to the outcome if we were to keep the intervention constant while varying the mediator by the amount it would change if the intervention had changed?" This would require manipulating the mediator independently of the treatment. Manipulation of mediators is in principle possible but practically extremely challenging outside a controlled laboratory environment. Humphreys and Weinstein (2009) and Humphreys and Jacobs (2017) provide a discussion of manipulation of mediators.

Without manipulation of the mediator in a way that is independent of the treatment (i.e. exogenous manipulation), it is difficult to estimate the indirect effect properly. The apparently simple regression-based mediation analysis is usually problematic. For example, consider the regression of *Y*, the outcome of interest, on the treatment, D, and one mediator variable, *M*, and an error term, $e_i$:

$$Y_i = a + d\,D_i + b\,M_i + e_i \qquad (2)$$

Assume the treatment *D* is randomly allocated. The variable *M,* however, is not randomly allocated, nor is it a pre-determined fixed characteristic when the treatment is given (an example is parents' beliefs about the importance of interacting with their children in the example above). This makes the variable M effectively an outcome of the treatment itself. Even though the treatment *D* is properly randomised, there is a possibility that the error term is related to the right-hand side variables, introducing selection bias. A correlation between the error term and the right-hand side variables implies that the coefficient estimates of *d* and *b* can be biased and cannot be interpreted as causal estimates.

How can we learn the "causal chain" when randomisation of the mediator is not feasible? One approach to extricate elements of that "causal chain" is to design targeted "mechanism experiments" (Ludwig et al., 2011). These experiments are aimed at testing specific pathways of impact by estimating the impact of *M* changing on the outcome. This type of experiment is selective in scope: their goal is to learn about the pathway(s) of impact, rather than the overall effect of the treatment. This can be useful in relatively established treatments that have proven successful in a variety of situations. Ruling out mechanisms is a useful piece of information.

A second approach is an "implicit mediator analysis" in which there are different classes or intensity of treatments whose attributes affect one (or more) mediators in a known fashion. If the researchers are able to postulate in which ways various treatments change the underlying mediator(s), then factorial designs with random allocation of subjects into different treatment arms can offer some information about the mechanisms of impact.

A third approach is to study more explicitly the process of change of the outcome of interest. This can be done using qualitative case studies that, for example, focus on a selective group of subjects throughout the intervention[2].

The formalisation of theories of how the outcome of interest is generated and influenced by the intervention can be done in various ways. An alternative that is amenable to quantification is the formalisation of how the outcome of interest is generated in a formal structure (a model) which determines how inputs and conditions lead to observable outcomes of interest. With a formalised model and appropriate data and econometric tools, the links of the causal chain can be identified and estimated quantitatively.

## 2.3 Knowledge about Long-Term Effects and the Effects of Scaling-Up

One key evaluation question of interest in policy design is to understand under which conditions the intervention will work. A typical RCT is a small-scale, local and time-limited intervention. However, most often than not, the policy effects we would like to learn about are the effects of a policy on beneficiaries if the policy was permanent and applied to the entire population of interest.

Many interventions may work differently when applied at an aggregate scale, for example, an entire region or country. This is because many interventions alter supply and demand, which leads to changes in prices against which people respond and make economic decisions. One example is job training programmes. These programmes change the supply of skilled versus unskilled labour. As the job training coverage increases, skilled labour becomes readily available, and generates downwards pressure on wages as more skilled people are competing for jobs. The intervention may affect also those who haven't participated in it: since wages of high-skilled labour decrease, workers who have acquired higher skills without the training will be worse off than before the intervention. This crowding-out effect is a negative externality of increasing labour market participation in the high-skill sector due to the introduction of the job training programme.

A typical RCT would not capture these effects. Wages adjust only if the supply of high-skill labour is sufficiently large: this would not happen when the intervention is small-scale or geographically limited. Thus, a typical small-scale RCT may overstate the social benefits of the training, since crowding out in the labour market is ignored. When interactions among individuals induced by the policy intervention are ignored we talk about *partial* equilibrium effects. This is in contrast with approaches that take into account these interactions and place the analysis in a market setting (*general* equilibrium effect).

For some type of interventions, interactions between people are crucial to the impact of the intervention and can enhance or decrease it. Immunisation interventions are a case

[2] A survey of qualitative methods and the related mixed methods literature is outside the scope of this review, which is primarily concerned with quantitative approaches to quantify the causal impact of interventions. Denzin and Lincoln (2017) present a comprehensive survey of qualitative methods; a review of mixed methods is presented in Abbas and Charles (2003).

in point: immunisation has direct effects on the treated but those not treated also benefit from the reduced disease transmission, creating positive externalities (Miguel and Kremer, 2004). Another example are interventions involving social learning: think about peace education programmes aimed at reducing outbreaks of social conflict by changing prejudices and social norms about the perceived "enemies". For such a policy to be effective, a "critical-mass" majority in both parties has to favour peaceful interactions before a reduction of violent conflict outbreaks can be observed. If only a few people are participating in the education programme, the intervention may hardly show any effect: a few pacifistic individuals on their own cannot stop masses spurred by hatred and revenge. Another example can be found in White (2003) who models the critical mass required for participatory projects to succeed. A final important example is that of a subsidy to college enrolment that might improve the education of a group of students with limited resources. If such an intervention is implemented on a large scale it might affect the supply of a certain type of skills in the job market and, therefore, change the impact of the intervention relative to what had been estimated in an RCT of limited scale.

The question of timing is also crucial in the information basis upon which policies are designed. The question of timing can take many forms: how long the beneficiaries should be exposed to the intervention before benefiting from it? When do we evaluate? How is the impact effect evolving over time, does it dilute or does it amplify?

One key learning input for policy design is the long-run responses to a given intervention. Most RCTs are limited to a short period after the intervention took place and thus give insights over the short-run responses. This may overestimate (or underestimate) the impact on the long-run. Take our earlier example of the job training programme: wages do not adjust immediately. As a result, in the short-run, newly high-skilled labour earns higher wages and are better off. The programme looks effective in increasing treatment group's wages. In the medium-run, once wages in the labour market adjust, this effect is likely to disappear.

Intermediate outcomes may also take time to develop. Early Child Development programmes are a case in point because they could lead to long-run outcomes beyond those originally intended by the intervention. For example, early child stimulation programmes are associated with higher educational attainment and higher earning in adulthood (Grantham-McGregor and others, 2007). Another example comes from road building projects. Impacts of new roads in rural areas include development that takes time to emerge, such as changes from agriculture to non-agricultural income-generating activities or increases in secondary schooling after the primary schooling is completed (Mu and van de Walle, 2011; Van de Walle, 2009).

Section 3

# Complements and Alternatives to RCTs for (Causal) Evaluation Questions

When the interest is to gauge the causal impact of an intervention, RCTs provide a source of credible exogenous variation which is a necessary element to identify the causal relationship of interest. But what if an RCT is not feasible? What if we want to learn about the conditions under which the policy generates the impact (in addition to measuring the treatment effect)? What if we are concerned about the spill-over effects of an intervention in the wider context? Given how people interact in a particular setting, could we have achieved the same impact by another intervention?

Impact evaluation may be concerned with all of these aspects. To learn about these aspects, we need to able to understand three elements: the impact, the mechanisms of impact and the general equilibrium effects. In an RCT minded approach, the random allocation of treatment (or even multiple treatment arms) provides the exogenous variation necessary to quantify the causal impact of the intervention. Yet, in order to fully understand that estimate, one needs a framework of how inputs lead to changes in the primary outcome of interest (i.e. the impact), how these changes are generated (i.e. the mechanisms of impact) and how other outcomes may be affected by the policy intervention and feedback into the outcome of interest (i.e. general equilibrium effects). Modelling provides this framework: it makes the links between relevant inputs and outputs explicit and informs empirical testing. In other words, the model embeds the logic of relationships, while the RCT's estimates quantify their value in specific contexts.

When an RCT is available, the random allocation of treatment provides the mean to estimate a credible causal estimate of the average treatment effect. A model helps to make sense of the RCT's average treatment effect by introducing it in a coherent and logical framework describing how people interact in their environment. Once the mechanisms of impact are made explicit in the model, which ones are relevant can be tested using data. This informs why the ATE we observe coming from the RCT originates. By treating explicitly all relevant actors' interactions (including the actors not directly targeted by the intervention), the model can also help us learn about the general equilibrium effects and guide simulations of future scenarios.

However, RCTs are not always available and often, on their own, are too narrow a method. This section presents alternative methods that, in some cases, complement RCTs and, in others, offer alternatives to them. Their focus is on one specific element of the general framework. First, and most importantly, when a trial is not feasible, one needs a method to find and interpret alternative exogenous variations to be able to quantify the causal impact. Section 3A presents methods (Matching techniques, Difference-in-Difference, Instrumental variable, Control function and Synthetic control method) that create "control" and "treatment" groups when RCTs are not available. Secondly, Section 3B illustrates a method (Decomposition techniques) that helps unpack

where the impact comes from in absence of controlled multiple treatment arms seen in RCTs with factorial designs. Third, Section 3C discusses how structural modelling can bring these elements together in a unified framework and also addresses the question of spill-over effects. Table 1 summarises these alternative approaches and provides one reference for each approach as an illustrative application.

**Table 1: Approaches alternative to RCTs**

| Evaluation question | RCTs minded approach design | Alternative methods (in conjunction with RCTs or when RCT are not applicable) | Examples of use |
|---|---|---|---|
| This intervention causes outcome Y | Random allocation of treatment | Instrumental variable / Regression Discontinuity Propensity Score Method | Angrist, J.D., 1990. Lifetime earnings and the Vietnam era draft lottery: evidence from social security administrative records. *The American Economic Review*, pp.313-336.<br>Angrist, J.D. and Lavy, V., 1999. Using Maimonides' rule to estimate the effect of class size on scholastic achievement. *The Quarterly Journal of Economics*, *114*(2), pp.533-575.<br>Dehejia, R.H. and Wahba, S., 1999. Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American statistical Association*, *94*(448), pp.1053-1062. |
| | | Synthetic control method | Abadie, A. and Gardeazabal, J., 2003. The economic costs of conflict: A case study of the Basque Country. *The American Economic Review*, *93*(1), pp.113-132. |
| Uncovering individual treatment effects | X | Participatory evaluation | Pouw, N., Dietz, T., Bélemvire, A., de Groot, D., Millar, D., Obeng, F., Rijneveld, W., Van der Geest, K., Vlaminck, Z. and Zaal, F., 2017. |

| | | | |
|---|---|---|---|
| | | | Participatory Assessment of Development Interventions: Lessons Learned from a New Evaluation Methodology in Ghana and Burkina Faso. *American Journal of Evaluation*, *38*(1), pp.47-59. |
| Uncovering the causal mechanisms of impact | Random manipulation of treatment and mediator variables | Mechanism experiments | Ludwig, J., Kling, J.R. and Mullainathan, S., 2011. Mechanism experiments and policy evaluations. *The Journal of Economic Perspectives*, *25*(3), pp.17-38. |
| | | Structural modelling | Attanasio, O.P., Meghir, C. and Santiago, A., 2012. Education choices in Mexico: using a structural model and a randomized experiment to evaluate Progresa. *The Review of Economic Studies*, *79*(1), pp.37-66. |
| Uncovering the conditions under which impact occurs (e.g. intensity of treatment, generalisability) | Multiple treatment arms with random allocation | Decomposition techniques | Attanasio, O., Cattan, S., Fitzsimons, E., Meghir, C. and Rubio-Codina, M., 2015. Estimating the production function for human capital: Results from a randomized control trial in Colombia (No. w20965). National Bureau of Economic Research. |
| | | | O'Donnell, O., Nicolás, Á.L. and Van Doorslaer, E., 2009. Growing richer and taller: Explaining change in the distribution of child nutritional status during Vietnam's economic boom. |

| | | | |
|---|---|---|---|
| | | | *Journal of Development Economics*, *88*(1), pp.45-58. |
| | | Post-experimental stratification / reweighting techniques | Stuart, E.A., Cole, S.R., Bradshaw, C.P. and Leaf, P.J., 2011. The use of propensity scores to assess the generalizability of results from randomized trials. Journal of the Royal Statistical Society: Series A (Statistics in Society), 174(2), pp.369-386.<br><br>Hotz, V.J., Imbens, G.W. and Mortimer, J.H., 2005. Predicting the efficacy of future training programs using past experiences at other locations. Journal of Econometrics, 125(1), pp.241-270. |
| Understanding the impact of scaling up/ long term effects | X | General equilibrium modelling | Coady, D.P. and Harris, R.L., 2004. Evaluating transfer programmes within a general equilibrium framework. *The Economic Journal*, *114*(498), pp.778-799. |

# 3A. Identify the Causal Impact in Absence of RCTs

When an RCT is not available or infeasible, a plausible alternative is to think of the reason why a comparison between individuals (or communities) who receive the programme and those who do not, does not yield an estimate of the (average) impact of the intervention under study. In other words, one needs to understand why and how such a simple comparison would yield a biased estimate of the impact and the relationship between treatment assignment and variables that might be related to the outcome of interest. What is needed, as we argue below, is a model of the outcome of interest and how that can be related to treatment, as well as other variables. Under some assumptions, such a model can be used to derive credible estimates of the impact of certain interventions.

## 3A.1 Matching Techniques

Some of the issues in identifying causal relationships can be expressed within a simple regression framework. Suppose the outcome of interest is given by the following expression:

$$Y_i = a + d\,D_i + e_i \qquad (3)$$

where $e_i$ is an unobserved term. From a statistical point of view, the problem in identifying the parameter *d,* which represents the impact of the intervention $D_i$, arises from the possibility that the intervention is correlated with the unobserved variable $e_i$. This correlation would induce a bias in the estimation of *d* if one were to estimate equation (3) by OLS.

Matching methods assume that, whilst the unobserved component $e_i$ might be correlated with the intervention, controlling for enough variables, makes this correlation disappear. In particular, we can re-write the equation above as:

$$Y_i = a + d\,D_i + c\,X_i + u_i \qquad (4)$$

where it is now assumed that the term $u_i$ is uncorrelated with D, conditional on X.

An example can explain this approach (Barlow *et al*., 1981). Suppose that a given intervention is targeted to students of different socio-economic backgrounds and that the sample including treated and non-treated individuals includes students of different backgrounds. If background is relevant for the outcome of interest, a simple comparison of treated and non-treated is invalid if more students of a certain background are assigned the intervention. If a measure of background is the only relevant variable in this context, one can condition on it to obtain an unbiased estimate of the impact, comparing students with the same

background or, if the relationship between the outcome of interest and background is linear, using an equation such as (4).

The assumption of linearity used in equation (4) is a strong one. In many situations, such a specification is extremely restrictive as it excludes non-linear terms and possible interactions between different variables. On the other hand, a full non-parametric analysis might require a very large sample. A method that has received considerable attention is that of Propensity Score Matching (PSM), which provides a useful tool to condition for the effect of a potentially large vector X on the estimate of the impact (Rosenbaum and Rubin, 1983). Such a method consists in estimating the probability of treatment exposure as a function of a potentially large vector of variables X, and then consider an equation such as (4) where the vector X is substituted with the probability of receiving treatment. A clear discussion of this method can be found in Caliendo and Kopeinig (2008).

### 3A.2 Difference in Difference

Suppose that not enough control variables are available to make the correlation between D and $u_i$ zero conditional on X. It is possible, however, if one observes individuals at different points in time, before and after the intervention, that the term $u_{it}$ (where the subscript *t* indicates time), can be written as:

$$u_{it} = \varepsilon_i + \eta_{it} \tag{5}$$

where $\varepsilon_i$ is correlated with D but constant over time and $\eta_{it}$ is uncorrelated with D. In such a situation, one can take first differences of the equation determining the outcome of interest to eliminate $\varepsilon_i$ and obtain consistent estimates of treatment impacts.

$$\Delta Y_{it} = d\, D_{it} + c\Delta X_{it} + \Delta\eta_{it} \tag{6}$$

Equations like (6) have been used extensively in the literature. Notice that equation (6) combines matching and difference in difference. The assumption needed is that changes in the unobserved component are uncorrelated with the treatment, conditional on the changes in X.

### 3A.3 Instrumental Variable and Control Function Approach

Identifying causal relationships can be complicated because of unobserved omitted variables and/or because the treatment of interest and the outcome have some feedback on each other: the treatment may cause the outcome to change, and changes in the outcome feed back onto the treatment. In such cases, the method of instrumental variable (IV) is an alternative to identify exogenous variation in a variable of interest (treatment) and use it to make a causal inference of treatment effect on an outcome. IV solves the problem of missing or unknown control variables and the circular feedback loop between treatment and outcome by identifying the variation in 'treatment' driven by

factors that are unrelated to the outcome of interest and using only that variation to estimate the causal effect of the treatment.

Take the example of transport infrastructure and its effect on urbanisation. If motorways are used intensively and congestion becomes more frequent, then this will lead to an expansion of the road network in order to relieve the strain on the infrastructure. Improvements in the infrastructure network enable people to relocate and spread out their activities (e.g. living in greener suburban areas and commuting to the workplace in the city centre), thus "improved infrastructure" *causes* more urbanisation. At the same time, this phenomenon of "urban sprawl" induces more people to use infrastructure, putting pressure on infrastructure improvements. In this sense, urbanisation *causes* infrastructure improvements. This kind of feedback mechanism leads to a circular relationship between the explanatory variable of interest (the treatment) and the outcome variable, which is referred to in the literature as "reverse causality". Obviously, infrastructure projects are not randomly located and it would be unthinkable to do so. The construction of roads, railways and power plants is planned years in advance, often based on the (expected) need, cost and opportunity.

Another example is the impact of economic growth on the onset of civil war. Studying the relationship between changes in standards of living on the onset of conflict is complicated because civil wars cause economies to grow more slowly (a "reverse causality" problem) and there are likely to be unobservable variables that affect both economic growth and conflict outbreak, like government institutional quality, which is difficult to measure. Better institutions lead to less conflict and more economic growth. If we observe that high economic growth is associated with less conflict, is that effect the result of economic growth or the concurrent presence of better institutions? Omitting institution quality creates an "omitted variable bias" problem and this makes the ATE not causal. The IV method offers an alternative to true randomisation and aims to break the feedback loop of reverse causality and the bias arising from omitted variables.

The main idea of the IV method is to model explicitly the process of exposure to the "treatment" of interest (e.g. infrastructural improvement or economic growth). The main assumption for IV to work is that the instrumental variable determines the assignment to the "treatment" but not the outcome of interest. This condition is important: the instrumental variable affects the outcome only through its link with the "treatment", it must have no direct influence on the outcome. This condition is called in the econometric jargon "exclusion restriction" because it excludes direct effects on the outcome. It is an assumption because it is untestable. In other words, the researcher exploits the exogenous variation in the IV to identify exogenous variation in the "treatment" variable and track changes in the outcome of interest. The reliability of the IV method is dependent on the validity of the exclusion restriction assumption.

It is interesting to note that randomisation, when feasible, constitutes the perfect instrument. A variable that determines (in this case exclusively) treatment assignment and is, by construction, uncorrelated to the outcome of interest.

From an econometric point of view, the exogenous variation in the variable of interest, X, is identified by regressing X on the instrumental variable and recovering the predicted values of X, i.e. what X would look like if it was only determined by the exogenous variation of the instrumental variable. The predicted value for X, instead of the actual value of X, is then used to estimate the causal effect of X on Y in a standard regression analysis.

Alternatively, instead of using the predicted value of X, one can recover the predicted residuals from the regression of X on the instrument and add the actual value of X and the predicted residual in the equation of Y. The predicted residuals absorb all the other unobserved factors that explain the variation in X beyond the instrumental variable. This method is called *control function approach*. In the linear case, instrumental variable and control function approach are identical. In the case of non-linear models, the methods differ and the control function typically is easier to implement.

In the examples above, the researchers should choose an instrument which is only related to the "cause" of interest, but not the "outcome":

Infrastructure improvements could be instrumented with projected construction costs, which may sometimes increase unexpectedly and turn out be prohibitively high, such that some projects (which were planned to address the needs of urbanisation) are now cancelled. Construction costs are (arguably) independent of the degree of urbanisation.

Economic living standards could be instrumented with occurrences of natural disasters, such as floods, forest/bush fires or avalanches, or climatic conditions such as above average rainfall. In regions which have experienced the substantial destruction of capital through natural causes or unusual rainfall (particularly for agricultural regions), living standards will be lower. Such catastrophes occur independently of conflict or peace in a region.

In both cases, the treatment and control groups are implicitly defined by the instrumental variable. The instrument (e.g. natural disasters) tries to identify the treatment (e.g. low economic standards) where the outcome (e.g. conflict) cannot be the (main) cause. If conflict arises because of poor living standards, then conflict would be more likely to occur in regions after a natural disaster, since the living standards would be lower. The researchers can predict changes in living standards based on natural disasters (this step identifies the treated group and it is called the first stage) and use this predicted variation to estimate the causal impact of living standards on the outbreak of conflict (second stage). The control group would be those regions without natural disasters.

The fact that the validity of an instrument (especially the exclusion restriction) cannot, in most situations, be tested empirically is a major limitation. Instead, the researcher has to rely on causal chains and argues against potential objections. For example, the assumption that natural disasters and conflict are not directly related could be contested if conflict deters public investments into preventative

measures such as dams or warning systems, which would reduce the economic impact of natural disasters.

### 3A.4 Synthetic Control Method

This method provides a statistically based way to choose comparison units (Abadie et al.,2015). It is particularly useful when the units of analysis are aggregate entities, like regions, districts or states for which one single comparison often does not exist. The main idea of the synthetic control method is to create a suitable counterfactual by using a combination of comparison units. The "synthetic control" is an artificial control unit created as a weighted average of the (pre-intervention) characteristics of a pool of comparison units. The key of the synthetic control approach is to choose the weights such that the difference between the pre-intervention characteristics of the treated unit and the synthetic control is minimised. Formally, this can be represented as follows: let $X_1$ be a vector containing the value of the pre-intervention characteristics of the treated unit and $X_0$ be a matrix collecting the values of the same characteristics for all the pool of possible comparison units. The synthetic control aims to choose a set of weights, **W***, that minimises the difference between the pre-intervention characteristics of the treated units and the synthetic control, that is $X_1$-$X_0$W. The synthetic control estimator of the treatment effect is the difference in (post-intervention) outcomes between the treatment unit and the synthetic control unit.

In the context of policy impact evaluation, this method can be useful in contexts in which the policy under evaluation is an aggregate phenomenon that impacts an entire region and thus it is difficult to find a proper comparison region. One example is the impact of conflict on economic development. Abadie and Gardeazabal (2003) use the synthetic control method to estimate the economic cost of ETA terrorism in the Basque Country. Using time-series data on economic growth predictors from other Spanish regions not affected by conflict (e.g. investment ratio, population density, sectoral shares and human capital indicators), they create a synthetic Basque Country that best reproduces the trajectory of real per capita GDP of the Basque Country in the absence of conflict. The GDP gap between the *actual* GDP figure in the post- ETA terrorism period and the synthetic prediction of GDP without terrorism offers an estimate of the economic costs of terroristic conflict (see figure 1 taken from Abadie and Gardeazabal, 2003). These estimates can then be subject to further econometric analysis, such as impulse response functions in order to show, for example, at which point the impact of terrorism is highest and how long it takes for the economic impact of terrorism to fade away.

The main limitation of the synthetic control is that inference is based on observed characteristics only. Unobservable factors driving the outcome variable limits the causal interpretation as is common in observation studies without valid randomisation. Thus, inference requires a battery of "falsification tests" of alternative hypothesis or cross-validation as is customary in observational

studies.  A second limitation concerns the data requirement over the pre-treatment period: the reliability of the method centres on the ability to track the treated unit's outcome of interest and observable characteristics over a long period of time prior to the intervention, thus the availability of these data is essential.

**Figure 1: Per capita GDP for the Basque Country (Abadie and Gardeazabal, 2003)**

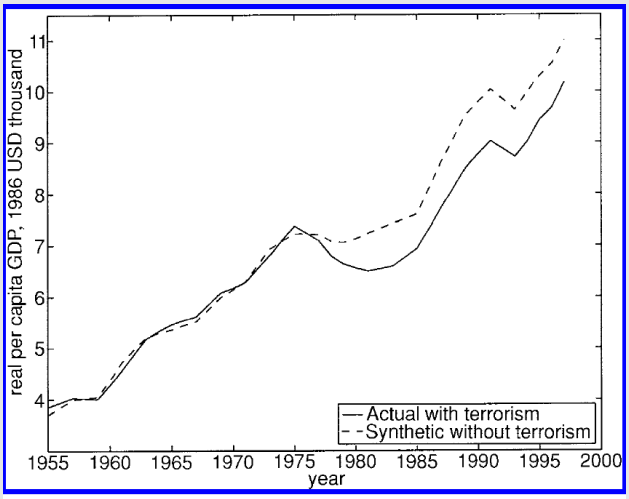


FIGURE 1. PER CAPITA GDP FOR THE BASQUE COUNTRY

# 3B. Learn about the Mechanisms of Impact

## 3B.1 Decomposition Methods

Decomposition methods refer to methods aiming to unpack systematically what drives variation in outcomes. Whenever we observe outcome differences between two groups, these may depend on differences in the endowments between the two groups or differences in the strengths of association between the endowments and the outcomes. For example, two regions may have different child malnutrition rates because they have different levels of income (i.e. differences in endowment) and/or different degrees of responsiveness of child malnutrition to income. The degree of responsiveness may depend on the environment and the availability of services, which may or may not be observable: for example, changes in the level of immunisation change the way child malnutrition is responsive to income.

Interventions may create both endowment changes and responsiveness changes. Think about a child stimulation programme aiming to increase child cognitive skills. In addition to impacting directly children cognitive skills, the stimulation intervention may increase the time parents spend with their kids (i.e. level of endowment) by raising awareness about the importance of stimulation activities, and it may change the responsiveness to parental time investment (e.g. because the quality of parent-child interactions has improved). Decomposition methods aim to measure the relative contribution of changes in endowments (the *covariate effect* in a regression jargon) versus changes in the way endowments are related to outcomes (the *coefficient effect*).

To gauge the relative contribution of different effects, the main idea of decomposition methods is to create a counterfactual distribution of the outcome that would have prevailed under different scenarios. Think for example of outcome distributions pre- and post-intervention. To measure the covariate effect, the decomposition method simulates the outcome distribution that would have prevailed in the post-intervention period had the coefficients of responsiveness been the same as in the pre-intervention period (while instead, the endowment distribution has changed). This simulated distribution imposes the same coefficient responses occurring in pre-intervention period to the post-intervention period. Thus, the covariate effect is the difference between the post-intervention simulated distribution, which imposes no differences in responsiveness across periods, and the outcome distribution arising in the pre-intervention period. To measure the coefficient effect, the decomposition method simulates the outcome distribution that would have prevailed in the pre-intervention period, had the endowments been distributed as in the post-intervention period. Thus, the only element allowed to change across periods are the coefficients. The coefficient effect is measured by the difference between the outcome distribution observed in the post-intervention period and the simulated pre-intervention distribution (with post-intervention endowments).

Often the interest is on the contribution of one single covariate that changes post-intervention. With panel data, it is possible to resample and create a simulated pre-intervention distribution under the counterfactual that only the covariate of interest is distributed as in the post-intervention period while the coefficients and the other covariates (assuming these are uncorrelated with the covariate of interest) remain the same as in the pre-intervention period. The contribution of the covariate of interest will be measured by the gap between the distribution in the pre-intervention period and the counterfactual pre-intervention distribution.

Decomposition techniques differ in the way they estimate the simulated distributions and which difference across distributions they focus on. The Oaxaca-Blinder decomposition technique uses linear regression analysis to provide differences in *means* across pre- and post-intervention distributions (Oaxaca, 1973). The Machado and Mata (2005) decomposition technique uses quantile regressions to build the counterfactual distribution and provides relative contributions of covariate and coefficient effects at various quantiles of

pre- and post-intervention outcome distributions. These techniques can be used to identify the contributions of coefficient and covariate effects in explaining changes in outcome over time (e.g. O'Donnell, 2009) or comparative case-studies explaining differences across states or regions (e.g. Cavatorta *et al*. 2015).

The techniques have two main limitations. First, the techniques alone are statistical methods and do not aim to address causality. Thus, the causal interpretation of the covariate and coefficient effects depends on the ability of the model to control for unobservable influences. The main assumption is that conditional on all observable factors included in the specification, the distribution of unobservable factors does not systematically influence the outcome distribution. The limit to the causal interpretation due to unobservable confounding variables is common to traditional observational studies. This drawback can be solved by designing a random allocation of covariates of interest or using the distribution of the covariate(s) of interest that is plausibly exogenous to outcomes. Second, the counterfactual distributions are based on resampling and simulations and thus requires large datasets in order to produce reliable and precise simulated distributions.

# 3C. Understanding Changes in Behaviour in a Unifying Framework

## 3C.1 Structural Modelling

In a model of behaviour, individual agents (these can be individuals, households or firms) make choices to maximise (or minimise) some objective (which can be utility or profit) expressed as a function, subject to a number of constraints they face. These constraints can be resources, biological constraints, cognitive constraints, technological constraints, the nature of markets and the information agents have access to. A structural model is a model of behaviour that aims to identify how agents make their decisions and their reactions to the various constraints they face. These responses are captured by parameters that can be estimated using the data. The researcher has to make some choices about which of these parameters are invariant to changes in the constraints (that is, structural) and which are not.

In the context of policy evaluation, these models can be estimated using the data prior to the introduction of a policy or programme. This offers a picture of how things work in absence of the policy. Then, this picture is modified by introducing the policy and the model is used to predict the behavioural effects arising from the introduction of the policy. For example, suppose we are interested in evaluating the introduction of a health insurance intervention that subsidises the price of health care in hospitals on the choice of health providers. First, we

estimate how people choose health providers in the current status-quo (i.e. in absence of the policy) as a function of the price. Then, introducing the policy can be done by "manipulating" the price by the amount of the subsidy and observe what the model predicts in terms of health care provider choices.

There are large gains in adopting structural modelling in the context of an RCT. In this context, if the structural model is estimated on the baseline pre-intervention data on control and treatment subjects, the RCT can be used, among other things, to assess the model's validity. In order to validate the model, the "simulated" impact of the policy using the structural model is then compared with the impact (i.e., the ATE estimate) estimated from the experiment. If the two estimates are sufficiently close, then this gives confidence in the validity of the model to represent behaviour. An RCT can also be used to estimate a more general version of the structural model than could be possible to estimate using non-experimental data.

If the model is judged to represent reality reasonably well, either through a validation exercise or because it passes several diagnostic tests, then it can be used to simulate the impact of a variety of counterfactual policies. They may reflect a menu of different policies available to policy-makers or modifications of a given policy that policymakers may want to learn about. The model and the mechanisms that it identifies can also be used to extrapolate the results obtained in a certain context to slightly different situations.

Two examples of this approach are Todd and Wolpin (2006) and Attanasio et al. (2012). They estimate a structural model of household decisions upon whether enrolling their children to school, taking into account economic incentives, such as a cash transfer from the government conditional on school enrolment, the opportunity cost of going to school (that is, the wages in the child labour market) and the distance to schools. The models are estimated using the baseline data from the RCT evaluation of PROGRESA, a conditional cash transfer programme in Mexico. The availability of the random variation in the cash transfer availability allows for the identification and estimation of some mechanisms of impact (for example, the effect of the cash transfer on the decision to go to school or not can be identified separately from the effect of wages in the labour market).

The availability of the data from the RCT allows also the validation of the model's ability to predict the impact of the grant on schooling decisions that actually occurs in the field and the simulation of policy experiments of interest. Attanasio et al. (2012) use the estimated model to simulate the effect of different cash transfer structure on schooling decisions: for example, they perform a policy experiment whereby transfers to primary schools are eliminated in favour of increased transfers to secondary schools, using the same amount of government budget. Another policy experiment is the elimination of the transfer in favour of reducing the distance to schools by building more schools. In this way, the model helps to extrapolate the results to slightly different versions of the policy. The RCT results cannot say anything about these counterfactual policies.

Structural models are also useful to identify and test specific mechanisms of impact of an intervention. This is particularly relevant for social policy interventions, whereby is if often difficult, if not impossible, to manipulate mediator variables independently of the treatment. By formalising how outcome of interest is generated and how the intervention changes one or more aspects of this process, a structural model can help identify some of the reasons why an intervention works.

Attanasio et al. (2015) provide an example in the context of early child stimulation and its impact on cognitive and socio-emotional skills. They formalise a theoretical framework of how cognitive and socio-emotional are created, that is a "production function" of human capital. They assume that the development of children skills is influenced by the stock of cognitive and socio-emotional skills already existing, the parent's cognitive and socio-emotional skills and the parental investment (material and time investments). These inputs come together and affect the accumulation of skills. The formalisation of the model has one main advantage: it identifies the contribution of each input into the accumulation of skills in a set of estimable parameters. These parameters can be made a function of the intervention such that one could study whether differences in outcomes between treatment and control comes from differences in the level of inputs (say, parental investments) or differences in the association between the input and the outcome (the estimated parameters). These differences can shed light on the mechanisms of impact of an intervention. For example, Attanasio et al. (2015) find that the parameters of the production function of children's skills are not shifted by the stimulation intervention, but the level of parental investment increased in treatment areas. In other words, the stimulation intervention does not affect the way skills are produced but it improves skills accumulation by increasing the material and time resources given by parents to their children.

Quantitative models of behaviour obviously have limitations. Reality is complex, models tend to be simpler representations of reality. There is always a trade-off between the tractability of the model and how close the model is to reality. Modelling requires assumptions (e.g., which factors in the subject's decision-making are included and which ones are excluded). The ability to identify the parameters of a certain model requires variation in data which is credibly defined as exogenous. In this sense, the availability of an RCT can help in identifying more general models. On the other hand, using the variation of the RCT for identification may preclude useful exercises.

The credibility of "simulated" impact of new policies using structural modelling relies heavily on the validity of the modelling assumptions. These assumptions are difficult to test in general. The experimental variation provided by RCT is an advantage in this regard. The alternative is to use out-of-sample predictions. However, these require the availability of data that were not used in the estimation of the models. This is possible if data are large-N datasets but in practice, there is a trade-off between achieving precise model estimates and

using less data (to maintain a sufficiently large number of observations out-of-sample).

## 3C.2 General Equilibrium Modelling

A general equilibrium (GE) model is an extension of a structural model since the behaviour of individuals is characterised by a set of formal objectives and constraints. However, a GE approach models all interactions within an economy as a closed system. The economy consists of different types of agents, such as households, firms and the government, which pursue their own objectives (e.g. utility, profit or welfare maximisation) and face different constraints. The agents can exchange commodities and resources (such as labour, capital and consumption goods) on the respective markets. In a general equilibrium, supply and demand on all markets jointly determine the prices for which the commodities are traded in the economy.

Since individual agents are participating in multiple markets (e.g. workers earn wages for supplying labour to the labour market and use this labour income to purchase consumption commodities on the goods market), a change of prices (or supply/demand) in one market will affect agents' behaviour, which in turn changes market outcomes in other markets.

In the context of impact evaluation, a general equilibrium analysis thus helps to understand so-called "second-round" or "indirect" effects of a policy measure, which arise from the interdependencies of markets. In contrast to RCT, where an intervention is designed and analysed on a small scale in isolation, general equilibrium models can draw a more holistic picture of the intervention's impact on the economy, and thereby provide insights about scaling up the intervention in the longer run.

Consider for example a poverty alleviation measure in the form of targeted cash transfers to poor households in Mexico (Coady and Harris, 2004). Using a computable general equilibrium (CGE) the authors try to assess the overall effect on welfare of such a policy measure on the Mexican economy. In their model, the transfer program needs to be financed domestically by tax revenues; i.e. they assume a budget constraint for the government, but the choice of taxes will in turn impact welfare.

Coady and Harris (2004) define welfare as an aggregate function of individuals' well-being (utility) and derive mathematical expressions for welfare changes, depending on the financing of the transfers. These expressions are then fed with Mexican household data to compare the direct and indirect effects of different value-added tax (VAT) or subsidy regimes as means of program financing.

The direct effect of the transfer stems from the redistribution of income from wealthy to poor. Without a social aversion against income inequality in the economy, i.e. when an additional peso of income is of equal social value to a poor or a rich household, this direct effect is neutral on welfare. Evidently, the more society values income equality, the greater the direct benefit of the

transfer. However, funding such a redistribution measure through consumption taxes (VAT) or subsidies results in two indirect effects.

The first component is a reallocation effect: assuming that the recipients of the transfer (poorer households) have a different consumption pattern than its contributors (wealthier households), the transfer will also shift the aggregate pattern of consumption accordingly. Furthermore, if goods are not uniformly taxed (e.g. basic consumption goods may have lower tax rates as opposed to luxury goods), a change in the aggregate consumption pattern will feed back into the total tax revenue that the government collects.

The second indirect effect arises from the tax collection through distorting tax instruments. A change in tax (or subsidy) rate, needed to finance the transfer program, changes the effective price for consumers (producers), who will adjust their quantities demanded (supplied); this change of the market outcomes may again feed back into the tax revenue.

Both of these "second-round" effects highlight that credible financing of the transfer program in the long run, also has to account for behavioural responses outside the target group. In this case, the authors find that, for any social preference over income inequality, targeted cash transfers can be financed most efficiently by first abandoning (already) distortionary food subsidies instead of changing the VAT-regime.

**Table 2: Social Cost of Public Funds (Coady and Harris, 2004)**

**Table 2. Social cost of public funds**

| | | Cost of raising a unit of revenue ($\lambda_j$)[a] | | | | | |
|---|---|---|---|---|---|---|---|
| Inequality aversion[b] | Benefit ($\lambda_D$) | Food subsidies | PVAT (0,7.5, 15) | HVAT (0,5,16) | TVAT (0,11) | BVAT (7,10) | SVAT (8.3) |
| ε = 0 | 1 | 0.625 | 1.061 | 1.071 | 1.051 | 0.969 | 0.955 |
| ε = 0.5 | 1.242 | 0.468 | 0.732 | 0.751 | 0.718 | 0.668 | 0.685 |
| ε = 1 | 1.584 | 0.397 | 0.611 | 0.633 | 0.602 | 0.560 | 0.599 |
| ε = 2 | 2.792 | 0.395 | 0.658 | 0.679 | 0.664 | 0.612 | 0.690 |
| ε = 3 | 5.448 | 0.557 | 1.023 | 1.045 | 1.054 | 0.970 | 1.109 |
| ε = 4 | 11.549 | 0.996 | 1.962 | 1.988 | 2.042 | 1.882 | 2.155 |
| ε = 5 | 26.011 | 2.060 | 4.227 | 4.263 | 4.425 | 4.082 | 4.671 |

Suppose, no distortionary subsidies were in place, and the government has to adjust VAT to finance the transfer program; e.g. VAT can be charged at different rates for different goods, or at a uniform rate. The trade-off between equity (direct effect) and efficiency (indirect effects), and hence which tax regime would be desired, will crucially depend on how society is assumed to value the well-being of poorer compared to richer households when redistributing resources. One could quantify these trade-offs by comparing the cost of public funds across alternative ways of financing the programme and assuming different values of society's distributional concerns. An example of these simulations based on Coady and Harris (2004) is presented in Table 2.

Coady and Harris (2004) formalise a general, yet arbitrary specification to characterise this inequality aversion. As with any structural model, a general equilibrium model simplifies the mechanisms in the real world. The assumptions on functional forms or parameters are therefore highly debatable. This clearly limits the applicability of general equilibrium models for evaluation. Quantifying an intervention with a general equilibrium model (GEM) for a specific context does not offer an objective and reliable estimate of the total impact. Nonetheless, GEM may be useful to point qualitatively at general indirect effects and - together with robustness tests of model specifications - can offer a way of comparing alternatives, such as different tax/subsidy regimes.

Section 4

# Data and Measurement

The discussion above on the use of matching method stressed that one possibility to identify the parameters of interest is to have enough information on confounder variables so that, conditional on those variables, treatment assignment is random or, at least, unrelated to the outcome of interest. In this sense, the identification problem can be expressed as a problem of not having enough information about the sample being used. If one had all the variables that drive treatment assignment and are related to outcomes, matching methods could be a viable approach. One alternative, therefore, is to use data that are rich enough to cover the relationship between assignment and treatment. This reasoning is analogous to Goldberger (1991) discussion of 'micronumerosity'.

The use of richer and more useful data set has received renewed attention in evaluation. This is particularly true for situations where it is possible to design surveys that include the measurement of variables that could be relevant for such a purpose.

Alternative data sources can be particularly relevant in situations where randomisation is not an option. Administrative data, data on weather patterns, satellite images and so on, might turn out to be particularly useful in many dimensions and have recently received renewed attention.

The use of new data sources and new measures is also relevant to the identification of the mechanisms that generate the observed impacts. When identifying and estimating structural models, additional and rich data sources can be very useful in that they allow the use of richer and more realistic model that can be identified empirically using weaker structural assumptions.

In our opinion, there are three different ways in which the availability, or possibly the construction, of new data and measurements, can be very useful. We list them here.

Measurement error.  Perfect data often do not exist. In such a situation, it can be more useful to have two or three different measures of a variable of interest (be it an outcome or a control variable) than pursuing the perfect measure. It is profitable to model explicitly the presence of measurement error and take it into account by using several measures whose measurement error is uncorrelated. Two useful contributions in this respect are by Schennach (2004) and Cuhna et al. (2010). The need of having the measurement errors of different variables to be uncorrelated can be incorporated into the design of new surveys.

Administrative and historical data. In many situations, administrative data can be extremely useful to estimate the impact of an intervention or establish causal relations, especially when these data can be interpreted within the context of a natural experiment. A recent celebrated example are the studies of the impact of the development of railroads in India on development, pursued by Dave Donaldson and collaborators (see e.g. Burgess and Donaldson, 2010), for which he was recently awarded the John Bates Clark Medal.

Additional measures. In many evaluations, it is very important to measure not only the outcomes of interest but a number of contextual variables that might be extremely useful in identifying the mechanisms behind certain impacts (or lack of). In some cases, this involves designing new measurement tools. An example, for instance, is about measuring people's expectations (for instance about the returns to a given activity) or attitudes or beliefs.

Section 5

# Conclusions

The methodology of RCTs have received considerable attention in international development (as well as other fields) because they offer a rigorous way to identify causal impacts of policy interventions. However, there is more in policy evaluation than the mere identification of causal impacts, and RCTs are not always feasible. This paper reviews methods that, in some cases, complement what can be learned from RCTs and, in other cases, offer alternatives to RCTs. We review three broad classes of methods: first, methods that create "control" and "treatment" like conditions in order to identify and interpret the causal impact of certain phenomena when RCTs are impossible. This class of methods include Matching techniques, Difference-in-Difference, Instrumental Variables, Control Function approaches and Synthetic Control Methods. The second class comprises methods that help unpack where the impact is coming from and why it occurs, such as Decomposition Techniques. The third class comprises one

method that offers a general framework to analyse these aspects altogether and interpret the interactions that intervention causes on the wider context in which it is introduced: structural modelling. We see the use of structural models as important because they make clear the conceptual framework used to study a certain intervention or policy. This is useful to spell out the assumption that are needed to identify the causal relations under study and, importantly, to give an interpretation to the impacts that are estimated in a rigorous fashion, such as when RCTs are available. In this sense, structural models and RCTs complement each other. The former offers an interpretation of the latter, while the presence of the latter allows at times the identification of more flexible models and/or can be used to test some of the assumptions on which they are based.

We argue that there is an important distinction between attempting to estimate the actual causal impact of an intervention from the more ambitious goal to understand why and how things work (or don't). The use of structural models offers a mean to achieve that goal. The model embeds the logic of relationships between input (e.g. an intervention) and output (an outcome of interest). Estimating a model provides a quantification of the importance of these relationships in specific contexts. Whenever we want to learn about the why and how things work, there is scope for a fruitful synergy between RCTs + structural models. RCTs provide a rigorous source of exogenous variation to estimate important causal relationships. Models provide the theoretical framework to interpret the ATE from RCTs and put them into the wider context into which they occur.

The synergy of "RCT+ structural models" has several advantages: firstly, it provides an effective step towards the goal of accumulation of knowledge of "what works" that is transportable and generalisable. Differently from ATEs, models that make sense and work well can be transported in different contexts. Therefore, the lessons learned from the synergy of "RCT+ structural models" go beyond what can be learned from one single case study in one context. For example, important relationships emerged from structural models used in one context can be easily replicated and tested in different contexts. Secondly, structural models that work well can be used to simulate alternative policies scenarios, without the need to embark on costly fishing expeditions of policy experiments in the field. Finally, structural models can be moulded relatively easily to simulate scaling up effects and to encompass other general equilibrium effects.

In conclusion, our main message is that structural modelling is a productive framework and approach to incorporate specific results generated from single RCTs. This does not mean that structural modelling is superior to other methods. Yet, if the interest is in understanding the conditions under which a policy generates impact and the broader changes a policy introduction may produce, structural models offer a practical and rigorous way to draw conclusions on these fronts.

# References


Abadie, A. and Gardeazabal, J., 2003. The economic costs of conflict: A case study of the Basque Country. *The American Economic Review*, *93*(1), pp.113-132.

Abadie, A., Diamond, A. and Hainmueller, J., 2015. Comparative politics and the synthetic control method. *American Journal of Political Science*, *59*(2), pp.495-510.

Abbas, T. and Charles, T., 2003. Handbook of mixed methods in social & behavioral research.

Achour, H. and Belloumi, M., 2016. Investigating the causal relationship between transport infrastructure, transport energy consumption and economic growth in Tunisia. *Renewable and Sustainable Energy Reviews*, *56*, pp.988-998.

Angrist, J.D., 1990. Lifetime earnings and the Vietnam era draft lottery: evidence from social security administrative records. *The American Economic Review*, pp.313-336.

Attanasio, O.P., Meghir, C. and Santiago, A., 2012. Education choices in Mexico: using a structural model and a randomized experiment to evaluate Progresa. *The Review of Economic Studies*, *79*(1), pp.37-66.

Attanasio, O., Cattan, S., Fitzsimons, E., Meghir, C. and Rubio-Codina, M., 2015. *Estimating the production function for human capital: Results from a randomized control trial in Colombia* (No. w20965). National Bureau of Economic Research.

Attanasio, O.P., Fernández, C., Fitzsimons, E.O., Grantham-McGregor, S.M., Meghir, C. and Rubio-Codina, M., 2014. Using the infrastructure of a conditional cash transfer program to deliver a scalable integrated early child development program in Colombia: cluster randomized controlled trial. *BMJ*, *349*, pg. 5785.

Barnow, B., Cain, G. and Goldberger, A., 1981. Selection on observables. *Evaluation Studies Review Annual*, *5*(1), pp.43-59.

Bonell, C.P., Hargreaves, J.R., Cousens, S.N., Ross, D.A., Hayes, R.J., Petticrew, M. and Kirkwood, B., 2009. Alternatives to randomisation in the evaluation of public-health interventions: design challenges and solutions. *Journal of Epidemiology & Community Health*, 65 pp.582-587

Burgess, R. and Donaldson, D., 2010. Can openness mitigate the effects of weather shocks? Evidence from India's famine era. *The American economic review*, *100*(2), pp.449-453.

Cavatorta, E., Shankar, B. and Flores-Martinez, A., 2015. Explaining cross-state disparities in child nutrition in rural India. *World Development*, *76*, pp.216-237.

Caliendo, M. and Kopeinig, S., 2008. Some practical guidance for the implementation of propensity score matching. *Journal of economic surveys*, *22*(1), pp.31-72.

Coady, D.P. and Harris, R.L., 2004. Evaluating transfer programmes within a general equilibrium framework. *The Economic Journal*, *114*(498), pp.778-799.

Cunha, F., Heckman, J.J. and Schennach, S.M., 2010. Estimating the technology of cognitive and noncognitive skill formation. *Econometrica*, *78*(3), pp.883-931.

Denzin, Norman K., and Yvonna S. Lincoln, eds. 2017. *The Sage handbook of qualitative research*. Sage.

Faber, B. and Gaubert, C., 2016. *Tourism and Economic Development: Evidence from Mexico's Coastline* (No. w22300). National Bureau of Economic Research.

Gertler, P., Heckman, J., Pinto, R., Zanolini, A., Vermeersch, C., Walker, S., Chang, S.M. and Grantham-McGregor, S., 2014. Labor market returns to an early childhood stimulation intervention in Jamaica. *Science*, *344*(6187), pp.998-1001.

Grantham-McGregor, S.M., Powell, C.A., Walker, S.P. and Himes, J.H., 1991. Nutritional supplementation, psychosocial stimulation, and mental development of stunted children: the Jamaican Study. *The Lancet*, *338*(8758), pp.1-5.

Goldberger, A. S. (1991). *A Course in Econometrics*. Harvard University Press, Cambridge MA.

Humphreys, M., 2015. Reflections on the ethics of social experimentation. *Journal of Globalization and Development*, *6*(1), pp.87-112.

Humphreys, M. and Weinstein, J.M., 2009. Field experiments and the political economy of development. *Annual Review of Political Science*, *12*, pp.367-378.

Humphreys, Macartan, and Alan Jacobs. 2017. Qualitative Inference from Causal Models. Unpublished manuscript.

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

Machado, J.A. and Mata, J., 2005. Counterfactual decomposition of changes in wage distributions using quantile regression. *Journal of applied Econometrics*, *20*(4), pp.445-465.

Miguel, E. and Kremer, M., 2004. Worms: identifying impacts on education and health in the presence of treatment externalities. *Econometrica*, *72*(1), pp.159-217.

Mu, R. and Van de Walle, D., 2011. Rural roads and local market development in Vietnam. *The Journal of Development Studies*, *47*(5), pp.709-734.

Oaxaca, R., 1973. Male-female wage differentials in urban labor markets. *International economic review*, pp.693-709.

O'Donnell, O., Nicolás, Á.L. and Van Doorslaer, E., 2009. Growing richer and taller: Explaining change in the distribution of child nutritional status during Vietnam's economic boom. *Journal of Development Economics*, *88*(1), pp.45-58.

Rosenzweig, M.R. and Wolpin, K.I., 2000. Natural “natural experiments" in economics. *Journal of Economic Literature*, *38*(4), pp.827-874.

Rosenbaum, P.R. and Rubin, D.B., 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), pp.41-55.

Schennach, S.M., 2004. Estimation of nonlinear models with measurement error. *Econometrica*, *72*(1), pp.33-75.

Thomson, H., Hoskins, R., Petticrew, M., Ogilvie, D., Craig, N., Quinn, T. and Lindsay, G., 2004. Evaluating the health effects of social interventions. *bmj*, *328*(7434), pp.282-5.

Todd, P.E. and Wolpin, K.I., 2006. Assessing the Impact of a School Subsidy Program in Mexico: Using a Social Experiment to Validate a Dynamic Behavioral Model of Child Schooling and Fertility. *The American Economic Review*, pp.1384-1417.

Van de Walle, D., 2009. Impact evaluation of rural road projects. *Journal of development effectiveness*, *1*(1), pp.15-36.

White, H., 2003. Social organization, civic responsibility and collective action: game theory models of community participation in development projects. *Oxford Development Studies*, *31*(2), pp.149-158.

# Contact us

Centre of Excellence for Development Impact and Learning
London International Development Centre
36 Gordon Square
WC1H 0PD
www.cedilprogramme.org
@CEDIL2017