Impact evaluation methods in public economics

Pomeranz, Dina

Abstract: Recent years have seen a large expansion in the use of rigorous impact evaluation techniques. Increasingly, public administrations are collaborating with academic economists and other quantitative social scientists to apply such rigorous methods to the study of public finance. These developments allow for more reliable measurements of the effects of different policy options on the behavioral responses of citizens, firm owners, or public officials. They can help decision makers in tax administrations, public procurement offices, and other public agencies design programs informed by well-founded evidence. This article provides an introductory overview of the most frequently used impact evaluation methods. It is aimed at facilitating communication and collaboration between practitioners and academics by introducing key vocabulary and concepts used in rigorous impact evaluation methods, starting with randomized controlled trials and comparing them with other methods ranging from simple pre–post analysis to difference-in-differences, matching estimations, and regression discontinuity designs.

DOI: https://doi.org/10.1177/1091142115614392

Impact Evaluation Methods in Public Finance

A Brief Introduction to Randomized Evaluations and Comparison with Other Methods

May 2015

Dina Pomeranz
Harvard University and NBER[1]

Abstract

Recent years have seen a large expansion in the use of rigorous impact evaluation techniques. Increasingly, public administrations are collaborating with academic economists and other quantitative social scientists to apply such rigorous methods to the study of public finance. These developments allow for increasingly reliable measurements of the effects of different policy options on the behavioral responses of citizens, firm owners, or public officials. They can help decision makers in tax administrations, public procurement offices and other public agencies design programs informed by reliable evidence. This paper provides an introductory overview of the most frequently used impact evaluations methods. It is aimed at facilitating communication and collaboration between practitioners and academics through an introduction to key vocabulary and concepts used in rigorous impact evaluation methods, starting with randomized controlled trials and comparing them with other methods ranging from simple pre-post analysis to difference-in-differences, matching estimations and regression discontinuity designs.

.

[1]I thank Michael Eddy and Stephanie Majerowicz for excellent research assistance, and the officials of the Chilean and Ecuadorian tax authorities, the Chilean procurement authority, and the Chilean National Comptroller Agency for helpful feedback and suggestions.

1. Introduction

Daily decisions made in public finance can affect the economy of an entire country. However, assessing the effectiveness of different policy options is challenging. Public officials are constantly confronted with a myriad of important questions related to policy impacts on the behavior of citizens, firms or public officials. What policies are most effective against tax evasion? How strongly will firm owners react to tax incentives? How can monitoring be optimized to improve the behavior and compliance of public procurement officials? What type of communication can motivate school officials to disperse educational infrastructure grants promptly?

Recent improvements of impact evaluation techniques allow for increasingly reliable answers to this type of question. A growing number of collaborations between public administrations and academics have facilitated the application of randomized evaluations and other quasi-experimental methods to questions of public finance. These studies often take advantage of already available administrative data, which considerably reduces the cost of their application. The objective of an impact evaluation is to provide information about the effects of current and potential policies. There are various evaluation methods, each with different degrees of validity. The quality of the evaluation is of utmost importance for obtaining informative, unbiased results.

This paper provides an overview of the most frequently used methods, in a language that is accessible both to academics and practitioners in public finance. It offers a brief summary of each method, its advantages and drawbacks, and the conditions under which the method produces valid results. In addition, it provides an introduction to key elements of the specialized terminology of impact evaluations in order to facilitate the communication between policy makers and academics looking to collaborate on these topics.

It is therefore useful to define some basic concepts before presenting the specific methods below. The objective of every impact evaluation is to demonstrate a ***causal effect***. The goal is to measure the impact of a program or policy on some outcome of interest. For example, what is the effect of a notification letter on tax payments? What's causing the impact can be a policy change or the implementation of a new program, such as the mailing of the notification letter. In the context of impact evaluations, the policy whose impact we want to analyze is often referred to as the ***treatment.*** Its effect is then the result that can be attributed directly to the treatment – such as a change in tax filings as a result of the notification letter.

The fundamental challenge of impact evaluation is that at any given moment, it is only possible to observe what happened given the policies in place, not what would have occurred without those policies. It is possible to observe tax filings of taxpayers that received a notification, but it is not possible to observe what those same taxpayers would have done in the absence of such a notification. This imaginary situation of what would have happened in the absence of the treatment is called the ***counterfactual***. Un-

derstanding the counterfactual is key to understanding the impact of a program. Figure 1 provides a graphical representation of this unobserved counterfactual.

[Figure 1]

Figure 1 represents the fundamental challenge of impact evaluations, which seek to measure the difference between the outcome that in fact occurred (shown in light/yellow dots) and a counterfactual that is never observed (shown with dark dots). In this example, we can see that the primary outcome increased more steeply after the intervention (light dots), than would have been the case without the intervention (dark dots). The impact is measured as the difference between the outcome that happened after treatment, and what would have happened without the treatment (the counterfactual).

If an accurate representation of the counterfactual existed, then impact evaluation would be easy. The impact of a program or policy would be the difference between the result observed with the program and the result that would have prevailed without the program – the counterfactual. Given that the counterfactual can never be observed in reality, each evaluation tries – in an explicit or implicit manner – to construct an estimate of the counterfactual to compare it to what occurred. The quality of that representation drives the quality of the impact evaluation.

Normally, the counterfactual estimate is represented by a group called the ***control group*** or ***comparison group***. The control group consists of people or firms that did not participate in the program, while the ***treatment group*** is the group that participated in the program. To measure the impact of the intervention, the outcomes of the treatment group are compared with the outcomes for the control group. An evaluation will produce reliable results if the control group is identical to the treatment group in all its characteristics – observable or not – except one: their exposure to the treatment. In this case, any difference after the intervention can be attributed to the program. In the absence of treatment, both groups would be the same.

All methods used to construct the comparison group rely on assumptions under which the control and treatment group would be comparable. When the assumptions are realistic, the control group is a good representation of the counterfactual. When these assumptions are not realistic, the resulting impact evaluation will be ***biased***. That means it may over- or under-estimate the true effect. A biased evaluation may result in poorly-informed policy decisions and generate losses in terms of effort, time, and public resources. It is therefore important to use high-quality methods in order to obtain reliable impact estimations, and to provide solid evidence for decision-making.

Bias can stem from a variety of reasons that make the treatment and comparison groups different. ***Selection bias*** is produced when those selected into the treatment group are different from those in the comparison group in a way that affects outcomes. This happens also when people who take up a treatment

are different from those who do not (self-selection). Bias can also come about when an external factor affects those in the treatment differently from those in the comparison group. This is sometimes referred to as ***omitted variable bias.*** It biases the conclusion that is reached by comparing the treated group to a comparison group that no longer represents a valid counterfactual**.**

The focus on making the estimation accurate and unbiased is known as ***internal validity***. Internal validity indicates the extent to which a causal conclusion based on a study is warranted, i.e. the extent to which a study avoids the risk of bias. Well-executed randomized evaluations have very high internal validity. Other methods described below have larger risks of bias, and consequently, lower internal validity. These will be discussed in more detail below.

In contrast, ***external validit***y refers to the extent to which the causal findings of a study can be generalized or extrapolated to other situations and settings. For instance, in the case of public finance, external validity may refer to the question whether the findings of an evaluation in one region are informative for a potential nation-wide rollout of a policy, or even for other countries or continents. External validity can to some degree be assessed based on specific knowledge of the setting in question, or one can explicitly test for it through replication of the same analysis in different settings. See Banerjee and Duflo (2009) and Duflo, Glennerster and Kremer (2007) for a more in-depth discussion.

The remainder of the paper discusses characteristics, strengths and limitations of different evaluation methods.[2] Section 2 starts with randomized evaluation as the benchmark to which the other methods can be compared. Sections 3-4 discuss simple difference and simple pre-post analysis. These methods require the strongest assumptions and are most likely to yield biased results. Sections 5-6 present difference-in-differences analysis, matching procedures and propensity scores. Depending on the setting, these methods can yield reliable impact estimations, but they have to be applied selectively and with great care to ensure their underlying assumptions are met. Section 7 provides an introduction to the regression discontinuity design. This method can, under certain circumstances, deliver causal estimates that are just as valid as those from randomized evaluations, with the caveat that they estimate the effect only for a specific subsection of the population. Section 8 concludes.

2. Randomized Evaluation

The goal of randomized evaluations - also called experimental evaluations, randomized controlled trials (RCTs), or randomized field experiments – is to create an ideal comparison group by design from the beginning of the intervention. Study participants, which can be individuals, firms, or entire public entities or localities, are randomly assigned to either receive the treatment or be in the comparison group. This random assignment ensures that (on average) there is no difference between the individuals in the

[2] For a more in-depth treatment of any of these methods, see for example, Angrist and Pischke (2009; 2015), Imbens and Wooldridge (2009) and Gertler et al., (2011).

treatment and control group, except for the fact that one group has been randomly chosen to participate in the program and the other has not. We can therefore rule out that the impact measured is due to a systematic difference between the treatment and control group that would have existed even without the application of the treatment (Duflo, Glennerster and Kremer, 2007). Randomized evaluations are thus often seen as the ideal way to conduct an impact evaluation. It is for this reason that in the evaluation of new medicines and in natural science laboratory research, this method is used almost exclusively.[3]

Another benefit of randomized evaluations is that they allow researchers to identify the effect of a particular component of a larger program. To do so, one can vary one particular factor in each treatment group and compare it to the control group. This way, the casual impact of a particular component of a program or policy can be identified in a way that is difficult otherwise (Banerjee and Duflo, 2009). For instance, studies about what policies can improve access to education and school learning sought to measure the specific effects of textbooks, (Glewwe, Kremer and Moulin, 2009) class-size (Angrist and Lavy, 1999), and student health (Miguel and Kremer, 2004). Randomized evaluations that manipulate one factor at a time, while holding the all other elements of the classroom environment constant, can measure the individual impact of each factor. This isolation of specific factors can make it possible to test particular mechanisms through which a policy has an effect (Ludwig, Kling, and Mullainathan, 2011).

Importantly, randomized assignment requires that the evaluation is designed before the program has begun. For this reason, this method is also called ***prospective evaluation***. In a random process, individuals (or other entities like schools, firms, or villages) are assigned to the treatment group and those not selected are part of the control group. This generates two groups that are similar both in terms of observable characteristics, such as education levels, and unobservable ones, such as motivation. Therefore, any difference that arises later between the treatment and control groups can be attributed to the program and not to other factors. For this reason, if designed and applied adequately, a randomized evaluation is the most valid method for measuring the impact of a program and requires the fewest additional assumptions.

Randomization in practice

This section will lay out a brief overview of the different steps involved in setting up and implementing a randomized field study.[4] The first step is to choose a program, population, and main outcome variables of interest. Ideally, this will be a program that is of interest to the policy maker, in the sense that

[3] In terms of terminology, it is important to distinguish between a randomized evaluation and a random sample: Many studies use random samples to obtain representative information about a population. A random sample does not try to measure impact. The distinctive characteristic of a randomized evaluation is that the treatment is assigned randomly.

[4] For a detailed description of the steps involved in randomized controlled trials under different scenarios, see Glennerster and Takavarasha (2013).

learning about its effectiveness or aspects of its effectiveness will feed into the decision-making process of the public entity.

Second, prior to starting the evaluation, it is useful to calculate statistical estimates to determine the size of the treatment and control groups required for reliably measuring the impact on outcome variables of interest. This analysis is called ***power calculation,*** since it estimates how many observations are needed to have enough statistical power to detect a meaningful effect.

How do we determine the number of participants required in a randomized study? According to the law of large numbers, the greater the number of individuals included in a study, the more likely it is that both groups will be similar. This is one of the reasons why sample size is important. A larger sample is always better since it reduces the likelihood of having unbalanced groups. Moreover, a larger sample improves the precision of one's impact estimates, i.e., it increases the likelihood of detecting the true impact of a program. Nevertheless, a bigger study can be more costly and is not always feasible. Therefore, it is recommended that the statistical power be calculated to determine the minimum sample size necessary for measuring the impact on the main outcome variables of interest.

Statistical power calculations incorporate the different factors that affect the number of required participants. Among the factors to be considered are the variance of the variable of interest and the minimum effect expected to be detected. The smaller the size of the effect one wishes to detect, the larger the number of observations needed. In addition, the higher the variance in the outcome of interest, the larger the number of observations needed to distinguish the true effect size from simple noise in the data. Finally, the randomization design can affect the necessary group size. If the randomization is performed at the group level (clustered randomization), more observations will be necessary than if the randomization is done at the individual level. (Clustered randomization is explained.)

The third step in a randomized evaluation is the random assignment of treatment. The randomization process can be as simple as tossing a coin or a conducting a lottery. In order to make the process transparent and replicable, the random assignment is often implemented using a statistical software such as Stata. It is important that the randomization process be truly random and not just "seemingly" arbitrary. For example, assigning the treatment to people whose surnames start with the letters "A-L" and leaving those starting with "M-Z" as control may seem random, but it is not. Such assignment requires the assumption that the individuals whose surnames start with the letters "A-L" are the same as those that start with "M-Z". Nevertheless, it is possible that the families whose surnames start with the letters "A-L" are different from the families with a last name starting with the letters "M-Z". For example, the ethnic composition may vary. To avoid this situation, an automatic method like using a computer program to generate random numbers that determine treatment assignment is recommended.

A computer also simplifies more complex randomization processes, like ***stratified randomization***. Stratified randomization is recommended when the number of potential participants is small, to en-

sure that both groups are balanced with respect to the most important variables. In stratifying, the sample is divided into subgroups of similar characteristics, with participants within each subgroup randomized to treatment and control, such that the proportion in treatment and control is the same for each subgroup For example, if the population is divided by gender, if 30% of men and 30% of women are assigned the treatment, this assignment will be perfectly balanced in terms of gender. The treatment group will have the same gender composition as the control group.

As mentioned above, an often-used randomization design is ***clustered randomization***. In this procedure, the randomization is not conducted at the level of an individual, but at the level of groups of individuals. This is particularly useful for situations in which it can be expected that the treatment will have spillover effects on others in the same group. For example, when testing a new textbook, random assignment at the student level may not be possible, as the teacher will be teaching from the same book to the entire class. The assignment should then be done at the class level. Or if tax officials wanted to test a new communication strategy towards small firms, they might worry that tax accountants, which work for several firms could, share information across the firms they work for. To remedy this, the randomized assignment could be done at the accountant level, such that all firms that share the same accountant would be in the treatment group, or all in the control. The randomization should in that case be conducted at the accountant level, i.e., randomize among groups of firms that each share a given accountant.

It is not necessary for both groups to be the same size. However, it is important to verify that the groups are balanced with respect to the main outcome variables of interest. That is, the average characteristics (e.g., average firm revenue, industry composition, or percent women) are the same in the treatment and the control group. In the academic literature, experimental studies usually include a balance table that shows that the main characteristics are similar across the two groups.

The fourth step in a randomized evaluation should – whenever possible – be a pilot phase of the planned intervention. A small-scale pilot implementation of the program to be evaluated can provide enormous benefits for the preparation of the large-scale intervention. In practice, the lessons learned from the pilot are often what ends up making the difference between a successful, informative randomized study and an unsuccessful one. The pilots allows the researchers and policy makers to learn about unforeseen challenges at a small scale, when they can still be remedied, and avoid unexpected problems later. This applies both to the implementation of the program itself, as well as the data collection process, the internal communication in the public agency about the intervention, etc. This logic of piloting and testing the intervention before conducting the large-scale program evaluation is also consistent with the practices used in the Silicon Valley type technology start-up environment, where it is often known as the “Lean Start-Up” approach (Reis 2011).

Finally, the implementation of the program or policy to be evaluated is carried out. During this step, it is important to make sure that the random assignment of individuals to each group is respected and

that no participant is moved from one group to another.[5] The most important aspect in this process is to make sure that there is no difference between the treatment and control group except the application of the program. Sometimes, well-meaning officials misunderstand the idea of the control group and think that all other interventions towards the control groups should also be halted until the study ends. However, this would amount to treating the control group differently from the treatment group. For instance, imagine a tax authority wants to test a new communication strategy, by sending specific letter messages to randomly selected group of tax payers and comparing their behavior to a control group. If officials now decided to halt all auditing activities in the control group, but continue to apply such audits to the treatment group (or vice versa), the validity of the study would be lost. In this case, the two groups would not only differ in terms of receiving the treatment, but also in terms of their risk of being audited. When looking at the final difference between the two groups, it would be impossible to establish whether the difference stems from the treatment or from the effects of the audits.

Experiences of randomized evaluations in public finance

Recent years have seen a strong increase in the use of randomized field experiments to study many different aspects of tax administration. A pioneering collaboration of this nature was undertaken by Coleman (1996), Slemrod et al. (2001) and Blumenthal et al. (2001) with the tax authority of Minnesota in the mid 1990s. Many academics have followed their example and a growing number of tax authorities, while initially reluctant, have experienced the benefits of such collaborations. In the meantime, randomized experiments have been conducted by tax authorities in the Argentina, Australia, Austria, Chile, Denmark, Ecuador, Finland, Germany, Israel, Mexico, Peru, Switzerland, USA, Venezuela (Hallsworth, 2014), and plans for such projects are underway in Kenya, Liberia, Uganda and other countries around the world.

One frequently used type of intervention consists of sending letter messages to taxpayers in order to test different hypotheses about how taxpayer behavior. The most frequently used outcome variables are amount of tax paid, since tax administrations already have access to this data, and it is the first order concern for tax administration. A growing number of recent studies have measured the impact of randomly sending letter or text messages on the behavior of individual taxpayers (Coleman 1996; Slemrod et. al 2001; Blumenthal et al. 2001; Trogler 2004, 2012; Wenzel 2005, 2006; OECD 2010; Kleven et. al 2011; Fellner et. al 2013; Hallsworth et. al 2014; Dwenger et. al 2014), such as property owners (Wenzel and

[5]In case that the randomization is not respected in the implementation process, it is possible to use the "Intent-to-Treat" methodology, and use instrumental variables to observe the "Treatment-on-the-Treated" effect. For example, this method could be used if tax payers who were supposed to be audited as a result of being in the treatment group could not found when the audit was to be carried out, or if letters were sent to the tax payer as a treatment but were not received. It is very important that even if this happens, the original random assignment is used when working with the data to conduct the impact the evaluation; that is, those that were assigned the treatment are compared to those assigned to be the control. It is never valid to compare those who were in fact treated with those with those that were meant to be treated, but that ultimately did not participate in the program, because these two groups will no longer be identical ex ante.

Taylor 2004; Castro and Scartascini 2013; Del Carpio 2013), or firms (Hasseldine et. al 2007; Iyer et. al 2010; Ariel 2012; Pomeranz 2013; Harju et. al 2013; Ortegan and Sanguinetti 2013). Some letters have tested behavioral responses to either audit threats or motivational messages. Others have evaluated the importance of the wording, such as the simplicity and clarity of the message (Bhargava and Manoli, 2011). Other studies include additional measures such as face-to-face visits (Gangl et. al 2014). For an excellent overview on the use of randomized field experiments to increase tax compliance, see Hallsworth (2014).

In collaboration with the tax authority in Chile, we employed this type of randomized letter message experiment to develop unbiased inputs for a risk indicator to predict what types of taxpayers are more likely to react to an increase in the audit probability (Pomeranz, Marshall and Castellon, 2014). Typical inputs into such risk indicators suffer a self-fulfilling circle problem: information about high evasion is typically found in types of taxpayers that were audited more frequently in the past, as a result of these audits. The risk indicators therefore end up having a self-referential problem. We developed a method that gets around this problem, by using randomized deterrence letter messages. Tax authorities can apply this method to target audit activities towards categories of taxpayers that can be expected to respond particularly strongly.

In addition to studying different communication and auditing strategies, randomized studies can also be used to study behavioral responses of tax payers to the tax structure itself. In collaboration with the Chilean tax authority, we evaluated the role of third-party information for value added tax (VAT) compliance (Pomeranz, forthcoming). The results show that the VAT can indeed have important "self-enforcing" properties. However, these properties are only activated if the audit probability is high enough that taxpayers take the risk of detection seriously. In this case, the third-party information can lead to important spillover effects that multiply the effectiveness tax enforcement measures.

Taxation is by no means the only area of public finance, in which randomized experiments play a growing role. Public procurement is another area of growth for these types of study. Projects are currently under way in procurement agencies in Brazil, Chile and Colombia among others. One of the few randomized studies in this area that has already been completed is Litschig and Zamboni (2013). They study whether a randomized increase in the audit risk deters corruption and waste in local public procurement in Brazil and find that a 20 percentage point increase in the audit risk reduces the incidence of corruption and mismanagement of local procurement by 17 percentage points.

Finally, governments may also want to study many other aspects related to the effectiveness government spending. For example in the area of savings, randomized evaluations in very different settings found (by randomly varying the savings interest rate) that subsidizing interests rates to encourage the poor are not very effective (Kast, Meier and Pomeranz, 2014; Karlan and Zinman, 2014), but that follow up and feedback messages have may be more impactful (Karlan et al 2010; Kast and Pomeranz, 2014). Stud-

ies that provided randomly selected low-income individuals access to free savings accounts may help the poor to cope with economic shocks (Kast and Pomeranz, 2014) and increase monetary assets (Prina, 2013), and increase investments in health and education (Prina, 2013; Dupas and Robinson, 2013) suggesting that there may be a role for governments to play in facilitating access to such accounts for the poor. Many studies have also been conducted in the area of education, health, and other key public expenditures.

Summary on randomized evaluations

Randomized evaluations allow for estimating the effect of a program or policy on the behavior of those affected by it. The fact that participants are randomly assigned to treatment makes it possible to measure the effect by simply comparing the outcomes of those assigned to the treatment those that were not. The counterfactual is represented by the comparison group, which is selected randomly before the start of the program, among a group of potential participants. Estimates obtained through randomized evaluations have extremely high internal validity in that they require very few additional assumptions. For this reasons, randomized evaluations are often referred to as the "gold standard" in impact evaluations. The key assumption of this method is that the randomization is valid. If that is the case, the treatment and comparison groups are in expectation statistically identical along both observable and unobservable characteristics. In addition, it is important that no other treatment is applied to one group and not the other. One practical drawback is that the random assignment has to be done before the program, and as a result, it is not possible to carry out retrospective randomized evaluations. In addition, in certain cases, random assignment to a particular treatment may not be practically, politically or ethically feasible.

The following sections describe other evaluation methods that try to construct an approximation of the counterfactual in circumstances where randomization is not possible. The validity of each method will depend on how similar the treatment group is to the control group before the intervention.

3. Simple Difference: Comparing the Treated to the Untreated

The ***simple difference*** method is one of the most frequently used methods employed to measure impacts. However, in many circumstances, its application will not provide unbiased results. This section describes how simple differences work, and what assumptions need to hold for them to be valid. Understanding the limits of simple differences will also further illustrate the benefits of having a valid comparison group in order to be able to obtain unbiased impact evaluations.

The simple differences methodology is straightforward: comparing the group that received the program with another that did not. The comparison group in this case corresponds to people or entities that did not participate in the program. That is, the assumption is that the comparison group represents a valid counterfactual of what would have happened to those who received the program, had they not re-

ceived the program. Unfortunately, in many cases, this assumption is not realistic. In many programs, there is a selection process that determines who receives the treatment. For example, consider an audit program in which only tax payers identified as high risk are selected. This assignment is not random and introduces selection bias.

To illustrate this situation with a concrete example, suppose someone wants to measure the impact of a program that offers free tutoring for children who have difficulty in school. This was the case in the study by Banerjee et al. (2007), which evaluated the effect of offering separate classes to the weakest students, where they were tutored by young women (so-called Balsakhi) in basic reading, writing and math to help them catch up with their peers. If this study simply compared the grades of children that received help from a tutor with those that did not, the results might be misleading. It is very well possible that the children with tutors would be found to have lower grades than those without tutors. However, concluding, based on this observation, that the tutors hurt the academic achievement of the kids would probably be erroneous. It is likely that selection bias was introduced prior to the start of the program: children who had lower grades might have been more likely to receive the help of a tutor. In this case, the selection bias introduces an underestimate of the impact. Because the treated group had lower grades to begin with, when comparing those that receive the help of a tutor to those who do not, it may appear that the tutoring even had a negative effect on grades.

Despite the potential serious concerns with selection bias, simple differences are often popular because they can be conducted in a retrospective manner, even after the program has been concluded, and they do not require a lot of data (for example no data on the situation of the participants prior to the program start). Newspapers and government documents therefore frequently report such difference as evidence for the benefit (or lack of benefit) of certain programs. Based on the discussion above, such statements have to be treated with much caution.

Summary on simple differences

Analysis based on simple differences measure the impact by comparing the post-treatment situation of those that participated in a program with a comparison group that did not. The counterfactual is represented by those in the comparison group. The key assumption of this method is that those in the comparison group are identical to those that participated in the program, except for the effects of the program. A key advantage, and reason for its frequent use, is that this method does not require data on the situation prior to the treatment. However, a big drawback is that if the treated and comparison groups are different in any way prior to the program, the method may be biased and may under- or overestimate the real impact of a policy; that is, selection bias is introduced into the estimation.

4. Pre- vs. Post-Treatment Period

A ***pre-post comparison*** is a particular type of simple difference evaluation. Instead of using another group as a control group, the same group of people is compared before and after participating in the program. Therefore, a pre-post evaluation measures change over time taking into account the initial state of the group. In this case, the impact is measured as the difference between outcome variables of interest before and after an intervention. The pre-post analysis is frequently used in evaluating programs. In many cases, when there is data on outcomes prior to the intervention, this type of retrospective analysis seems convenient, particularly because it does not require information on people who did not participate in the program.

In the aforementioned example of a tutoring program, a pre-post evaluation would allow taking into account the initial grades of the students. However, the important question to assess the validity of a pre-post evaluation is the following: is the situation of the participants before the start of the program a good representation of the counterfactual? In other words, is it correct to assume that without the program, during this period, there would not have been any change in the results of the treated group?

[Figure 2]

Figure 2 represents this issue graphically. In a pre-post impact evaluation, the key assumption is that in the absence of the treatment, there would have been no change over time in the outcome variable. This implies that the value of the pre-treatment situation represents a valid counterfactual for the post-treatment period.

In the free tutoring program example, it is very unlikely that the children would not have improved their learning at all over time, even in the absence of a tutor. However, a simple pre-post evaluation would assume any difference over the time span of the program is due to the impact of the program. So in this type of evaluation, even the learning due to the normal development of the children would be attributed to the tutoring program. In other words, the estimates would have a positive bias: they would overestimate the true effect of the program.

In addition to such time trends that can be expected, there can also be unexpected "shocks" that change outcomes, but are not related to the program. For example, if there is an economic crisis during the implementation period of an auditing program, tax behavior may change independently of the auditing program. It is then not possible to know if the change over time is due to the crisis, the policy, or a combination of both. That is, the evaluation may be affected by omitted variable bias.

Experiences of pre-post comparison evaluations in public finance

While in many situations, a simple pre-post comparison will lead to biased results, there are certain settings in which a pre-post analysis can yield credible results, i.e. in which the pre-treatment situation provides a valid counter-factual for the post-treatment situation. One such example is Carrillo et al (2014). This study evaluates a policy by the Ecuadorian tax authority, which focused on all firms whose declared tax filings indicate a large discrepancy between their self-reported income and information about the firms' sales that the tax authority had from third party sources. Several years after the corresponding tax filings, the tax authority sent letters to firms with a particularly large discrepancy, asking them to amend their declaration. This lead to an immediate spike in the amendment rate, while other firms were very unlikely to make any amendments such a long time after the original filing. In this case, a valid counter-factual for the amended filing of firms that received a notification is the declaration that they had filed before, i.e. the situation that would have occurred had they not filed an amendment at that time. The underlying assumption in this case is that an amendment at that particular time is due to the receipt of the notification.

Summary on pre-post comparison

Pre-post analysis measures the change in outcomes over time for participants of a program. It compares the situation before and after a treatment. The counterfactual is represented by the same participants, but prior to the program. The key assumption of this method is that the program is the only factor that influenced a change in outcomes over that time period. Without the program, the outcomes would have remained the same. The benefit of this method is it does not require information on people that did not participate in the program. An important drawback is that many factors that vary over time can affect an outcome, which contradicts the key assumption made above. In particular, the pre-post comparison does not control for general time trends or shocks that are unrelated to the program but that affect outcomes.

5. Difference-in-Differences Estimations

A difference-in-differences evaluation combines the two previous methods (simple difference and pre-post) to take into account both the differences between the two groups and changes over time. The effect is calculated by measuring the change over time for the treated group and for the comparison group and then taking the difference between these two differences.

[Table 1]

Table 1 is shows a numerical illustration of the tutoring example. It displays average grades of the children with and without the tutoring program, before and after the program (on a scale of 0 to 100). As

we can see, the treated group that receives a tutor has lower grades than the untreated group, both before and after the treatment. So a simple difference would have introduced a negative bias into the analysis. The numbers also illustrate that the grades of both groups improved over time. So a simple pre-post analysis would have introduced a positive bias. When we take the difference between the two differences, we see that the grades of those who received a tutor improved by 6.82 points more than the grades of those who did not receive a tutor.

In notation of multivariate regressions, the difference-in-differences estimator is represented by the interaction term between the treatment group and the post treatment period:

$$Y_{it} = \alpha + \beta_1 T_i + \beta_2 post_t + \beta_3 T_i * post_t + \epsilon_{it},$$

where Y_{it} represents the variable of interest for individual i in period t, T_i is a binary variable indicating whether or not individual i participated in the program, and $post_t$ is a binary variable indicating the period following the program. β_3 is the difference-in-differences estimator and ϵ_{it} represents the error term.

In essence, the difference-in-differences estimation uses change over time for the untreated group as the counterfactual for the change over time for the treated group. That is, it controls for all the characteristics that do not change over time (both observable and unobservable) and for all the changes over time that affect the treated and untreated group in the same manner.

The key assumption is that without the program, the change over time would have been the same in both groups. This is often referred to as the common or ***parallel trend assumption***. This assumption is violated if in the absence of the program, the treated group would have had a different trend over time than the comparison group.[6] These concepts are illustrated graphically in Figure 3.

[Figure 3]

In the case of the student tutoring example, the assumption implies that without the additional help, the children with a tutor and those without one would have improved their scholarly achievements at the same rate. However, it is possible that even without the program, the children who were originally behind – and where therefore more likely to receive a tutor – would have improved more than the other children, given that they had more room to improve. On the other hand, since these children had a harder time learning, it also possible that they would have fallen even further behind. That is, the difference-in-differences estimate could in this case be upward or downward biased. This is not possible to assess, since we do not know how much the children with a tutor would have improved without a tutor. That is, we cannot test the parallel trend assumption.

[6] See Meyer (1995) for a discussion of the parrallel trend assumption.

Experiences of difference-in-differences in public finance

Duflo (2001) provides great illustration of the application difference-in-differences estimation in practices. The paper takes advantage of variation in school construction across regions and time to identify the impact of building schools on school attendance in Indonesia. It illustrates well, how many assumptions need to be taken into account when conducting this type of estimation in a reliable manner.

On the topic of tax administration, Naritomi (2015) uses a difference-in-difference approach to study the effectiveness of incentives for final consumers to ask firms for a receipt, by comparing declared revenues of retail versus wholesale firms, before and after the policy change. She finds that providing consumers a financial incentive to ask for a receipt proofs to be effective in boosting declared sales and taxes by firms. Incentives in the forms of lotteries seem to be particularly effective, suggesting that consumers might be affected by behavioral biases. There is also a large literature in taxation, particularly focusing on the US and other highly developed countries, using difference-in-difference estimation to analyze the impacts of tax changes on individual behavior such as labor supply, and on firm behavior such as investment. Reviewing this entire literature is beyond the scope of this paper.

A recent study by Lewis-Faupel et al. (2014) applies difference-in-differences estimation in the area of public procurement. The study exploits regional and time variation in the adoption of electronic procurement systems across India and Indonesia in order to test the effect of e-procurement on the cost and quality of infrastructure provision. The fact that both countries rolled out the treatment gradually by region allowed the authors to carry out a difference-in-differences strategy, comparing states that were treated first with those that followed later. They find no effect on the prices paid by the government, but significant improvement in quality.

Summary difference-in-differences analysis

Difference-in-differences analysis compares the change in outcomes over time of those that participated in the program to those that did not. The change for those who do not participate in the program represents the counterfactual of the change for those that did participate in the program. The key assumption of this method is the assumption of common trends. It assumes that without the program, both groups would have had identical trajectories over time. The benefit of this method is it controls for all the characteristics that do not change over time (both observable and unobservable) and for all the changes over time that affect the treated and untreated group in the same manner. The drawback is that it is typically impossible to assess whether the two groups would have developed in the same way in the absence of the program. If this is not the case, the analysis will be biased. When longer time series of date are available, the assumption can be tested to some degree by showing that over a long pre-treatment period, the two groups had the same changes over time, and only when the treatment started did the time trends of the two groups diverge.

6. Matching Procedures and Propensity Scores

Matching procedures are based on the original objective of constructing a representation of the counterfactual and attempting to create a control group that is as similar as possible to the treatment group. The idea of matching is to construct a comparison group that is as similar as possible in terms of the observable characteristics prior to the program.

There are several matching methods. In the basic case, each individual in the treated group is matched to an individual with the same observable characteristics in the untreated group. To estimate the impact of a program, the method compares the outcomes between the treatment group and the control group, which is composed of individuals with characteristics identical to the treated individuals. Given that both groups have the same observable characteristics before the program, it is expected that the only difference after the program will be due to having been exposed to the program.

In the tutoring program example, for instance, it is possible to find children who did not sign up for the program, but who had the same grades on average as children who received the help of a tutor before the intervention. This way, a group can be created of all those that were treated and a group with identical non-treated peers; that is, individuals that are not treated but have the same observable characteristics.

[Figure 4]

Figure 4 shows the matching process with three characteristics: ages, pre-test score and gender. This is an example of a direct matching process for the tutoring example discussed above. In Figure 4, students in the treatment group are matched to children who did not receive a tutor. The matched students from the non-treated list then serve as the comparison group.

In certain cases, matching can be a better method than difference-in-differences because the process of finding peers ensures that the two groups are identical in the observable characteristics that are considered important. The key assumption in this case is that those who do not participate are, on average, identical to their matched peers, except for having participated in the program.

The question is therefore whether it is reasonable to assume that the treated group is identical non-participants that are similar according to observable characteristics. The challenge is that matching can never control for *unobserved* variables. In the tutoring program example, there is a non-random reason that two children with the same grades received a different treatment. Maybe the teacher knew that some students had more potential than others, or maybe some students had more proactive parents who were pushing for their child to receive a tutor. If there are such differences that the available data cannot measure, then the selection bias problem arises again, even though on observed characteristics, the two

matched groups are identical. It is likely, for example, that in the absence of the tutoring program, children with more proactive parents would have improved more than their classmates with the same grades.

In this context, the benefits of randomized treatment assignment become apparent. Randomized assignment ensures that the treatment and comparison groups are similar not only along observable but also along unobserved characteristics. Apart from the fact that some characteristics cannot be observed, another challenge in matching is that it requires individuals with the same characteristics in both the treated group and the untreated group. This requirement is called the common support condition. In the tutoring program example, if all students with very low grades received help from a tutor, it would not be possible to match based on grades.

Finally, the larger the number of characteristics that are included in the matching, the harder it is to use one-to-one matching. With many observed characteristics, it may be impossible to find an identical student that did not have a tutor. For these reasons ***"Propensity Score Matching" (PSM)*** was developed. PSM allows matching with many characteristics. Based on the observable characteristics of individuals, their propensity (or probability) of being in the treated group is estimated. In this way, the number of characteristics is reduced to a single score, ranging from 0 to 1, which predicts the probability of participating in the program. In effect, the propensity score is a weighted average of the included characteristics. The matching is then done between individuals that have the same score: that is, the same likelihood of participating in the program. For a detailed guide for implementing matching techniques see Imbens (2014).

Summary of matching

Matching methods compare outcomes of treated individuals with those of similar individuals that were not treated. In exact matching, participants are matched with individuals that are identical along selected characteristics but that did not participate in the treatment. In propensity score matching, participants of the program are compared to those that did not participate, and that according to their observable characteristics, had the same probability of participating in the program. The key assumption of this method is that those who participate in the program are, on average, identical to their matched peers, except for having participated in the program. It assumes that when matching people on observable characteristics, they will also be comparable along unobserved dimensions. The benefit of this method is it controls for observed characteristics that do not change over time. The drawback is that it is typically impossible to rule out that there are not also other, unobserved characteristics that differ between the groups, which would bias the impact estimation. Knowing the likelihood that unobservable characteristics will be important in this context requires fully understanding how the participants of the program were selected and what factors other than the program are likely to have affected our outcomes of interest.

7. Regression discontinuity

Regression discontinuity design (RDD) is a methodology that allows making causal conclusions that are nearly as reliable as the randomized control trial. It can only be applied in cases where a program or policy has a specific threshold that determines who is eligible to participate. An RDD uses the fact that the individuals or entities just barely above the threshold are basically identical to individuals just below. Under certain assumptions, it is therefore possible to interpret the difference between the outcomes of the individuals just below the threshold – who are therefore not eligible - and the outcomes of those just above – and who are therefore eligible.

Assume, for example, that a tax authority sends a notification letter to all firms whose declared tax filings indicates a large discrepancy between their self-reported income and information about that firms' sales that the tax authority has from third party sources. The tax authority therefore suspects these firms of cheating. However, the tax authority does not want to send out too many notifications, and decides to send notifications to all firms with discrepancies in tax obligations that are greater than $500. That is, whether a firm receives a notification is determined by whether it has more or less than $500 in discrepancies. The regression discontinuity design will then compare firms that had discrepancies a bit smaller than $500 to firms that have discrepancies just a bit larger than this cut-off.

Figure 5 displays the concept of a regression discontinuity evaluation. The solid line represents the relationship between the size of the difference and the tax amount declared: the larger the difference, the more tax is declared. Taxpayers above the cutoff value (in our example $500 in discrepancies) are included in the treatment, i.e. they receive a notification. Under certain assumptions, the sharp increase around the cutoff in the amount of taxes declared can then be attributed to the treatment.

[Figure 5]

One of the most important assumptions for the use of a regression discontinuity design is that there was no strategic change in the behavior of the firms around the threshold. If, for instance, it was known prior to the mailing of the notifications that the threshold was set at $500, then firms might be able to manipulate their discrepancy to be just below that cut-off. Those who do so may be particularly shrewd, well informed, or otherwise different. In that case, there would be a difference between the firms just under the threshold and those just above. Such a difference around the threshold again introduces selection bias. The manipulation around the threshold is referred to as a behavioral response to the threshold. The good news is that the assumption that there is no behavioral response to the threshold can be tested. If a manipulation occurred, there would be a higher concentration of firms (bunching) just below the threshold, which can be verified. In the same manner, it is possible to verify that there are no differences in the key characteristics between the firms just above or below the threshold.

Finally, a regression discontinuity design also requires that no other programs or policies are applied to the same threshold. For example, if the firms with differences greater than 500 dollars are also visited by an auditor, it would not be possible to distinguish the impact of that visit from the impact of the notification. Knowing whether other things change at the same threshold requires good knowledge the institutional details of the context in which the intervention takes place. Both problems, the behavioral response to the threshold and other policies applied to the same threshold, are more frequent when the cutoff is known by everyone. Therefore, optimal thresholds for the use of this methodology are secret, or defined ex-post, and are applied in the implementation of a single program.

One limit of RDDs is that the estimation can only be applied to observations around the cutoff. It is not possible to know what the impact was for firms with discrepancies much larger than $500, or what it would have been for firms with much smaller discrepancies. How informative the insights of the RDD are will therefore depend on the context of the policy and on the extent to which the program affects people or entities that are far away from the threshold differently.

Experiences of regression discontinuity in public finance

RDDs are of particular interest for impact evaluations in the domain of public finance, since many policies related to public finance are organized around cut-offs. In tax administration, for instance, there are many policies that are applied according to some cutoff, and frequently the administrative data required for the analysis already exists. Similarly, audit rules for public procurement, tax evasion, or the observation of other public finance regulations are often applied to certain scoring rules with a cut-off, above which entities have a higher risk of being audited.

In an ongoing study, we apply this method to procurement practices in Chile (Gerardino, Litschig and Pomeranz, 2014). We exploit a scoring rule of the national comptroller agency that creates a cutoff in the audit probability of public procurement entities at certain thresholds. The study then analyzes the impacts of audits on the public procurement process by comparing public entities that fell just below the cutoff to entities that were just above. This analysis allows the comptroller agency to continue learning and optimizing their methods.

Summary of regression discontinuity designs

RDDs compare the outcomes of individuals who are just below a threshold that qualifies them for the treatment with the results of the individuals that are just above this threshold (or cut-off). Outcomes of individual (or entities) who fall just below the threshold represent the counterfactual of the individuals who fall just above. The key assumption is that the individuals just above the threshold are identical to those who fall just below. This implies that there is no manipulation around the threshold and that no other policies are applied based on the same cutoff. This is more likely to be the case when the exact

threshold is not known ex ante. RDDs produce very reliable impact estimations. In public administration, there are many policies that are applied according to some cutoff, and frequently the administrative data required for the analysis already exists. The key weakness of RDDs is that the effect can only be estimated for individuals or entities close to the cutoff.

8. Conclusion

Rigorous impact evaluations have enjoyed a large expansion in recent years, both in their methodological developments and in their practical applications. This paper aims to provide an introductory overview for those interested in their use. Among the methods covered, randomized evaluations and regression discontinuity designs provide the most rigorous, causally valid estimates. If these methods are not available, difference-in-differences estimation or matching methods may provide an alternative. These latter methods are more likely to suffer from selection bias or omitted variable bias, and therefore have to be applied with more caution. Finally, simple differences and pre-post analysis, while being frequently applied in practice by the media or policy makers, due to their conceptual simplicity, are also the most prone to estimation biases and therefore generally the least reliable of the quantitative methods described in this paper.

In all cases, the reliability and validity of the estimates will depend to a large degree on the careful execution of the analysis, and on a good knowledge of the specific context of the program that is being evaluated. This is why the increasing number of collaborations between academics, who specialize in the rigorous analysis, and public officials, who are experts of the practical context under analysis, hold so much promise to grow the knowledge resulting from impact evaluations in public finance.

References

Abadie, Alberto, Alexis Diamond, and Jens Hainmueller. "Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program." Journal of the American Statistical Association 105.490 (June 2010): 493-505.

Abadie, Alberto. "Semiparametric Difference-in-Differences Estimators." *Review of Economic Studies*. 72. (2005): 1-19.

Abdul Lateef Jamil Poverty Action Lab (J-PAL). "Why Randomize? Case Study" (2015) www.povertyactionlab.org

Angrist, Joshua D., and Jörn-Steffen Pischke. *Mastering 'Metrics: The Path from Cause to Effect.* Princeton, NJ: Princeton University Press, 2015.

Angrist, Joshua D., and Jörn-Steffen Pischke. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton, NJ: Princeton University Press, 2009.

Angrist, Joshua D., and Victor Lavy. "Using Maimonides' Rule To Estimate The Effect Of Class Size On Scholastic Achievement." *Quarterly Journal of Economics*, (1999) 114 (2): 533-575.

Ariel, B. "Deterrence and Moral Persuasion Effects on Corporate Tax Compliance: Findings from a Randomized Controlled Trial." *Criminology* 50, (2012): 27–69.

Banerjee, Abhijit, and Esther Duflo. "The Experimental Approach to Development Economics." *Annual Reviews of Economics*. 1. (2009): 151-178.

Banerjee, Abhijit, Shawn Cole, Esther Duflo, and Leigh Linden. "Remedying Education: Evidence from Two Randomized Experiments in India." *Quarterly Journal of Economics*, 122(3): 1235-1264, 2007.

Bhargava, Saurabh and Dayanand Manoli. "Why are Benefits Left on the Table? Assessing the Role of Information, Complexity, and Stigma on Take-up with an IRS Field Experiment." Working Paper (2011).

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan. "How Much Should We Trust Differences-In-Differences Estimates?" *Quarterly Journal of Economics*. 119.1 (2004): 249-275.

Blumenthal, M., Christian, C., and Slemrod, J. "Do Normative Appeals Affect Tax Compliance? Evidence from a Controlled Experiment in Minnesota." *National Tax Journal* 54 (2001): 125–36.

Carrillo, Paul, Dina Pomeranz, and Monica Singhal. "Dodging the Taxman: Firm Misreporting and Limits to Tax Enforcement." NBER Working Paper, No. 20624, October 2014.

Castro, L., and Scartascini, C. "Tax Compliance and Enforcement in the Pampas. Evidence from a Field Experiment." Inter-American Development Bank Working Paper Series, Washington, DC, Inter-American Development Bank, 2013.

Chetty, Raj, John Friedman and Jonah Rockoff. "Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates." American Economic Review 104(9): 2593–2632, 2014a.

_____. "Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood", *American Economic Review* 104(9): 2633–2679, 2014b.

Coleman, S. 'The Minnesota Income Tax Compliance Experiment: State Tax Results', MPRA Paper No. 4827, University of Munich (1996).

Dehejia, Rajeev, and Sadek Wahba. "Causal Effects in Non-Experimental Studies: Re-Evaluating the Evaluation of Training Programs." *Journal of the American Statistical Association*. 94. (1999): 1053-1062.

Del Carpio, L. "Are the Neighbors Cheating? Evidence from a Social Norm Experiment on Property Taxes in Peru." Princeton, NJ, Princeton University Working Paper, 2013.

Duflo, Esther, Rachel Glennerster, and Michael Kremer. "Using Randomization in Development Economics Research: A Toolkit." *Handbook of Development Economics*. 4. (2007): 3895-3962.

Duflo, Esther. "Schooling and Labor Market Consequences of School Construction in Indonesia: Evidence from an Unusual Policy Experiment." *American Economic Review*. 91. (2001): 795-813.

Dupas, Pascaline and Jonathan Robinson. "Why Don't the Poor Save More? Evidence from Health Savings Experiments." American Economic Review 2013, 103.4 (2013): 1138–1171

Dwenger, N., Kleven, H., Rasul, I., and Rincke, J. "Extrinsic and Intrinsic Motivations for Tax Compliance: Evidence from a Field Experiment in Germany." Working Paper, Max Planck Institute for Tax Law and Public Finance, 2014.

Fellner, G., Sausgruber, R., and Traxler, C. "Testing Enforcement Strategies in the Field: Threat, Moral Appeal and Social Information." *Journal of the European Economic Association*, **11** (2013): 634–60.

Gangl, K., Torgler, B., Kirchler, E., and Hoffmann, E. "Effects of Supervision on Tax Compliance." *Economics Letters* **123** (3): 378–82, 2014.

Gerardino, Maria Paula, Stephan Litschig, and Dina Pomeranz. "Monitoring Public Procurement: Evidence from a Regression Discontinuity Design in Chile." Working Paper, September 2014.

Gertler, Paul Sebastian Martinez, Patrick Premand, Laura B. Rawlings, Christel M. J. Vermeersch. *Impact Evaluation in Practice*. Washington, D.C: World Bank Group, 2011.

Glennerster, Rachel, and Kudzai Takavarasha. *Running Randomized Evaluations: A Practical Guide*. Princeton University Press, 2013.

Glewwe, Paul, Michael Kremer, and Sylvie Moulin. "Many Children Left Behind? Textbooks and Test Scores in Kenya." *American Economic Journal: Applied Economics*, 1(1): 112-35, 2009.

Hallsworth, M. List, J. A., Metcalfe, R. D., and Vlaev, I. "The Behavioralist as Tax Collector: Using Natural Field Experiments to Enhance Tax Compliance." NBER Working Paper No. 20007, 2014.

Hallsworth, Michael. "The use of field experiments to increase tax compliance." Oxford Review of Economic Policy, Vol. 30, No 4 (2014): 658–679.

Harju, J., Kosonen, T., and Ropponen, O. "Do Honest Hairdressers Get a Haircut? On Tax Rate and Tax Evasion." Government Institute for Economic Research (VATT) Working Paper, 2013.

Hasseldine, J. James, S., and Toumi, M. "Persuasive Communications: Tax Compliance Enforcement Strategies for Sole Proprietors." *Contemporary Accounting Research* **24** (2007): 171–9

Imbens, Guido and Thomas Lemieux. "Regression Discontinuity Designs: A Guide to Practice." *Journal of Econometrics*. 142. (2008): 615-635.

Imbens, Guido, and Jeffrey Wooldridge. "Recent Developments in the Econometrics of Program Evaluation." *Journal of Economic Literature*. 47.1 (2009): 5-86.

Imbens, Guido. "Matching Papers in Practice: Three Examples." NBER Working Paper No. 19959 (2014).

Iyer, G. S., Reckers, P. M., and Sanders, D. L. "Increasing Tax Compliance in Washington State: A Field Experiment." *National Tax Journal*, **63**(1): 7–32, 2010.

Karlan, Dean and Jonathan Zinman. "Price and Control Elasticities of Demand for Savings." Working Paper, January 2014.

Karlan, Dean, Margaret McConnell, Sendhil Mullainathan, and Jonathan Zinman. "Getting to the Top of Mind: How Reminders Increase Saving." NBER Working Paper No. 16205, June 2010.

Kast, Felipe, and Dina Pomeranz. "Saving More to Borrow Less: Experimental Evidence from Access to Formal Savings Accounts in Chile." *NBER Working Paper Series*, No. 20239, June 2014.

Kast, Felipe, Stephan Meier, and Dina Pomeranz. "Under-Savers Anonymous: Evidence on Self-Help Groups and Peer Pressure as a Savings Commitment Device." *Harvard Business School Working Paper*, No. 12-060, January 2014.

Kleven, H. J., Knudsen, M. B., Kreiner, C. T., Pedersen, S., and Saez, E. “Unwilling or Unable to Cheat? Evidence from a Tax Audit Experiment in Denmark.” *Econometrica*, **79** (2011): 651–92.

Lee, David, and Thomas Lemieux. "Regression Discontinuity Designs in Economics." *Journal of Economic Literature*. 48.2 (2010): 281–355.

Lewis-Faupel, Sean, Yusuf Neggers, Benjamin A. Olken, and Rohini Pande. “Can Electronic Procurement Improve Infrastructure Provision? Evidence from Public Works in India and Indonesia.” NBER Working Paper No. 20344 (2014).

Litschig, Stephan and Yves Zamboni. "Audit Risk and Rent Extraction: Evidence from a Randomized Evaluation in Brazil." BGSE Working Paper 554 (2013).

Ludwig, Jens, Jeffrey Kling, and Sendhil Mullainathan. "Mechanism Experiments and Policy Evaluations." *Journal of Economic Perspectives*. Forthcoming (2011).

Meyer, Bruce D. “Natural and Quasi-Experiments in Economics.” *Journal of Business & Economic Statistics*, Vol. 13, No. 2, JBES Symposium on Program and Policy Evaluation (Apr., 1995), pp. 151-161

Miguel, Edward and Michael Kremer. “Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities,” *Econometrica*, Vol. 72 (1), pp. 159-217, 2004.

Naritomi, Joanna. “Consumers as Tax Auditors.” Working Paper, April 2015.

OECD. “Understanding and Influencing Taxpayers’ Compliance Behaviour.” Paris, Organization for Economic Cooperation and Development, 2010.

Ortega, D., and Sanguinetti, P. “Deterrence and Reciprocity Effects on Tax Compliance: Experimental Evidence from Venezuela.” CAF Working Paper No. 2013/08, 2013.

Pomeranz, Dina, Cristobal Marshall, and Pamela Castellon. "Randomized Tax Enforcement Messages: A Policy Tool for Improving Audit Strategies." *Tax Administration Review*, no. 36 (January 2014): 1–21.

Pomeranz, Dina. "No Taxation without Information: Deterrence and Self-Enforcement in the Value Added Tax." *American Economic Review* (forthcoming).

Prina, Silvia. “Banking the Poor via Savings Accounts: Evidence from a Field Experiment.” Working Paper, 2013.

Ries, Eric. *The Lean Startup: How Today's Entrepreneurs Use Continuous Innovation to Create Radically Successful Businesses*. New York, NY: Crown Business Inc., 2011.

Slemrod, J. Blumenthal, M., and Christian, C. “Taxpayer Response to an Increased Probability of Audit: Evidence from a Controlled Experiment in Minnesota.” Journal of Public Economics 79 (2001): 455–83.

Torgler, B. “A Field Experiment on Moral Suasion and Tax Compliance Focusing on under-Declaration and over-Deduction.” School of Economics and Finance, Queensland University of Technology, Discussion Paper and Working Paper Series (2012).

_____. “Moral Suasion: An Alternative Tax Policy Strategy? Evidence from a Controlled Field Experiment in Switzerland.” *Economics of Governance* 5 (2004): 235–53

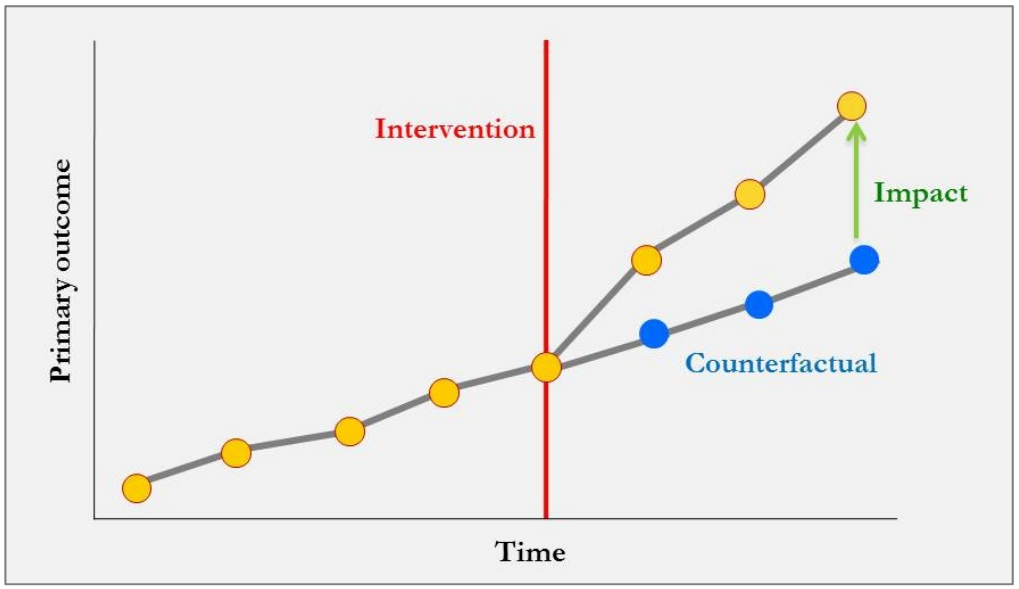
Wenzel M. Taylor, N. “An Experimental Evaluation of Tax-reporting Schedules: A Case of Evidence-based Tax Administration.” *Journal of Public Economics*, **88** (2004): 2785–99.

Wenzel, M. “A Letter from the Tax Office: Compliance Effects of Informational and Interpersonal Justice.” *Social Justice Research* 19 (2006): 345–64.

_____. “Misperceptions of Social Norms About Tax Compliance: From Theory to Intervention.” *Journal of Economic Psychology* 26 (2005): 862–83

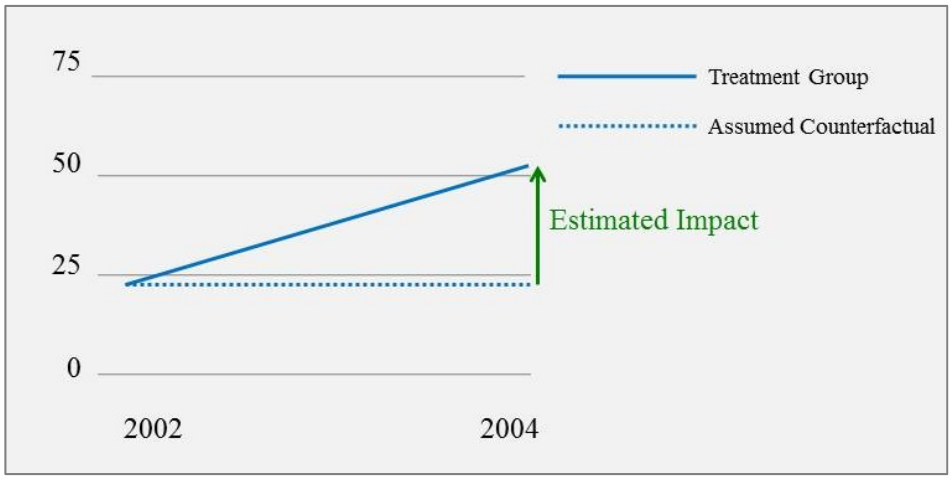
Tables and Figures

Figure 1. Counterfactual



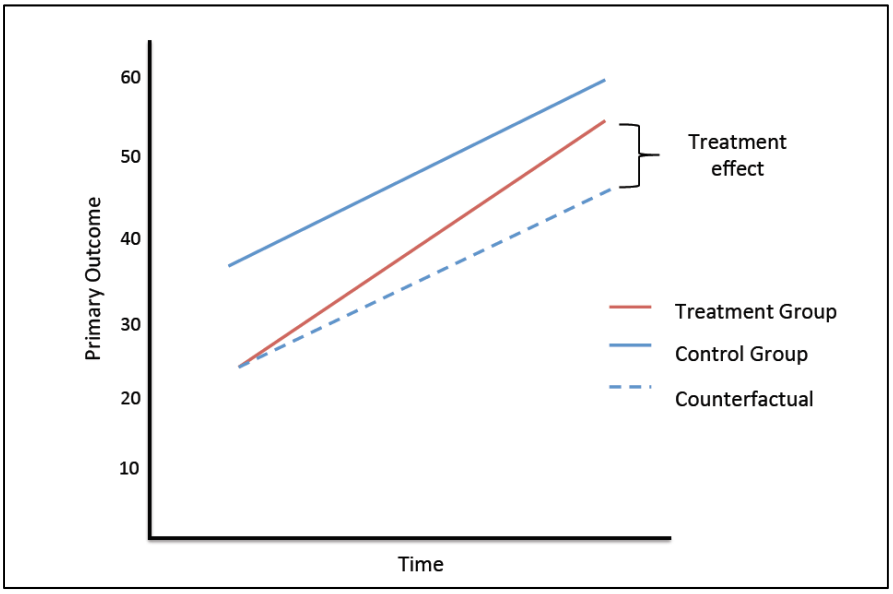
This figure represents the fundamental challenge of impact evaluation, which seeks to measure the difference between the outcome that occurred (shown in light/yellow dots) and a counterfactual that is never observed (shown with dark/blue dots). Impact evaluation techniques therefore - implicitly or explicitly - attempt to construct an estimation of the counterfactual in order to be able to measure the impact. This is often done through the use of a control group. Source: Abdul Lateef Jamil Poverty Action Lab (2015)

Figure 2. Counterfactual Assumption for Pre-Post



In a pre-post impact evaluation, the key assumption is that in the absence of the treatment, there would have been no change in the outcome variable such that the value of the pre-treatment situation represents a valid counterfactual for the post-treatment period. Source: Abdul Lateef Jamil Poverty Action Lab (2015)

Figure 3. Counterfactual Assumptions in Difference-in-Differences Analysis



This figure displays the logic and assumptions underlying the difference-in-differences analysis. The counterfactual of the change for those that did participate in the program is the change for those who do not participate in the program represents (represented by the dashed line). The key assumption is therefore that in the absence of the treatment, the two groups would have followed the same trend over time. If this holds, the treatment effect can be measured as the difference between the differences, as indicated in the figure. Source: Abdul Lateef Jamil Poverty Action Lab (2015)

Figure 4. Matching Process in the Tutor Example

Tutoring Group

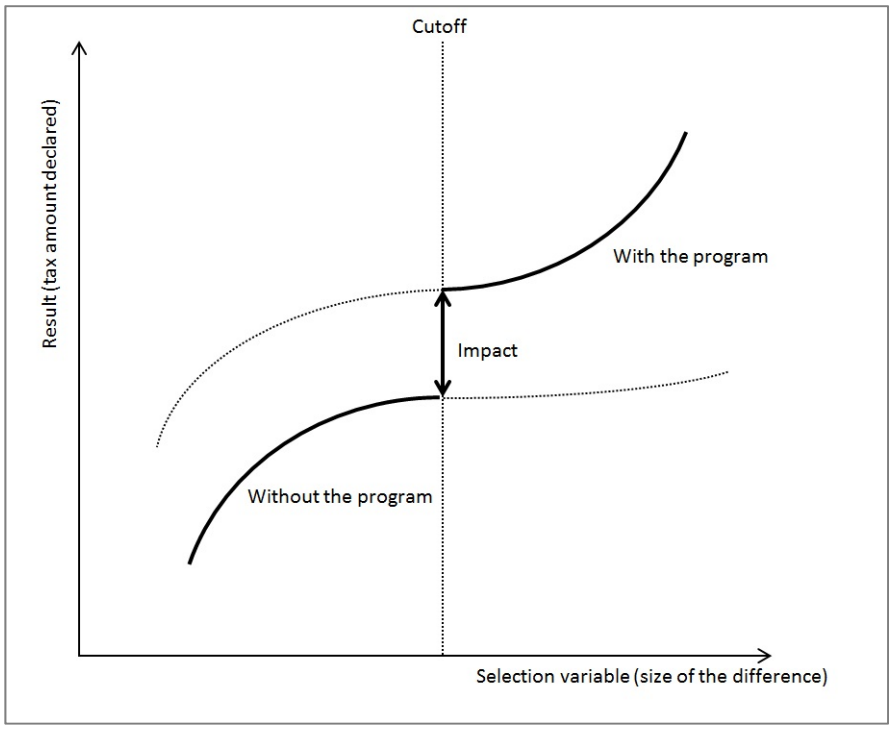
Age	Pre-Test Score	Gender
10	48	Female
10	55	Male
9	84	Male
8	14	Male
7	42	Female
10	82	Female
10	22	Female
8	53	Female
9	69	Female
8	51	Female
7	13	Female
10	62	Male

Non-Tutoring Group

Age	Pre-Test Score	Gender
10	55	Male
9	76	Female
8	81	Female
8	51	Female
10	32	Female
8	67	Male
7	64	Male
6	67	Female
10	42	Female
6	77	Male
8	93	Female
10	22	Female

This is an example of a direct matching process for the tutoring example discussed throughout this paper. This example matches students in the treatment group to children who did not receive a tutor along three observable dimensions: age, pre-test-score and gender. The matched students from the non-treated list then serve as the comparison group. Source: Abdul Lateef Jamil Poverty Action Lab (2015)

Figure 5. Illustration of Regression Discontinuity Design



This figure provides a graphical representation of an RDD. Individuals or entities above a certain cutoff value of the selection variable are included in the treatment, and those below the cutoff are not. That is, there is a discontinuity along the selection variable, above which the treatment is applied. If the required assumptions for a RDD are met, the sharp increase in the outcome variable at the cutoff can then be attributed to the treatment. Source: Abdul Lateef Jamil Poverty Action Lab (2015)

Table 1. Estimating Difference-in-Differences

	Result before the program	Result after the program	Differences
Treated group	24.80	51.22	26.42
Untreated group	36.67	56.27	19.60
Difference-in-differences estimate			6.82

This table provides a numerical example a difference-in-differences estimation. The numbers are illustrative for the tutoring example and represent grades of the children with and without the tutoring program, before and after the program. Source: Abdul Lateef Jamil Poverty Action Lab (2015)