EMPLOYMENT POLICY PRIMER

World Bank
Employment Policy Primer

December 2009 ■ No. 12

A PRACTICIONER'S GUIDE TO EVALUATING THE IMPACTS OF LABOR MARKET PROGRAMS*

1. Introduction

This note provides an introduction to the impact evaluation of labor market programs, with particular reference to developing countries. Its focus is on the main issues that need to be considered when planning an impact evaluation, including the importance of rigorous design for an evaluation, and on the statistical techniques used to estimate program impacts. To help the exposition, a prototype of a training program is referred to intermittently throughout the note. This hypothetical training program, which we call *Get-to-Work*, provides training to the unemployed to help them find work.

The note describes some general issues that are important for any impact evaluation of employment programs, both in the design and analysis stages, regardless of the specific evaluation techniques used. It then describes the main evaluation techniques, including the data requirements and the main assumptions invoked by each.

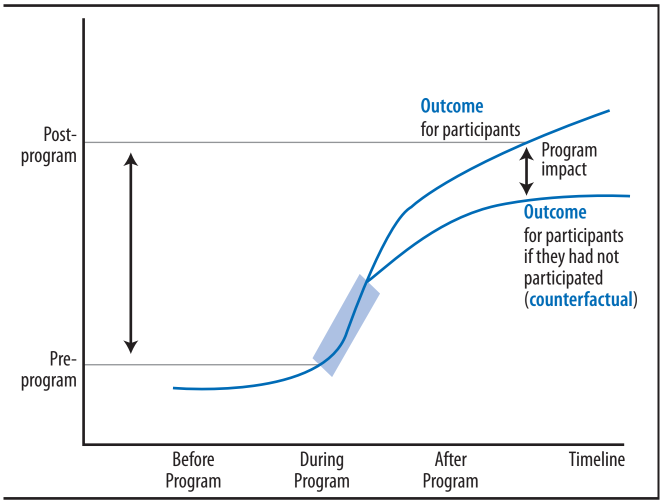
2. The evaluation of employment programs

An impact evaluation of a labor market program is a quantitative analysis that specifies one or more outcome variables of interest (for example earnings, employment) and that estimates the difference a program makes to the levels of these outcome variables. Figure 1 depicts this graphically. So an impact evaluation of *Get-to-Work*, for instance, might estimate the average earnings of participants, and the average earnings that participants would have earned had they *not* participated in the program. The difference between the two estimates of average earnings would be the impact of *Get-to-Work* on earnings.

2.1. The Counterfactual

Assume that participants in the *Get-to-Work* program earn an average of $200 per week six months after the program starts. One might be tempted to think that this is the impact of *Get-to-Work* on earnings six months after the start of the program. However, some

Figure 1: The Evaluation Problem



Note: The impact of a program is the difference in the outcome variable for participants and what participants would have obtained if they had not participated.

* Prepared by Emla Fitzsimons (Institute for Fiscal Studies, UCL) and Marcos Vera-Hernández (Institute for Fiscal Studies, UCL,). Peer reviewed by Silvia Paruzzolo (HDNSP, World Bank), Patrick Premand (HDNSP, World Bank) and Rita Almeida (HDNSP, World Bank)

participants could have found a job and earned money even if they had not participated in *Get-to-Work*. Let's say that the average earnings of participants had they not participated in *Get-to-Work* is $130. In this case, the average impact of *Get-to-Work* is $70, which represents the difference the program makes to the average earnings of participants.

We note that estimating the average earnings of participants ($200 in the above example) is relatively straightforward. It involves using observed data to compute the average earnings of participants. However, estimating the average earnings of participants *if they had not participated* (usually called the counterfactual, and equal to $130 in the above example) is considerably more complicated, as it is not observed. To get around this, one uses data on the earnings of non-participants. However, in general one cannot expect the average earnings of non-participants to be the same as what participants would have earned had they not participated. This is because participants and non-participants are usually different in terms of their background characteristics. So the difference in earnings between participants and non-participants not only reflects the impact of the program (the objective), but also differences that are unrelated to the program (i.e. differences in background characteristics that yield a bias).

2.2. Sources of Bias

One important type of bias is what is known as selection bias. This may occur when, in comparing the outcomes of participants and non-participants, the analyst fails to observe some particular background differences between the two groups that affect their outcomes. In our example, imagine that the participants and non-participants of *Get-to-Work* are identical in most characteristics except motivation, with non-participants being less motivated than participants. For this reason, let's assume that the earnings of non-participants are $20 lower than the earnings participants would have had in the absence of the program, so $110 (= $130-$20). What will be observed in the administrative or survey data are the average earnings of participants ($200) and the average earnings of non-participants ($110). The true counterfactual of $130 is not observed. So by simply subtracting the average earnings of non-participants from those of participants, one obtains a program impact of $90 (=$200-$110). This comprises the true impact of the *Get-to-Work* program ($70) and the impact of differences in motivation ($20). The latter is a bias. In reality, it is not possible to separate out the true impact and the bias using data in this way.

A different type of bias emerges when the impact is estimated by comparing the outcomes of participants only, before and after the program. In order to understand this source of bias, imagine that the Get-to-Work participants earned $190 on average before starting training and, as above, earned $200 after finishing training, say a year later. The estimated impact of $10 obtained by comparing earnings before and after participation in *Get-to-Work* underestimates the true impact of the program of $70. The bias is due to the fact that earnings can vary from year to year for many reasons, such as prevailing economic conditions, so one cannot attribute any differences between the two years to the program only. Note again that with actual data one would not be able to pinpoint this bias because the true impact is unknown—it is what we are trying to estimate!

There are different evaluation techniques that can be applied to data in order to estimate the counterfactual and thus the impact of a program. Before entering into a discussion of the different techniques, a summary of the issues that are common across most evaluation techniques is provided.

3. Issues common to most evaluation techniques

3.1. Individual- and cluster-level evaluations

Two basic schemes can be used to evaluate a program: an individual-level or a cluster-level evaluation. There is no hard and fast rule as to which approach is more appropriate. It depends on such issues as political considerations, the nature of the question being asked, the extent of externalities, and logistical concerns. The individual-level approach consists of using data from participants and non-participants living in the same communities. So within each community, one must observe both participants and non-participants.

In a cluster-level analysis of labor market programs in developing countries, a cluster is generally a geographical concept, such as a village or community. This type of analysis requires one to distinguish between treatment communities (communities where the program is in place by the time of the evaluation) and control communities (communities where the program is not in place by the time of the evaluation). This distinction comes naturally in a pilot phase of a program, before it is rolled out nationally. A sample of individuals that verifies eligibility requirements is drawn from both types of communities. It is very important that the sample is collected in the same way in both types of community (that is, that the same sampling scheme is followed). This means that the sample is not based on whether an individual participates or not in the program, as we only know this in treatment communities and not in control ones. Note that the results of a cluster-level evaluation refer to the group of eligible individuals and not to the group of participants. This is because one cannot know who the participants would have been in control communities. Consequently, if not many participants take up the program, there will be little change in the outcomes of the set of eligible individuals living in treatment communities, and the estimated impact will be small. This is intuitive: if a program attracts few people, the difference that the program makes—its impact—is likely to be small.[1]

A cluster-level evaluation is generally more expensive than an individual-level one because it requires a larger sample and the collection of data from a significant number of communities.[2] However, there are at least two important instances when it is more appropriate than an individual-level evaluation.

The first is the existence of spillovers from participants to non-participants. Let's take a couple of examples of how these might occur and their implications: (1) Assume that the evaluation of *Get-to-Work* is carried out at the individual level by comparing earnings of participants and non-participants living in the same communities. Participants are likely to chat to non-participants and it is not inconceivable that they convey to them some of what they have learned in the training sessions. If this transfer of information, or spillovers, is very strong, it is plausible that there might be no difference in the skills of participants and non-participants, and that the earnings of both groups increase by the same amount as a result of *Get-to-Work*. This would lead one to the erroneous conclusion that the program has no impact on earnings. (2) Spillovers might also occur through the labour market. In small labour markets, non-participants may be harmed by the fact that a group of individuals is participating in *Get-to-Work*. For example, employers might now prefer to employ individuals who participate in *Get-to-Work* and as a result the demand for and earnings of non-participants might drop. Let's say they drop by \$20. In our example, this means that the earnings of non-participants are \$110 rather than \$130, and yields an estimate of the impact of the program of \$90 (= \$200-\$110) rather than \$70 for participants and –\$20 for non-participants.

The second instance where a cluster-level evaluation might be preferred to an individual-level one is when it is not politically feasible to exclude some eligible individuals in the community from participating in a program. In other words, it might not be feasible to have participants and non-participants living in the same community. In general, one would choose a cluster-level evaluation for programs that are to be implemented in small communities and where take-up of the program is expected to be reasonably high.

3.2. Treatment and Intention to Treat

There are two types of impacts that an evaluation can estimate: the impact of the treatment itself, and the impact of being offered the treatment (regardless of whether it was received or not). The latter is called the intention to treat impact. These two impacts are different because not everyone who is offered it ends up participating in the program. It might depend on the way the program is marketed, targeted and delivered. The impact of *Get-to-Work* could be reasonably high, but the intention to treat might be low if take up of the program is low. Clearly, these effects are important for policy and this is why the intention to treat impact might be interesting in many settings. In some circumstances, upon being offered to participate in *Get-to-Work*, the individual might decline to participate but the offer itself might motivate him/her to look for alternative training possibilities. In such an instance, *Get-to-Work* has a positive intention to treat impact on the individual even if the individual does not participate in *Get-to-Work*.

[1] The lower the take-up rate, the larger the sample required to estimate the impact. To understand why, assume that only 5% of the eligible individuals participate in Get-to-Work. This means that the impact of the program amongst eligible individuals would be \$3.50 (=0.05 x \$70). As this is a small number, one would need a large sample to obtain an estimate that is statistically different from zero.

[2] The larger sample required is not only related to the possibility of low take-up of the program, but also to the existence of variables that vary at the community level and that affect the relevant outcomes.

3.3. General data requirements

Unfortunately there is no simple data checklist that applies to all evaluation techniques, though most techniques have some overlap. First, data are required on the outcomes of interest. And second, data on background characteristics, or 'covariates', are generally necessary. These two data requirements are discussed in more detail below. Any other technique-specific data requirements are outlined in the relevant parts of section 4.

3.3.1. Outcome variables

It is useful to consider two types of variables: outcomes and covariates. Outcome variables are the ones that policymakers have in mind as the ultimate goals of the program, i.e. the actual impacts/benefits/changes for participants during or after the program. In most evaluations of labour market programs, the final outcome variables include earnings and employment. They could also include some variables that measure job quality, such as whether the employment contract is written and formal, and whether it includes social security. It is also very important to track intermediate outcomes, which can be thought of as variables that could be directly affected by the program and that are a first step towards achieving the final outcomes. They are sometimes called *mechanisms/channels*. In the context of *Get-to-Work*, these intermediate outcomes might include skills, contacts in the labour market, and motivation to search for a job, for instance. Understanding the impact of the program on these intermediate outcomes might be useful in order to understand how the program is making a difference (channels or pathways), and might help one to improve the program or to understand how feasible it is to extrapolate it to other settings.

3.3.2. Covariates

As mentioned in section 2, differences in background characteristics between participants and non-participants might yield biased estimates of the impact of the program. Covariates, often referred to as observed background characteristics or control variables, can affect directly the outcome variable of interest, and can also affect the relationship between the program and the outcome. So they are controlled for in the analysis in order to strip out their effects from the estimated impact. In contrast, unobserved background characteristics are variables that are not included in the analysis, usually because the information is not available (either because it is difficult or costly to measure).

One important requirement is that the covariates used in the analysis are not affected by the program, or if one suspects that they are, that pre-program values of the covariates are used instead. Otherwise one might underestimate the effect of the program. To take an example, one might be tempted to use house conditions (e.g. construction materials of the house) measured during or after a program in order to purge the impact estimates of the effect of background characteristics (here, house conditions, which may be a proxy for wealth). However, the potential problem with this is that participants may improve house conditions with the extra earnings that they obtain due to the program. In this case, using dwelling conditions during or after the program may underestimate the impact of the program. However, there is no problem in using measures of dwelling conditions from before the program starts, as the program could not have affected them. In fact, one of the most important reasons for carrying out a baseline (or pre-program) survey is to collect a large set of variables that could not have been affected by the program and that can be used as covariates in the impact evaluation. In the absence of a baseline survey, the set of potential covariates will be much smaller as it is feasible that many variables may have been affected by participation in the program.[3]

Covariates can also be put to other uses including (1) to estimate whether the impact varies for different groups of the population (as grouped according to values of particular observed characteristics), (2) to obtain a measure of how different participants and non-participants are in observed background characteristics, (3) to corroborate hypotheses on the nature of sample selection (for instance, if one believed that participants were more motivated than non-participants, this belief would be backed up by observing that participants had attended more training courses than non-participants before the start of *Get-to-Work)*.

It is very important that information on outcomes and covariates is obtained from the same sources and

[3] In the absence of a baseline, one could ask retrospective information on housing conditions before the start of the program. However, in practice, the collection of retrospective data might be affected by a recall bias. Pre-program administrative records, where available, are often good sources of covariates.

in a consistent manner for both participants and non-participants. In relation to this point, one should never rely on program guidelines as a reliable source of data! For instance, even if the program is supposed to provide work to participants for 4 hours per day, in reality guidelines are unlikely to be followed to the letter. So it is important to collect information on hours worked, from participants as well as non-participants. Another reason that the same questionnaire should always be applied to all, regardless of participation status, is that it avoids so called "framing effects".[4]

4. Techniques for impact evaluation

In this section, the different techniques to estimate the counterfactual, and hence the impact of a program such as *Get-to-Work*, are described. It starts with the experimental technique of randomization, often considered to be the gold-standard of evaluation techniques. This is because participants and non-participants are chosen at random, so any differences in their outcomes can be attributed to the program. It then describes quasi-experimental techniques in which the allocation of the program between participants and non-participants is not random but is decided by the individuals themselves together with politicians, policymakers and/or the institutions running the program. It is important to understand the objectives and logistical constraints of those who are involved in allocating the program between participants and non-participants because this can often help the analyst to assess in what particular dimensions participants and non-participants are likely to be different. Quasi-experimental techniques, which include matching, difference-in-differences, instrumental variables and regression discontinuity, are used to overcome the problem that participants and non-participants may be different for reasons other than program participation, and so differences in their outcomes cannot be attributed solely to the program.

4.1. Randomization and its variants

This technique, in its pure form, allocates participants and non-participants randomly. One could think of it as tossing a coin to decide who the participants and non-participants are. The main advantage of randomization is that, because (if carried out properly and assuming the sample is large enough) the only difference between participants and non-participants is the toss of a coin, there are no differences in background characteristics between them. In this case, a simple difference in means between participants and non-participants estimates the impact of a program on the outcomes.[5]

However, sometimes political considerations get in the way of randomizing: no government wants to become unpopular by excluding eligible unemployed individuals from a training program (even though the program has not yet been evaluated so it is not known whether it works!). A more politically palatable approach is to withhold the program from controls for a set period and then extend it to them. This at least allows for an evaluation of short-term effects. That aside, when operationally feasible, randomization is the preferred technique for impact evaluation because it estimates the causal impact of a program under weak and very plausible assumptions. However, when carrying out the randomization, it is important to choose program participants who are representative of the population that would be targeted were the program to be rolled out on a large scale. For instance, it is better to randomize among the group of individuals who have expressed an interest in participating in the program, than to randomize among individuals listed in an unemployment register, who may have very different characteristics from the population that would be targeted by the program were it rolled out after the evaluation.

When access to the program is universal, a group of individuals cannot be excluded from participating in the program in order to serve as a control group. Moreover, as alluded to already, governments may not want to risk popularity by excluding eligible people from the program. In such circumstances, an encouragement design would allow one to still carry out a credible evaluation.

[4] Framing refers to the manner in which questions are posed. Naturally, there will be instances where some questions may not be applicable to non-participants. For example, if the program involves making payments to participants then any questions relating to these payments will not apply to non-participants.

[5] When the evaluation is carried out at the cluster level, it is entire clusters (for instance, communities) that are randomized in and out. In this case, if there are few clusters, one could end up with differences in background characteristics between clusters. One way to reduce the risk of this happening is to first form pairs of clusters, with each pair consisting of the two clusters most similar to each other in terms of observed characteristics. Then the treatment could be allocated randomly within pairs (so within each pair, one cluster would receive the treatment and the other would not).

The encouragement design consists of selecting a group of individuals and then randomizing an incentive to participate, or providing extra information about the program to this group, so that the subgroup that ends up with the incentive has a higher probability of participation.[6,7] For instance, in the context of *Get-to-Work*, a social worker could visit the homes of a randomly selected group of people on various occasions, inform them of the availability of *Get-to-Work,* provide them with information about the program, and ultimately encourage them to participate in it. One would expect the group of people that received the visit to be more likely to participate than the group of people that did not receive it.

Another variant is to randomize the treatment allocation within a group of applicants, and to allow individuals who are randomized out to participate only if they follow a time-consuming administrative procedure (see Box 1). Another variant is to allow drop outs to be replaced by individuals who were randomized out.[8] Both of these variants require larger sample sizes than pure randomization.

Strictly speaking, the randomization method only requires the collection of data on outcomes, as covariates are "in theory" not required to net out differences in background characteristics (there should be none!). However, it is advisable to collect information on covariates, not only to show that individuals or clusters that were randomized in and out do indeed have the same background characteristics, but also to allow the analyst to estimate the impact for different demographic groups (heterogeneous effects).[9] The use of covariates can also improve the precision of the estimates. In terms of other data that should be collected, information for each individual on whether (s)he is a participant or not and whether (s)he was randomized in or randomized out, is important. Finally, the collection of a baseline survey is not strictly necessary but is advisable particularly in cases where the program is randomized among a small group of clusters, as differences in background characteristics can still occur by chance and data from a baseline survey can help to correct for them using the methods that will be explained in section 4.3.

BOX 1: An example of Randomization

Training Disadvantaged Youth in Latin America: Evidence from a Randomized Trial, by Attanasio O, Kugler A, and Meghir C. 2008

What: Evaluate the impact of a training program for young people. The program combines in-class training with an internship in a firm

Where: Colombia

How: The program is oversubscribed. A random process divides applicants into participants and non-participants. The latter are allowed to participate only if they follow a time-consuming administrative procedure

Findings: The program raises earnings and employment for both males and females. The effects are larger for females

4.2. Matching

Matching estimates the counterfactual using data on non-participants, but giving more weight to non-participants who are more similar to participants, as measured by observable background characteristics (See Box 2).[10] Two steps are required to estimate the impact of *Get-to-Work* using matching. In the first step, the analyst discards participants for whom there is no *similar* non-participant. This has the implication that the impact estimates will only be representative of individuals who have not been discarded. In the second step, a weighted average of the earnings of non-participants is computed. As discussed, this average (which is the estimate of the counterfactual) is weighted in such a way that non-participants who are relatively more *similar* to participants receive a larger weight than those who are relatively less similar to participants. The impact is obtained from the difference between the average earnings of participants who were not discarded in the first

[6] When these variants are used, estimates are obtained through instrumental variables techniques. The instrumental variable in this instance is whether or not the individual was (randomly) encouraged to participate (see section 4.4. for a description).

[7] Unlike pure randomization, this design will not identify a representative effect of the population of participants. Rather, it will estimate the effect of the program for those individuals who participate with the incentive but who would not have participated without it. This might be a problem if one believes the impact to be very heterogeneous.

[8] Card *et al.* (2007) use this variant to evaluate a training program for youth in the Dominican Republic

[9] Moreover, if the data is rich in content it might be used to answer other interesting questions apart from the impact of the program. This is important because data collection usually has a high fixed cost.

[10] Matching is used by, for example, Jalan and Ravallion (2003) to estimate the impact of a workfare program in Argentina, and by Rodríguez-Planas and Jacob (2009) to estimate the impact of four types of active labor market policies in Romania (Training and Retraining, Self-Employment Assistance, Public Employment, and Public Employment and Relocation Services).

BOX 2: An example of Matching

Estimating the benefit incidence of an antipoverty program by propensity score matching, by Jalan, J and Ravallion M. 2003

What: Evaluate the impact of a workfare program in which participants work at relatively low wages in socially useful projects located in poor areas

Where: Argentina

How: A sample of participants, and a national sample that contains information on non-participants, are used. Similarity between participants and non-participants is measured by their estimated probability to participate (propensity score)

Findings: The program increases on average the earnings of participants by about half of the gross wage

step, and the estimate of the counterfactual obtained in the second step. There are many matching techniques and they differ in both the way they define similarity and the way the weights are computed. Common to all of the techniques is that similarity is defined using the values of the covariates. In general, the similarity between two individuals is measured by the differences in the values that their covariates take.

The crucial assumption in matching is that the set of covariates that the analyst uses, that is, the observed background characteristics, includes all of the variables that simultaneously affect both outcomes and program participation.[11] For instance, if participants in *Get-to-Work* are more motivated than non-participants, and it is true that more motivated individuals have higher earnings, then matching will not remove the selection bias described in section 2.2. This is why matching methods generally call for including a large set of covariates, and hence a baseline survey is highly desirable (see section 3.3).[12] Labor market histories are good candidates to include in the set of covariates because it is widely believed that they are a good proxy for variables that are important determinants of labor market outcomes but that are difficult to measure (such as motivation).

4.3. Difference-in-differences

Difference-in-differences is a technique that recognizes that participants and non-participants would have different average outcomes even in the absence of the program (see Box 3). In the *Get-to-Work* example, participants earn $20 more than non-participants even in the absence of the program due to their higher motivation.[13] The underlying assumption behind the technique is that differences in outcomes due to differences in unobserved background characteristics ($20) do not change over time. Consequently, one can account for them as follows. Differences unrelated to program participation (i.e. motivation) are estimated by going back in time and measuring the outcome variables of non-participants and would-be- participants, before the program started. As already discussed, *Get-to-Work* participants earned $190 in the year before the program started. As differences in motivation between them and non-participants remain constant over time, non-participants must have earned $170 (i.e. $20 less than participants) in the same period. The difference between these two pre-program earnings measures, $20 (= $190 – $170) is the so-called *pre-existing difference*. For ease of exposition, let's define the *contemporaneous difference* as the difference in the outcome variable between participants and non-participants at the time that the impact is estimated (while the program is active or when it has

BOX 3: An example of Difference-in-differences

Training Jump-Starting Self-Employment? Evidence among Welfare Participants in Argentina, by Almeida, R and Galasso E.

2007

What: Evaluate the impact of a program that promotes self-employment by providing technical assistance as well as grants for inputs and equipment

Where: Argentina

How: Use data for both participants and non-participants, both before and after the program. An important factor that explains non-participation is that some municipalities delayed their participation in the program

Findings: The program increases on average participants' total number of hours worked by just under 15

[11] Any unobserved background characteristic must not affect either the probability of participation in the program or the outcome variable (or both).

[12] However, one must note that this is at some cost. The more covariates that are included in the model, the more likely it is that there are participants for whom one cannot find a counterpart in the set of non-participants. This in turn restricts the set of individuals for whom the estimate of the program impact is representative.

[13] This could be either because there are differences in the socio-demographic composition of the group of participants versus non-participants, or differences in the effect that these variables have on outcomes. For instance, both of the following should have remained constant over time: (1) the difference between participants and non-participants in the percentage of individuals with secondary education , and (2) the difference between participants and non-participants in the returns to education.

finished). As explained in section 2, in general, the contemporaneous difference $90 (= $200 – $110) includes both the impact of the program and any other differences that are unrelated to program participation (i.e. motivation). If one assumes that the latter has not changed over time, one can subtract the estimate of *pre-existing difference* ($20) from the *contemporaneous difference* ($90) to obtain an estimate of the impact of the program ($70). This method of subtracting the *pre-existing difference* removes the bias as long as the assumption that the differences unrelated to program participation have remained constant over time holds. The basic method of difference-in-differences can be strengthened by using covariates in a regression or matching framework.[14]

In order to compute the *pre-existing difference,* the analyst needs to know who the participants and non-participants are at baseline, before the program has started. If panel data are going to be collected, one can wait until the follow-up survey (i.e. during or after the program) to collect this information. However, it should be noted that the difference-in-difference method does not always require the same individual to be followed over time (panel data). In some circumstances, one can perform the analysis using two repeated cross sections as long as one can identify at baseline who will go on to be participants and non-participants. For instance, sometimes participants are individuals in a particular age group living in certain communities, and non-participants are individuals in the same age group living in a different set of communities. In other instances, participation and non-participation status depends on membership of certain organizations. In these examples, one readily knows who the participants are, even at baseline before the program has started.[15] This is particularly useful when the analysis is being carried out using household surveys that are not exclusively collected for the evaluation of a given program.

A common problem in the evaluation of labor market programs using difference-in-differences is the so-called *Ashenfelter dip* (Ashenfelter, 1978). This occurs when non-participants are drawn from the pool of individuals that does not apply to the program, and participants apply to the program because they have experienced a temporary drop in earnings. When this is the case, one risks over-estimating the pre-existing difference, which in turn biases the impact estimate. For this reason, it is advisable to draw the pool of non-participants from the pool of applicants. In this way, the pre-program drop in earnings will be similar across both groups and thus the pre-existing difference between their earnings will not be overestimated. It is also advisable to collect retrospective income and labour supply data, say from the previous year, so that the pre-existing difference is estimated using data that is relatively distant from the time of application to the program (and thus less likely to be affected by the impending program). Instrumental variables, discussed next, can also be used to solve the *Ashenfelter dip* (see Almeida and Galasso, 2007).

4.4. Instrumental Variables

Instrumental Variables is a technique that recognizes that participants and non-participants would have different average outcomes even in the absence of the program, most likely because they differ in background characteristics (see Box 4). Its underlying assumption is that there is at least one variable, called an instrumental variable, that (1) predicts program participation, (2) only affects outcomes through affecting program participation, and (3) is not correlated with the unobserved background characteristics that affect outcomes. The two last conditions cannot be tested and must be assessed on a case by case basis.[16] The methods that will be outlined in the postscript would be useful to analyze their plausibility. It must be emphasized that the evaluation will only be as good as the instrumental variable is in the sense of verifying the conditions above. A rich set of covariates will make it more likely that the last two conditions can be met.[17]

Though the instrumental variable can be continuous, the logic of the instrumental variable technique can be more easily understood by considering a binary instrument. In the context of *Get-to-Work,* assume that

[14] Almeida and Galasso (2007) use difference-in-differences within a regression framework to evaluate a program that promotes self-employment in Argentina, and Díaz and Jaramillo (2006) use difference-in-differences within a matching framework to evaluate a training program for youth in Peru.

[15] When this approach is used, one should check that the composition of participants and non-participants in terms of socio-demographic characteristics has not changed over time.

[16] Note that if the instrumental variable is correlated with many observed variables, it will be difficult to argue that it is not correlated with any unobserved variable.

[17] Bartik (2002) uses this technique to estimate the effect of state level welfare caseload on wages and unemployment rates of low skilled individuals.

there are two type of communities, ones where the program registration office is very centrally located, and others where individuals must travel very far to register and thus to participate. One would naturally expect program participation to be higher in communities where it the office is centrally located (condition 1 above). Assume further that travelling long distances for a training program does not affect the determinants of labor market outcomes apart from the effect it might have on participation in the program (condition 2), and that the unobserved characteristics of the labor markets and of the individuals living in communities where the office is centrally located are identical to those in communities where it is not. Under such circumstances, one can build the *Wald estimator* which is the simplest version of an instrumental variable estimator.[18]

4.5. Regression Discontinuity

Sometimes, eligibility to participate in a program is determined on the basis of the value of an index. For instance, each individual in our *Get-to-Work* example could have an index which would take the value 0 for the poorest individual and 100 for the richest. The index level of most individuals would lie somewhere between 0 and 100, depending on their poverty level. Individuals would be eligible to participate in *Get-to-Work* if their index level is below a particular threshold, such as is determined by the program administrator. This implies that there is a large fall (or discontinuity) in the participation rate from households with a value slightly below the threshold, to households with a value slightly above the threshold.[19]

The logic of regression discontinuity is as follows. If there is a fall in the percentage of participating households in *Get-to-Work* at the threshold, and the program has an effect on labor market outcomes, then there should also be an abrupt change in labor market outcomes at the threshold. To implement this technique, one requires data on the outcomes of interest, the value of the index, the participation threshold, and whether or not the individual (or whatever the unit of observation) has participated in the program. In order to have sufficient power, large enough samples around the threshold are required (see Box 5).

BOX 4: An example of Instrumental Variables

Instrumental Variable Estimates of the Labor Market Spillover Effects of Welfare Reform, by Bartik, T 2002

What: Evaluate the impact of state level welfare caseload on wages and unemployment rates of low skilled individuals

Where: United States

How: Estimate instrumental variable regressions of wages and unemployment rate on welfare caseload, using labor market and political variables as instruments

Findings: Find some evidence of reduced employment and wages

BOX 5: An example of Regression Discontinuity

Can Entrepreneurial Activity be Taught? Quasi-Experimental Evidence from Central America, by Klinger, B and Schündeln M. 2007

What: Evaluate the impact of a program that provides entrepreneurship training

Where: El Salvador, Guatemala, and Nicaragua

How: Each applicant to the training program is given a score based on his/her entrepreneurial ability. Individuals with scores above a threshold participate in the training. The impact is estimated through the abrupt change in outcomes that occurs at the threshold

Findings: Participating in the training program significantly increases the probability of starting a new business or expanding an existing one

One advantage of this technique is that policy makers usually find it easy to justify because it favours those that need the program the most. However, some caveats are in order. First, the impact estimate is only informative for households with a value of the index around the threshold. Thus it is difficult to extrapolate findings to the general population. On this note, it requires big enough samples around the cut-off point to have sufficient power. Second, it is important that no other program induces a fall in participation at the threshold. Otherwise, one would inadvertently be considering the effect of all such programs. Finally, one should be wary of households or

[18] Intuitively, the Wald estimator takes the effect of the instrument on the outcome and divides it by the effect of the instrument on participation. In the example here, the numerator is the average earnings of eligible individuals living in communities in which the Get-to-Work office is centrally located, minus the average earnings of eligible individuals living in communities in which it is not. The denominator is the difference between the percentage of participants in communities where it is free and communities where it is not.

[19] Klinger and Schündeln (2007) use this technique to estimate the impact of a entrepreneurial training program on enterprise outcomes.

authorities misreporting the components of the means test in order to score lower and become eligible to participate. If such manipulation is present, which is usually manifest in a significant number of households with a score just below the threshold, then one should use household surveys that have not been collected for the specific purpose of determining eligibility.

Postscript: A tip for assessing the validity of the evaluation strategy

The analyst, facing a menu of evaluation techniques, is likely to meet with scepticism or even criticism at some point regarding the evaluation method chosen. Assumptions can always be questioned and challenged. It is important to include plenty of descriptive statistics in the analysis, to examine carefully pre-program as well as post-program data, and to look for support for assumptions invoked by the particular evaluation method used. But another way to assess the credibility of results is to repeat the exact same evaluation method, but using as an outcome variable one that should *not* be affected by the program! A particularly neat procedure is to take an outcome variable considered in the evaluation but relating instead to a period *before* the program started (i.e. the lagged value of the outcome).[20] If the evaluation strategy is indeed reasonable, one should find that the program has no effect on the lagged outcome variable, as the program was not in operation then. If one finds that it does have an effect, it suggests strongly that one of the main assumptions invoked by the technique is being violated. This approach calls for collecting retrospective information on outcome variables.

5. Conclusion

Readers of this note are no doubt well aware of the vast array of potential labor market programs and policies, not least of all in developing countries. But there is a need to know which programs work best, so that scarce resources can be allocated optimally. This note has been motivated by the importance of credible impact evaluations of labour market programs for guiding international organizations, policymakers, donors, and other potential stakeholders, as to which programs and policies work best.

Program design is of the utmost importance, not just for effective implementation and delivery of programs, but also for impact evaluation: the program should be designed in such a way as to allow for its credible evaluation. A pilot stage is important in order to test whether the program works and to quantify its effects. This is also an opportunity to try out different variants of a program in order to compare them and see which works best. Randomization is widely considered to be the gold standard of evaluation. When done properly, it makes the construction of the counterfactual—the situation *without* the program—very straightforward. But not all programs can be randomized, and this note has also discussed the main evaluation techniques used in instances when randomization is not possible. These evaluation techniques are extremely useful to know *before* the program is piloted, as they can influence strongly the types of data that need to be collected in order to evaluate the program properly, and the optimal way of designing the pilot. In essence, deciding which labor market programs to implement with limited resources can only be done with careful planning and rigorous evaluation.

[20] For the difference-in-differences technique, one would need this outcome to be measured two periods before the program started.

6. References

Almeida R, Galasso E. Jump-Starting Self-Employment? Evidence among Welfare Participants in Argentina. IZA Working Paper 2007, number 2902

Ashenfelter O. Estimating the Effect of Training Programs on Earnings. *The Review of Economics and Statistics* 1978; 60: 47–57

Attanasio O, Kugler A, Meghir C. Training Disadvantaged Youth in Latin America: Evidence from a Randmized Trial. *NBER Working Paper* 2008, number 13931

Bartik, T, Instrumental Variable Estimates of the Labor Market Spillover Effects of Welfare Reform, Upjohn Institute Staff Working Paper 2002, number 02–078

Card D, Ibarraran P, Regalia F, Rosas D, Soares Y. The Labor Market Impacts of Youth Training in the Dominican Republic: Evidence from a Randomized Evaluation. *NBER Working Paper* 2007, number 12883.

Díaz J, Jaramillo M. An Evaluation of the Peruvian "Youth Labor Training Program" Projoven. Office of Evaluation and Oversight 2006. Inter American Development Bank.

Jalan J, Ravallion M. Estimating the benefit incidence of an antipoverty program by

propensity score matching. *Journal of Business & Economic Statistics* 2003; 21: 19–30

Klinger, B., Schündeln, M. Can Entrepreneurial Activity be Taught? Quasi-Experimental Evidence from Central America. Center for International Development Working Paper 2007, number 153. Harvard University. http://cid.harvard.edu/cidwp/153.htm

Rodríguez-Planas N, Jacob B. Evaluating Active Labor Market Programs in Romania. *Empirical Economic,*forthcoming.

Appendix: Commented bibliography on policy evaluation review articles

Introductory articles:

Ravallion M. The Mystery of the Vanishing Benefits. Ms Speedy Analyst´s Introduction to Evaluation. *World Bank Economic Review* 2001; 15: 115–140.

Vera-Hernández M. Evaluar intervenciones sanitarias sin experimentos. Gaceta Sanitaria

2003; 17, 238–248. (In Spanish, available at *http://scielo.isciii.es/pdf/gs/v17n3/revision.pdf*)

Review article on cluster level evaluation:

Ukoumunne O, Gulliford M, Chinn S, Sterne J, Burney P. Methods for evaluating area wide and organization-based interventions in health and health care: a systematic review. Health Technology Assessment 1999; 3, No. 5. *http://www.hta.ac.uk/fullmono/mon305.pdf*

Articles that review several evaluation techniques:

Blundell, R. Dearden, L. and B. Sianesi. Evaluating the impact of education on earnings in the

UK: Models, Methods and Results from the NCDS. *Royal Statistical Society: Serie A;* 168, 437–512

Blundell R, Costa Dias M. Evaluation methods for non-experimental data. *Fiscal Studies* 2000; 21, 4:427–468.

Heckman, J, Navarro-Lozano, S. Using Matching, Instrumental Variables, and Control Functions to Estimate Economic Choice Models. *Review of Economic and Statistics* 2004; 86, 30–57

Heckman J, Lalonde R, Smith J, The econometrics of active labor market programs. In: Ashenfelter O, Card D, editors. Handbook of labor economics. Vol 3. Amsterdam: North Holland, 1999; p.1865–2097

Heckman J, Robb R. Alternative methods for evaluating the impact of interventions. An overview. *Journal of Econometrics* 1985; 30:239–267.

Imbens, G.W., Wooldridge, J. Recent Developments in the Econometrics of Program Evaluation. *Journal of Economic Literature* 2009; 47: 5–86.

Ravallion, M. Evaluating Anti-Poverty Programs. In: Evenson R.E. and Schultz, T.P. Handbook of Development Economics. Vol. 4. Amsterdam: North Holland, 2008(NOT SURE OF YEAR)

Review articles on Randomization:

Burtless G. The case for randomized field trials in economic and policy research. *Journal of Economic Perspectives* 1995; 9: 63–84.

Duflo, E. Glennerster. R., and Kremer, M. Using Randomization in Development Economics Research: A Toolkit. *CEPR working paper* 2007, number 6059.

Review articles on Matching:

Caliendo M, Kopeinig S. Some Practical Guidance for the Implementation of Propensity Score Matching. *Journal of Economic Surveys* 2008; 22: 31–72

Imbens, G. Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review. *The Review of Economic and Statistics* 2004; 86: 4–29.

Review articles on Instrumental Variables:

Angrist, J. Treatment Effects in Theory and in Practice. *Economic Journal* 2004: C52–C83

Heckman, H. Instrumental Variables. A Study of Implicit Behavioral Assumptions Used in Making Program Evaluations, *Journal of Human Resources 1997;* 32: 441–462

Review articles on Regression Discontinuity:

Imbens G, Lemieux T. Regression Discontinuity Designs: A Guide to Practice. J*ournal of Econometrics* 2008; 142: 615–635

Lee D, Lemieux T. Regression Discontinuity Designs in Economics. *NBER Working paper* 2009; number 14723.

EMPLOYMENT POLICY PRIMER

The findings, interpretations, and conclusions expressed herein are those of the author(s), and do not necessarily reflect the views of the International Bank for Reconstruction and Development / The World Bank and its affiliated organizations, or those of the Executive Directors of The World Bank or the governments they represent.