View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Research Papers in Economics

IZA

DISCUSSION PAPER SERIES

IZA DP No. 6085

Building in an Evaluation Component for Active Labor Market Programs: A Practitioner's Guide

David Card
Pablo Ibarrarán
Juan Miguel Villa

October 2011

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

Building in an Evaluation Component for Active Labor Market Programs: A Practitioner's Guide

David Card
University of California, Berkeley, NBER and IZA

Pablo Ibarrarán
Inter-American Development Bank and IZA

Juan Miguel Villa
University of Manchester

Discussion Paper No. 6085
October 2011

IZA

P.O. Box 7240
53072 Bonn
Germany

Phone: +49-228-3894-0
Fax: +49-228-3894-180
E-mail: iza@iza.org

Any opinions expressed here are those of the author(s) and not those of IZA. Research published in this series may include views on policy, but the institute itself takes no institutional policy positions.

The Institute for the Study of Labor (IZA) in Bonn is a local and virtual international research center and a place of communication between science, politics and business. IZA is an independent nonprofit organization supported by Deutsche Post Foundation. The center is associated with the University of Bonn and offers a stimulating research environment through its international network, workshops and conferences, data service, project support, research visits and doctoral program. IZA engages in (i) original and internationally competitive research in all fields of labor economics, (ii) development of policy concepts, and (iii) dissemination of research results and concepts to the interested public.

IZA Discussion Papers often represent preliminary work and are circulated to encourage discussion. Citation of such a paper should account for its provisional character. A revised version may be available directly from the author.

IZA Discussion Paper No. 6085
October 2011

ABSTRACT

Building in an Evaluation Component for Active Labor Market Programs: A Practitioner's Guide*

The guide outlines the main evaluation challenges associated with ALMP's, and shows how to obtain rigorous impact estimates using two leading evaluation approaches. The most credible and straightforward evaluation method is a *randomized* design, in which a group of potential participants is randomly divided into a treatment and a control group. Random assignment ensures that the two groups would have had similar experiences in the post-program period in the absence of the program intervention. The observed post-program difference therefore yields a reliable estimate of the program impact. The second approach is a *difference in differences* design that compares the *change* in outcomes between the participant group and a selected comparison group from before to after the completion of the program. In general the outcomes of the comparison group may differ from the outcomes of the participant group, even in the absence of the program intervention. If the difference observed prior to the program would have persisted in the absence of the program, however, then the *change* in the outcome gap between the two groups yields a reliable estimate of the program impact. This guideline reviews the various steps in the design and implementation of ALMP's, and in subsequent analysis of the program data, that will ensure a rigorous and informative impact evaluation using either of these two techniques.

JEL Classification: C21, C93, H43, I38, J24

Keywords: active labor market programs, policy evaluation, randomized trials, difference in difference, average treatment effect on the treated, development effectiveness

Corresponding author:

Pablo Ibarrarán
Office of Strategic Planning and Development Effectiveness
Inter-American Development Bank
1300 New York Ave, Stop E0805
Washington, DC 20577
USA
E-mail: pibarraran@iadb.org

* We would like to thank Paul Winters for his comments and revisions, as well as Roberto Flores Lima and Graciana Rucci for their comments.

Table of Contents

1. Introduction 5

1.1 Purpose of the Guide 5

1.2 The Goal of an Evaluation 6

1.3 Two Basic Designs: Randomization and Difference in Differences 6

1.4 Avoiding a "Broken Design" 8

2. Background Factors that Affect an ALMP Evaluation 10

2.1 Mandatory or Voluntary Programs 10

2.2 Outcomes During the Program and After the Program 10

2.3 Selection of Program Participants 11

2.4. Non-Compliance: No-Shows, Dropouts, and Crossover from Control to Program Group 13

2.5. Site Variation in Program Enrollment and Impact 14

3. Design Issues for an ALMP Evaluation 16

3.1. Design Issues for Randomized Evaluations 16

3.1.1 Recruitment and Screening 16

3.1.2 Baseline Survey and Randomization 17

3.1.3 Sample Sizes 18

3.1.4 Procedures for Filling Open Slots 20

3.1.5 Timing and Content of Baseline and Follow-up Surveys 20

3.2. Design Issues for Difference in Differences Evaluations 22

3.2.1 Choosing a Comparison Group 22

3.2.2 Data Sources 25

4. Monitoring Implementation and Collecting Site-Level Information 27

4.1. What Really Happened? The Importance of Monitoring 27

4.2. Contextual and Site Effects 28

5. Ex Post Analysis 29

5.1. Check List for a Randomized Design 29

5.1.1 Check of Randomization 29

5.1.2 Check of Response Rates to Follow-Up Survey 30

5.1.3 Analysis of Compliance /Cross-Over 31

5.1.4 Estimation of Intention to Treat Impacts, with and without Adjustments 31

5.1.5 Adjustment of Intention to Treat Estimates for Non-Compliance/Cross-Over 32

5.2. Check List for a Difference in Differences Design 33

5.2.1 Comparison of Pre-Program Outcomes for Participant Group

and Comparison Group 33

5.2.2 Check of Response Rates to Follow-Up Survey 34

5.2.3 Analysis of Compliance /Cross-Over 34

5.2.4 Estimation of Intention to Treat Impacts, with and without Adjustments 34

5.2.5 Adjustment of Intention to Treat Estimates for Non-Compliance/Cross-Over 35

6. Example: Dominican Republic's Programa Juventud y Empleo – PJE 36

6.1. The Program and its First Evaluation 36

6.2. Applying the Check List for a Randomized Design to the 2009 JE Cohort 40

6.2.1 Check of Randomization 40

6.2.2 Check of Response Rates to Follow-Up Survey 41

6.2.3 Analysis of Compliance/Cross-Over 44

6.2.4 Estimation of Intention to Treat Impacts, with and without Adjustments 44

6.2.5 Adjustment of Intention to Treat Impacts for Non-Compliance/Cross-Over 46

7. Conclusions 48

References 49

Appendix A 52

1. Introduction

1.1 Purpose of the Guide

This guide is intended for practitioners who are designing an Active Labor Market Program (ALMP) that incorporates a rigorous evaluation component as part of the design. It offers guidance in the design of an evaluation, as well as practical advice on planning for a successful implementation. It does *not* provide a summary of the pros and cons of ALPM interventions. It is assumed that the decision to implement an ALMP has already been made, and that the user's goal is to provide credible and reliable evidence on the effectiveness of the approach.[1] It is also assumed that the primary objective of the program is to improve the *post-program* labor market outcomes of participants.

Active labor market programs include three broad classes of interventions –training programs, subsidized employment programs, and job search assistance programs– that are used in many countries to help labor market participants find and retain better jobs.[2] Despite their widespread adoption, the effectiveness of these programs remains controversial.[3] With the increasing emphasis in many governments and international agencies on evidence-based policy advice, it is important to understand which programs actually "work" −generating gains for participants that are large enough to justify their costs−and which are less successful.

Although useful insights into likely program success can be gained from a careful process analysis, a full understanding of the effects of a program requires a formal impact evaluation. A credible evaluation requires a rigorous methodological foundation. It also has to be carefully implemented according to the original design. Further, all stages of the evaluation must be fully documented to avoid any ambiguity in the interpretation of the results.[4]

---

[1] For an overview of active labor market policy issues, see Betcherman et al. (2000, 2004).

[2] The distinguishing feature of an *active* labor market program is that it requires participation by program recipients. Benefit programs like unemployment insurance are not generally considered as "active" labor market programs.

[3] This is particularly true for developing countries, in particular LAC. In a recent meta-evaluation of ALMP (Card, Kluve and Weber, 2011) only a handful of evaluations referred to projects in developing countries.

[4] Ibarrarán and Rosas (2009) review impact evaluations of job training programs in Latin America. Because of limitations in the design, implementation, and documentation of these evaluations, however, it is impossible to reach strong conclusions or make precise recommendations.

1.2 The Goal of an Evaluation

The primary goal of an ALMP evaluation is to provide objective, scientifically based evidence on the post-program impacts of the program. In most cases an evaluation will attempt to measure the extent to which participation raised the employment and/or earnings of participants at some point after completion of the program.[5] In general terms the "effect" or "impact" of a program represents the difference between the actual outcomes of program participants and their *counterfactual* outcomes if they had not participated. The fundamental problem for any evaluation is that these counterfactual outcomes can never be directly observed. An *evaluation design* is a systematic method for *estimating* these outcomes for the program participants, thereby allowing the analyst to measure the program's effects.

1.3 Two Basic Designs: Randomization and Difference in Differences

This guide considers two basic designs that are widely used in evaluating ALMP's. The first is a randomized experimental design (also known as a randomized controlled trial). In a randomized design, a group of individuals who satisfy the eligibility requirements for the program are randomly divided into two groups: the *treatment group*, who are assigned to receive the program, and the *control group*, who are assigned to *not* receive the program. Because assignment is random, the treatment and control groups would be expected to have similar experiences in the post-program period in the absence of the program intervention. Randomization therefore provides a simple method for constructing a counterfactual for the treatment group, using the observed outcomes of the control group.[6] The estimate of the program effect from a randomized evaluation is simply the difference in post-program outcomes between the treatment group and the control group.

It is important to emphasize the role of *random* assignment in this design. Other ways of selecting the beneficiary group will typically lead to differences between the post-program outcomes of those who participated in the program and those that did not, even in the absence of the program effect. When this occurs, the difference between post-program outcomes of the participants and the non-participants includes *both* the effect of the program *and* any differences

[5] The question of what outcome(s) should be measured in a particular evaluation context is discussed in Section 3 of this guideline.

[6] See Duflo, Glennerster and Kremer (2006) for a discussion and toolkit on the use of randomized methods in development economics.

that would occur even if the program had no effect. The latter are known as "selection biases", since they depend on the precise way that the participants are selected. Selection biases are averted in a randomized design because participants and non-participants are selected at random, and differ (on average) only because of their participation status. From a practical perspective, randomization is also a fair method to assign benefits among equally deserving and eligible individuals. When, as is often the case, demand for benefits exceeds the program's capacity to deliver them, random assignment is a transparent way to allocate participants to the available slots.

Although a randomized design is widely considered the gold standard for program evaluation, in many situations randomization is infeasible or unacceptable. In this case the analyst must adopt a non-experimental design. The second basic design considered in this guide is the so-called *difference in differences* (DD) design. DD designs are widely used in ALMP evaluations, and are generally accepted as one of the best alternatives to a randomized design.[7]

The DD design uses information from a non-randomly selected group of program non-participants –known as the *comparison group*– to construct the counterfactual outcomes of the participant group. In general, the observed difference in the post-program outcomes of the participant group and the comparison group will include both the true treatment effect, and the selection bias component due to underlying differences between the two groups. The idea of a DD design is to compare the difference in outcomes between the participant group and the comparison groups in some period *after* the participants have completed the program with the difference that existed *before* the program. Provided that the selection bias component is constant over time, the *change* in the difference between the participant group and the comparison group from before to after (i.e., the difference in differences) is an "unbiased" estimate of the estimated effect of the program.[8]

The *assumption* in a DD design is that any pre-program differences between the treatment and comparison groups would have remained fixed in the absence of the program.

[7] A third design, the so-called "regression discontinuity" (RD) design, can be used in special cases where selection of the program participants involves a rule that depends on an observable factor (like participant earnings in the previous period), with a sharp break (or "discontinuity") in the rules at some threshold. RD designs can yield highly credible evaluations but have not been used in the ALMP context because of the general absence of sharp thresholds in the eligibility rules. See Lee and Lemieux (2009) and DiNardo and Lee (2010) for a discussion of RD designs.

[8] An "unbiased" estimate is one that on average yields the true value of the quantity being estimated.

Under this assumption, the post-program outcomes of the comparison group, adjusted for the pre-program differential, form a valid counterfactual for the post-program outcomes of the treatment group. Clearly, the integrity of a DD design hinges on whether it is really true that the pre-program difference between the participants and the comparison group would have persisted in the absence of the program. If so, the design is valid. If not, program impacts estimated from a DD design contain a selection bias component, and will be biased.

An extension of the difference in difference design involves the use of *matching techniques* to narrow the subset of the comparison group that is used to construct counterfactuals for the participant group. In a *matched difference in differences* design the comparison group is selected from a broader set of potential comparisons using a procedure that ensures that the pre-program characteristics of the selected comparison group closely match the pre-program characteristics of the treatment group.[9]

Note that one could use a DD approach to estimate program effects in a randomized design. Under the assumption of random assignment, however, the pre-program difference between the participant group and the comparison group will be close to zero, and there is little or no gain in using the DD, rather than simply the post-program difference, as an estimate of the impact of the program. Indeed, the existence of a pre-program difference in the outcomes of the treatment and control groups provides evidence of a potential failure of the random assignment process (see section 5, below).

1.4 Avoiding a "Broken Design"

Randomized and DD designs are straightforward evaluation methods that rely on simple comparisons to measure the impacts of a program. The simplicity and transparency of these designs enhance their credibility among policy makers. Since non-technical readers can understand the basis for the estimated impacts arising from these designs, they can have greater confidence in their validity and place greater weight on the findings.

Nevertheless, there are many details in the implementation of both designs that can go

[9] The extension to *matched difference in differences* is conceptually simple, and requires the same information as is required for a standard difference in differences (DD) design. Details on matching and matched DD analysis methods are provided in a separate guideline on Propensity Score Matching. (Heinrich, Maffioli and Vázquez, 2010)

wrong. When that happens, the evaluation design is compromised and the analyst has to invoke additional assumptions and/or use statistical techniques to try to fix the broken design. Unfortunately, the findings from a broken design often depend on the particular assumptions and/or techniques used in the analysis. The resulting ambiguity can lead to disagreements among analysts and confusion among policy makers. Many times the findings from a broken design are simply ignored.

Although many things can go wrong in an evaluation, experience suggests that careful attention to planning can reduce the chances for errors and omissions that lead to a broken design. The purpose of this guideline is to help evaluation designers plan ahead and be prepared for some of the most common problems that arise in the evaluation of active labor market programs. The next section begins with a background discussion of the common features of active labor market programs that need to be taken into account in any evaluation project. Section 3 then turns to a discussion of the basic elements in the two types of designs, including the design of data collection procedures. Section 4 discusses the critical importance of a careful monitoring system to oversee implementation (particularly of randomized evaluations), and the value of site-level contextual information. Section 5 presents a brief discussion of issues pertaining to the analysis and interpretation of findings from an evaluation. Section 6 presents an extended example of the application of the ideas to a “real” evaluation of an ALMP in Dominican Republic.

2. Background Factors that Affect an ALMP Evaluation

The nature of active labor market programs, and the types of people who choose to enroll in these programs, affect many aspects of the design of a valid evaluation. As background for a more detailed discussion of design issues in Section 3, this section presents a brief overview of some of the most common features of ALMP programs and participants. Understanding which of these features are likely to be present in a particular setting will help in formulating the appropriate design.

2.1 Mandatory or Voluntary Programs

A fundamental characteristic of an ALMP is whether the program is *mandatory* or *voluntary*. Mandatory programs are built into the unemployment compensation systems of many developed countries. In these settings, beneficiaries are *required* to participate in an active labor market program after receiving benefits for a certain amount of time.[10] Voluntary programs, in contrast, recruit participants from a wider pool of applicants who can decide whether or not to participate.

This guide focuses on the evaluation of voluntary active labor market programs.[11] Specifically, it is assumed that the ALMP in question is offered at a site or group of sites to people who present themselves as potential participants and satisfy the appropriate eligibility screening for the program (e.g., meet the age criteria for a youth program, or the family income criteria for an income-targeted program). The fact that the participants have willingly applied to participate in the program at a certain time has to be considered carefully in designing an evaluation, particularly if a non-experimental (DD) design is being used.

2.2 Outcomes During the Program and After the Program

Active labor market programs are distinguished from *passive* programs (like conventional unemployment benefits) by the requirement that participants attend classes, work in a subsidized job, or participate in workshops or similar activities. During the period of actual program participation there is a *mechanical* relationship between program status and outcomes like employment and income. In a classroom training program, for example, program participants generally have *lower* employment rates and earnings than they would have had in the absence of

[10] For example, see Sianesi (2004) for a discussion of mandatory ALMP's in Sweden.
[11] So far, only voluntary ALMP have been implemented in Latin America and the Caribbean.

the program.[12] By comparison, subsidized employment programs mechanically *increase* participants' employment and earnings while they are in the program. It is important to distinguish these "within-program" effects from any impact of the program on post-program outcomes. A key issue in designing an evaluation is making sure that the analysis can distinguish between the mechanical within-program effects and any lasting impact on post-program outcomes. Normally, this means that information on participants' outcomes must be available for a period well beyond the expected completion date for the program.

2.3 Selection of Program Participants

A key concern in the design of evaluations for voluntary ALMP's is the selection process that leads participants to volunteer for the program. In many cases, those who present themselves for a program have recently experienced a job interruption and are interested in participating in the program as a way to improve their chances for re-employment. As was first noted by Ashenfelter (1978), such people normally experience a rebound in employment and earnings, even in the absence of a program intervention. This rebound (or "mean reversion") effect makes a DD evaluation particularly challenging.

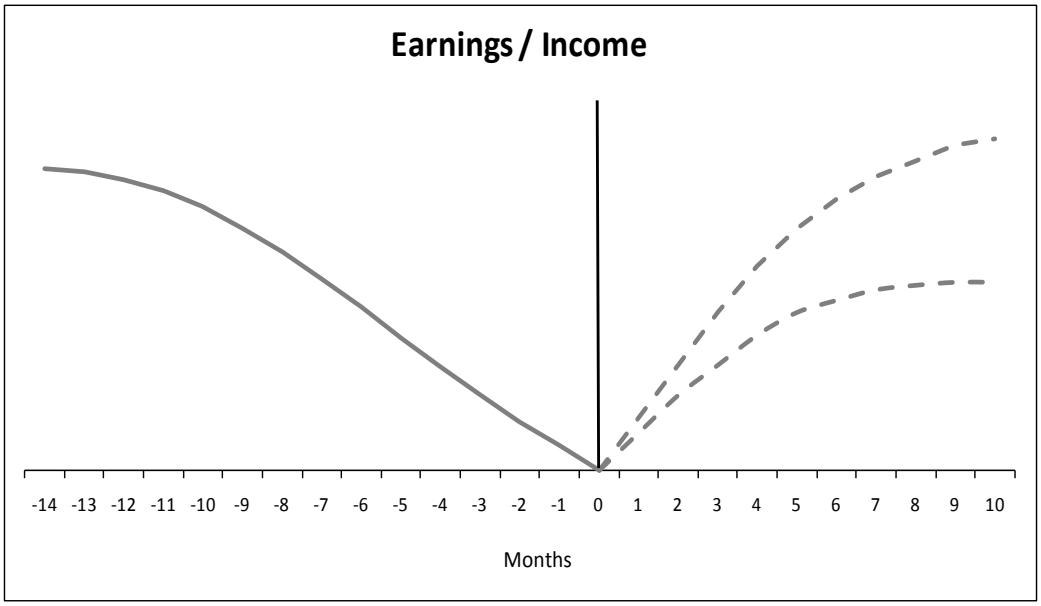
To understand the nature of the problem, consider a group of people who are all observed to be out of work at some reference date. Figure 1 shows the time-profile of the employment rate for such a group. Since everyone is currently unemployed, the employment rate in the reference month, denoted as "month 0" in the graph, is 0. Some of the group lost their job very recently (months -1, -2, etc.) whereas others have been searching for much longer. Consequently, the average employment rate of the group is declining steadily from month -12 (one year before the reference date) to 0% in the reference month. The figure also shows the hypothetical employment rates of the group in the months *after* the reference date (months 1, 2,...). Since some fraction of unemployed people typically find jobs and move into employment, the employment rate in these months is gradually rising, even in the absence of any program intervention.

Participants in many active labor market programs have been found to exhibit a pattern of declining employment and earnings in the period prior to entering the program, similar to pattern

[12] Such an effect is known in the ALMP literature as an incapacitation effect, and represents the opportunity cost of program participation, similar to the lower employment rate observed among people who are enrolled in school.

during months **-5** to **0** in Figure 1. This phenomenon, known as "Ashenfelter's dip", is most obvious in programs targeted to recent job losers, but is also evident in programs for other disadvantaged groups. Program selection based on current unemployment will often lead to a participant group whose labor market outcomes are temporarily depressed below their long-run or "permanent" level. Even if the program has no effect, we would expect to see a gain in the outcomes of the treatment group in the post-program period as these temporary factors are resolved.

Figure 1. The "Ashenfelter dip"



A valid evaluation requires some way of measuring the counterfactual outcomes for the program group, accounting for any expected rebound effect. A randomized design solves this problem by randomly assigning some potential enrollees to the treatment group and some to the control group. For a DD evaluation, however, there is no guarantee that the comparison group would be expected to experience a similar rebound. One solution is to find a comparison group that has experienced a parallel "Ashenfelter dip" and use their subsequent outcomes as the basis for the counterfactual. Another is to use information on pre-program earnings or employment for

several periods prior to program entry in forming a model-based version of a DD specification (see e.g., Card and Sullivan, 1988). In either case, information is needed on the outcomes of the treatment group *and* the comparison group for *several periods* prior to the program (typically for at least one year, and ideally longer).

Even when extensive data on pre-program outcomes are available, a second set of concerns arise when potential participants in a voluntary ALMP can assess their likely gains from participating in a program, and use this information in deciding whether to participate. In such cases it may be difficult to ensure that the outcomes of the comparison group adequately reflect the counterfactual outcomes of the program participants. A related problem arises when program operators select applicants who are most likely to have positive post-program outcomes, using information that is not available to the evaluation analyst.[13] Unless a similar screening process can be applied to the comparison group, a DD evaluation may overstate the impact of the program.

2.4. Non-Compliance: No-Shows, Dropouts, and Crossover from Control to Program Group

In many voluntary ALMPs a substantial fraction of people who are assigned to the program will either fail to register for the program (so-called *no-shows*) or will drop out prior to completion of the program (*dropouts*). Indeed, it is rare to achieve program completion rates over 80 percent and rates as low as 50 percent are common. Failure to anticipate the problems caused by no-shows and dropouts is one of the leading causes of a broken design in ALMP evaluations. Closely related to the problem of dropouts and no-shows is the possibility that people assigned to the control group manage to enroll in a very similar program (or in the same program at another site). While non-compliance (or "cross-over") by members of either the program group or the control group does not invalidate an evaluation design, it does complicate the interpretation of the results, and means that the evaluation has to collect data on the actual program participation rates of the program group and the control group.

The validity of a randomized design relies critically on the equivalence between the observed outcomes of the control group and the counterfactual outcomes of the treatment group. In most cases this equivalence is compromised when members from one group or the other are

[13] This process is known informally as *cream-skimming*, and may be particularly relevant in cases where program operators receive financial incentives based on the post-program outcomes of their clients.

dropped or lost. For this reason, the analysis of a randomized design should be based on a comparison of the initially-assigned treatment and control groups, using data on *everyone* who was initially assigned to these groups. In the experimental evaluation literature this is known as an *intention to treat* analysis.[14] As discussed in Section 5, below, results from an intention to treat analysis can usually be adjusted after the fact for the non-compliance or cross-over behavior of the groups.

A critical but easily overlooked principle is that *no-shows and dropouts have to be included* as part of the treatment group. A valid design requires post-program outcome data for *everyone* in the treatment group, regardless of whether they actually completed (or even started) the program.

Likewise, *everyone* assigned to the control group has to be followed and included in the analysis, including people who participate in the program at some later date (or participate in other similar programs).[15] The need for including everyone, including dropouts and no-shows, means that the information required to track people for inclusion in post-program surveys *must* be collected at the time of random assignment, since after that point some people inevitably "disappear."

The same general principles apply to DD designs. Outcomes for people who are part of the designated treatment group should be included regardless of whether they completed the program. Likewise, outcomes for everyone in the designated comparison group should be included in the construction of the DD counterfactual. As in a randomized design, the outcomes of the two groups can be adjusted after the fact for cross-over behavior.

2.5. Site Variation in Program Enrollment and Impact

Most ALMPs are offered at multiple sites, with site-to-site variation in the characteristics of the potential program enrollees and in the quality of program services.[16] This variation has a number of implications that need to be addressed in planning an ALMP evaluation. Most importantly,

[14] People assigned to the treatment group who fail to participate, or those assigned to the control group who actually participate, are known in this literature as "cross-overs".

[15] As discussed in Section 3, special attention should be paid to procedures for filling the slots left open by the no-shows and dropouts, to ensure that this process does not compromise the validity of the control group.

[16] For classroom programs, the training and ability of instructors may vary from site to site. The quality of on-the-job training and subsidized employment programs is determined in part by the characteristics of participating employers, which also vary from site to site.

when a program is implemented in multiple sites the design must include adequate resources for monitoring the implementation *at all sites*. Monitoring is particularly important in a randomized design because the validity of the design depends on proper implementation at each site. In particular, a randomized evaluation has to maintain balance across sites between the treatment and control groups (i.e., equal fractions of the two groups at each site). As discussed in section 3, the randomization procedure should be designed to ensure balanced assignment. Moreover, checks should be built in to the implementation process to ensure that sites actually enroll the full complement of people who are assigned to receive treatment at the site, and prevent people assigned to the control group from receiving services. It is also important to keep a record of the site information for both the treatment and control groups to allow an ex post analysis of the assignment process.

Site information can also be very useful in defining an appropriate comparison group for a DD design. As discussed in Section 3.2, comparison samples are often drawn from larger populations, using a matching procedure to construct a sample that has the same distribution across sites as the treatment group.

Finally, site variation can be an important in understanding differences in measured program effects. For example, if an evaluation is comparing two types of programs, and one type of program is mainly offered at one set of sites, while the other program is mainly offered at other sites, then site information is needed to separate the effects of site-specific effects from any differential impacts of the two programs.

3. Design Issues for an ALMP Evaluation

This section turns to an overview of the main steps in planning an evaluation design. For simplicity the two types of designs are treated separately, although there is obviously some overlap in the issues affecting the two designs.

3.1. Design Issues for Randomized Evaluations

Planning for a randomized evaluation has to address five basic questions:

i. How and where will potential members of the sample be recruited and screened?

ii. How and when will the baseline survey and randomization be conducted?

iii. What are the sample sizes for the treatment and control groups?

iv. What are the procedures for filling slots left open by no-shows and dropouts?

v. When will the baseline survey and follow-up survey(s) be conducted, and what questions will be included in these surveys?

3.1.1 Recruitment and Screening

Most ALMP evaluations involve programs that are offered by existing public or private-sector service providers at established sites. Typically the evaluation will recruit participants from the regular flow of clients at each site. In some cases, all the clients who appear at a specific set of sites in some time period will be incorporated into the evaluation. In other cases only a fraction of the clients who appear each day or week will be included. The size of the evaluation sample, relative to the flow of new clients at the program sites included in the evaluation, will determine how long the recruitment phase of the evaluation has to continue.

Because only a fraction of all the people recruited to a randomized design are assigned to actually receive the program, intake for an evaluation may disrupt the normal flow of clients into an ongoing program.[17] This is not a particular concern in a setting where there are many more applicants than available program slots: in these cases randomization serves as a convenient and objective rationing device. In settings where the regular flow of recruits is needed to fill the available program slots, however, program operators may object to having some of their

[17] For example, if 100 new clients present at the program sites each month, and there are 80 open program slots each month, then at most 40 people per month can be recruited into the evaluation: 20 will be assigned to the program (along with the other 60 new clients who are not part of the evaluation) and 20 to the control group.

potential clients allocated to the control group, and may try to over-ride the assignment process. *It is extremely important to know in advance whether this is likely to occur*. If so, planning for the evaluation may have to include a budget for extra recruitment effort to increase the flow of new clients, and extra resources to closely monitor compliance with recruiting protocols.

ALMP participants are often required to satisfy certain targeting criteria: for example, a program may be limited to unemployed men and women between the ages of 16 and 54 whose family incomes are below some threshold. Normally the same eligibility screening procedures and rules should be used to select participants for the evaluation. Whenever possible the eligibility screening data should be retained and made a part of the evaluation data set.

3.1.2 Baseline Survey and Randomization

Once participants for the evaluation have been screened, they have to complete the baseline survey and be randomized into the treatment and control groups.[18] Randomization *after* the baseline survey ensures that respondents are unaware of their program group status (i.e., treatment or control group) when they complete the survey, thus eliminating any concern that assignment status influences the answers to the baseline survey. The content of the baseline survey is discussed below. Normally, the survey should collect basic information on individuals in the evaluation (age, gender, education, family circumstances) as well as information on the same outcomes as will be considered in the evaluation (e.g., earnings and employment status in the period just prior to the baseline date). Such information will allow a check that randomization was correctly implemented, and can also be used to help interpret the outcome results.

The randomization process itself should be designed to meet two key objectives. First, the process must be carefully controlled and documented: normally the assignment process should be conducted through a centralized office using a random number generator (or some other verifiable randomizing device). Second, the process should be designed so *everyone* who has completed the baseline survey has the same probability of assignment to the treatment group (normally 50%).[19] Although more complex *stratified* designs are possible (e.g., designs that

[18]Prior to being randomly assigned it may be necessary to obtain the participants' informed consent that they agree to be a part of the experiment.

[19] If the experiment is designed to evaluate two alternative treatments that everyone should have equal probabilities of assignment to each of the treatment groups (e.g., 33% to treatment 1, 33% to treatment 2, and 33% to the

allow different probabilities of assignment to the treatment group at different program sites), these designs are more difficult to implement and analyze. In most cases the potential costs of a stratified design far outweighed the benefits.

After randomization individuals who are assigned to the treatment group enter the program and receive the normal services offered to participants. In many cases only a fraction of those who are assigned to the treatment group will actually complete the program, though as emphasized above *all those initially assigned to treatment* should be included as part of the group in all future surveys and analysis. Ideally, individuals who are assigned to the control group are actively prevented from receiving program services for a period of time (e.g. one or two year) after random assignment. Typically this means that names and addresses of control group members are entered into an "embargo list" of people who cannot receive services. A complete embargo may be difficult to enforce if members of the control group can apply at other sites outside the purview of the evaluation team. For this reason, the design should plan for the possibility of "cross-over", though as a general principle *all those initially assigned to the control group* should be included in the group in all future surveys and analysis.

3.1.3 Sample Sizes

Guidelines for the necessary sample sizes for a randomized ALMP evaluation are based on a standard power calculation. The main ingredient for this calculation is an estimate of the plausible *effect size* of the program (e.g., the effect of the program on the outcome of interest, expressed as a fraction of the standard deviation of this outcome). Given this value, and standard choices for the statistical significance level (e.g., 5%) and the adequacy of the power of the design (e.g., 0.80) it is straightforward to calculate the appropriate sample sizes for the treatment and control groups of a randomized design with equal-sized groups. The Table 1 shows the sample size required to measure a range of impacts.[20] Each row shows the employment rate of the control group, and each column represents the difference between treatment and controls. For example, if the employment rate is 50% in the comparison group, to detect a significant impact of 2.5 percentage points in employment, the required sample size is of 6,354 participants and the

control group).

[20] Under the standard assumptions (power=0.8, significance =0.5, equal-sized groups), using the *sampsi* command in STATA.

same number of non-participants.

Table 1. Sample Size Required to Detect Significant Impacts

		Impact of the Program					
		2.5%	5.0%	7.5%	10.0%	12.5%	15.0%
Employment Rate of Control Group	30%	5475	1417	650	376	247	176
	35%	5883	1511	688	396	259	183
	40%	6166	1574	713	408	265	186
	45%	6323	1605	723	412	266	186
	50%	6354	1605	719	408	262	183
	55%	6260	1574	702	396	254	176
	60%	6040	1511	671	376	240	165
	65%	5695	1417	625	349	221	151

In thinking about the effect size of interest for an active labor market program it is useful to place these programs in context. A very large body of research has shown that in most countries around the world each additional year of formal schooling is associated with a gain in earnings of about 10 percent. Arguably, a typical ALMP involves a smaller investment than a typical year of formal schooling, so an effect size of less than 10% is reasonable, and for less intensive programs, effect sizes of no more than 5% may be plausible.

The sample sizes indicated by a simple power calculation are based on a best case scenario, and ignore the problems caused by non-response, non-completion by the program group (i.e., no-shows and dropouts), and the receipt of program services by some members of the control group. If for example the expected rate of non-completion by the program group is 25%, and estimated non-response to the follow-up survey is 20%, then the sample sizes should be increased by 25-30% from the best case calculations.

More generally, larger sample sizes make it possible to conduct informative subgroup analyses within the overall experimental population. There is often wide interest among policy makers in the comparative effects of a program across sites or geographic areas, or between male and female or younger and older participants. With larger sample sizes there comparisons are more likely to be statistically informative.

3.1.4 Procedures for Filling Open Slots

No-shows and dropouts pose an operational challenge to many ALMP evaluations because service providers want to keep their programs full. In an ongoing program that has many more potential recruits than available slots, program operators often maintain a waiting list of eligible recruits. Slots left open by no-shows and dropouts are then assigned to people on the waiting list. If possible, similar procedures should be adopted in an evaluation. Thus, if people assigned to the treatment group drop out, and the program operators want to keep their programs full, the open slots should be filled from a waiting list. In cases where enrollment into the evaluation sample occurs over several months, the waiting list can be comprised of newly assigned members of the treatment group, assuming that intake for random assignment can "keep ahead" of the availability of slots for program participants. In cases where a list of newly assigned members of the program group cannot be maintained, it is preferable to fill open slots using people who are outside the evaluation (i.e., clients who are taken into the program by the regular process) rather than use members of the embargoed control group.[21]

3.1.5 Timing and Content of Baseline and Follow-Up Surveys

The baseline survey in a randomized evaluation should be conducted just prior to random assignment. The timing for the follow-up survey (or surveys) is less clear cut. Many ALMP evaluations use a one-year follow-up survey, in part because the terms of the evaluation contract often require a final report within two or three years. On the other hand, the existing ALMP literature suggests that the impact of more intensive programs, such as classroom training and on-the-job training programs, only tends to emerge after two or three years after entry into active labor market programs than after only a year (Card, Kluve, and Weber, 2009). A similar conclusion has been reached in a careful analysis of the long term effects of alternative programs for welfare recipients in the U.S. (Hotz, Imbens, and Klerman, 2006). Based on these studies, and consideration of the interruption effects of many ALMP's, a post-program horizon of at least two years is probably desirable for longer-duration ALMP's. Otherwise, the evaluation may fail to fully capture the long run benefits of the program. If the budget permits, a compromise is to

[21] In rare cases where the implementation plan failed to take account of recruiting limitations, the open slots may have to be filled by members of the originally-designated control group. In this case, re-assignment should be based on a randomization procedure to ensure the validity of the remaining control group.

conduct a first follow-up survey roughly one year after random assignment that can be used as the basis for a timely interim report, and a second follow-up survey 24-26 months after random assignment that can form the basis for a final report.

The content of the baseline and follow-up surveys should reflect the stated goals of the active labor market program being evaluated. Since most ALMP's are motivated by the goal of improving participants' earnings, the baseline and follow-up surveys should normally collect information on total earnings per week or month at the time of the survey, along with parallel information on hours of work, the number of jobs held, and the characteristics of each job (e.g., earnings, hours, industry, occupation, formality).[22] In cases where the focus of the program is on job skills, information on job characteristics is especially important, since it provides a basis for determining if participants are successful in obtaining jobs with higher skill requirements. Whether a job is in the formal sector is a particularly important characteristic of job outcomes in many Latin American and Caribbean (LAC) labor markets. Careful attention should be paid to developing appropriate ways to measure formality in the specific context.

Except in very special cases, the program effect at a single point in time is only a partial measure of the effect of an ALMP. A typical program has an interruption effect that may cause a negative impact on earnings while the participants are in the program. Once the program is completed, it may take several months for participants to catch up to members of the control group. These dynamic considerations suggest that a more complete analysis should examine the program's impact over the entire post-random assignment period. Normally, such an analysis can be based on retrospective questions contained in the follow-up survey (or surveys) asking about employment and earnings in each month since random assignment. The design of the retrospective calendar should take account of the fact that participants often will have entered *and* left the program at different calendar dates, and will be interviewed with different amounts of time since leaving the program.

3.2. Design Issues for Difference in Differences Evaluations

Planning for a DD evaluation has to address two basic questions:

[22]In standard economic models the benefit of program participation depends on its effect on the post-program wage per unit of time. In search theoretic models the benefit also depends on whether the program affects the arrival rate of job offers. A measure like earnings per month (or quarter) effectively summarizes both dimensions.

i. How will the comparison group(s) be defined?

ii. What data sources will be used to collect information on the two groups. If the answer includes specialized surveys, when will these be administered and what questions will be included?

3.2.1 Choosing a Comparison Group

The fundamental issue in a DD design is the choice of the comparison group. Most ALMP evaluations begin with a *potential* comparison group, then impose additional restrictions to select a final comparison group that better matches the characteristics of the actual program participants. For example, a potential comparison group may include people interviewed in an existing labor force survey (LFS) in a month close to the date that participants were recruited for the program. This potential group may then be further restricted by imposing age or labor force status restrictions.

There are four common methods of defining a *potential* comparison group. The first is based on location: in this case the potential comparison group consists of people who are similar to the program participants but are observed in other locations where the program is (or was) unavailable. The second is based on time: here, the potential comparison group is made up of similar people observed at a different time either before or after the program was in place. A third possibility is based on program eligibility rules: in this case the potential comparison group is composed of people from the same geographic area and time period who are as similar as possible to the program participants, but nevertheless ineligible to participate.[23] The fourth possibility is a potential comparison group made up of people from the same area and time period who could have participated but were either unaware of the program or chose not to participate.

Although the rankings of these alternatives will vary from context to context, a general observation is that comparisons based on location are often the most compelling choice. One reason is that the labor market outcomes of similar groups of people are often found to be very similar in different locations, once allowance is made for fixed location differentials. In a variety

[23]This method is appropriate for programs with sharp eligibility rules, like an age limit or family income threshold. Such programs can sometimes be evaluated using a regression discontinuity (RD) approach: see Lee and Lemieux (2009) and Imbens and Lemieux (2008) for recent discussions of RD evaluation methods.

of different countries, for example, the average earnings of young men who live in different counties or cities tend to be quite similar, apart from geographic differences that are relatively stable over time. A second reason is that rates of geographic mobility among the disadvantaged workers who are the typical clients for an ALMP tend to be relatively low. Few people will move to another city (or region) just to take advantage of an ALMP that is offered there. This means that the behavior of a comparison group of non-participants from areas where a program was not offered is likely to be free of selection biases associated with the decision *not* to participate in the program. Having such an "unselected" comparison group greatly simplifies a DD design.

Relative to a geographically-based comparison group, a comparison group comprised of people observed in a different time period (when the program was not offered) has the *disadvantage* that the outcomes of the participant group and the comparison group are observed at different time periods. *If* there is little temporal variation in the outcomes of interest this may not be a serious limitation. In many settings, however, the labor market outcomes of ALMP participants and other disadvantaged workers are heavily affected by seasonal and cyclical effects. Unless outcomes for the comparison group can be drawn from periods in the same season of the year as those of the participant group, and from years with the same general state of the overall economy, a temporally-based comparison group is likely to yield unreliable program impacts.

In some situations eligibility for an ALMP is limited to people with certain individual characteristics (like age or earnings). In these cases it may be possible to design a comparison group comprised of people who were "almost eligible". For example, if a program is restricted to people under the age of 25, people between the ages of 26 and 30 would be an obvious comparison group. To be valid, a comparison group of "ineligible" people has to be based on *fixed* and *readily verifiable* characteristics. Eligibility rules that based on a maximum education or income, for example, could be easily manipulated by under-reporting education or income, leading to a situation where some of the ineligible group actually participated in the program. In practice, eligibility for ALMP's is rarely based on sharp eligibility rules, and in cases where there is some potentially sharp criterion, the rules are only loosely enforced, or applicants are aware of the rules and circumvent them by misreporting their qualifications. This means it may be hard or even impossible to build a comparison groups based on presumed eligibility

characteristics.

Arguably, the *weakest* basis for a difference and differences evaluation design is a comparison group made up of people who could have participated, but for reasons unknown to the analyst did not. The problem with such a design is that *both* the participant group and the comparison group are self-selected. In one case individuals have actively chosen to join the program; in the other they have chosen to *not* join the program. Such two-sided selection can lead to very complex patterns of expected differences between the participant group and the comparison group, even in the absence of any program effect, making it very unlikely that the assumptions of a DD design are satisfied.[24]

Once a potential comparison group is selected, a second issue is how to narrow down the group to better reflect the counterfactual outcomes of the participants. For example, in settings where participants appear to exhibit an "Ashenfelter dip" in pre-program outcomes, it may be possible to select a comparison group with a similar time profile of earnings and/or employment in the pre-program period. This can be accomplished using a statistical matching procedure to select members from the potential comparison group whose labor market histories (and other characteristics) match those of the program participants.[25]

Refinement of the comparison group may not be necessary in all applications. For example, participants in labor market "insertion" programs for youths typically have had no significant earnings in any period prior to the program, and a comparison group of similar youths drawn from other regions may also show no significant earnings (or a relatively constant but low level of earnings). In this case a standard DD approach with a simple regression adjustment would be appropriate.

[24] Two sided selection problems arise in many other contexts (e.g., the decision of whether to work in a formal sector or informal sector job; whether to emigrate).

[25] There is an extensive methodological literature on the application of matching methods to the evaluation of ALMP's: see Heckman, Lalonde and Smith (1999); Lechner (2002); and Abadie (2005).

3.2.2 Data Sources

Closely related to the choice of the comparison group is the issue of data sources for the participant group and the comparison group. In fact, in many evaluations the choice of the comparison group is dictated by data availability. As discussed in Section 2, the data requirements for a credible DD evaluation include at least one year (and preferably more) of pre-program data for the participants, and comparable data for the comparison group.

There are three basic approaches to compiling the necessary data for a DD design. The first, widely used in studies of ALMP's in developed economies, relies on administrative data from the payroll tax system.[26] In countries with a small informal sector this is a simple, cost effective way to obtain reliable longitudinal data. In most developing countries, however, only a fraction of jobs are in the formal sector. Moreover, participants in voluntary ALMP's are often young, less-skilled workers who are relatively unlikely to find a job in the covered sector. For these reasons, administrative data have not been widely used for evaluating ALMP's in developing countries.

Instead, DD evaluations in developing countries usually have to rely on specialized baseline and follow-up surveys (as are used in randomized designs), or on a combination of specialized and existing surveys (for example, using an existing survey as the baseline survey for the comparison group, then administering a special baseline survey to members of the participant group, and conducting a follow-up survey for both groups).

If specialized surveys are used for both groups, the issues of the timing and content of these surveys are broadly similar to the issues for a randomized design. For a DD analysis the collection of similar outcomes data in the baseline and follow-up surveys is especially critical, since the *change* in outcomes is used to form the impact estimates. If possible the two surveys should include *identical* questions on labor market outcomes (i.e., employment and earnings) in the period prior to each survey. If there is any concern over a potential "Ashenfelter dip" in pre-program outcomes of the participants, the baseline survey must also collect information on the history of outcomes prior to the baseline (e.g., by filling in a retrospective calendar for the preceding 12 months or 4 quarters).

In cases where the baseline survey for the comparison group will be based on an existing

[26]This was the approach used in the early work in the area by Ashenfelter (1978) and Ashenfelter and Card (1985), and is still widely used in evaluation in Europe (see Card, Kluve and Weber, 2009).

survey, the available retrospective data may be limited. The evaluation design team then has to make a decision: can a credible DD analysis be conducted using the available data, or not? If retrospective data are unavailable (or are not detailed enough to allow comparisons of the trend in pre-program outcomes in the year prior to the baseline date) then the existing survey source *cannot* be used for a credible evaluation, and a specialized baseline for the comparison group will have to be developed.

A second consideration is that in collecting baseline data for a potential comparison group, it may be necessary to survey a relatively large number of individuals to obtain a useable sample whose pre-baseline data match those of the participant group. This is the advantage of using an existing survey (e.g., an existing Labor Force Survey) as the baseline for the comparison group: the comparison group can be drawn from respondents on the survey who match the eligibility rules for the program.

In cases where an existing survey will be used as the baseline for the comparison group, it is important to ensure that the questions on the baseline survey for the program group closely match the questions in the existing survey. It is well known that even minor changes in question wording can lead to quantitatively important differences in measured outcomes. Thus, identical wording should be used whenever possible.

4. Monitoring Implementation and Collecting Site-Level Information

In both randomized and DD evaluations it is very important to set up a comprehensive monitoring system to ensure the integrity of program implementation and (in the case of a randomized design) of the evaluation design. It is also important to collect some basic site-level information. This section briefly reviews the importance of implementation monitoring and the value of site-level information for a credible impact evaluation.

4.1. What Really Happened? The Importance of Monitoring

In order to learn from an impact evaluation, analysts and readers need to be very clear about the nature of the actual intervention they are studying. This can only happen with a sound monitoring system that records the "what, how, and when" of the intervention. In the case of a randomized evaluation, the monitoring system must also ensure the integrity of the random assignment process (for example, by ensuring that assignment protocols are rigorously followed at each site).

Assuming that the program was successfully implemented, the interpretation of the results of the evaluation requires detailed information on the nature of the program itself. Among the questions that are important to address with the implementation monitoring system are the following:

i) What was the typical content of the program? Was this content designed to help the participants immediately, or over the longer run?

ii) How long did the program typically last? At what point after initial program entry did most participants begin to search for a regular job?

iii) Were program participants in a subsidized employment program able to move directly from a subsidized job to a regular job at the same employer?

iv) What was the (approximate) cost of the program services delivered to participants?

Information on the nature and duration of the program is particularly important for understanding the likely dynamics pattern of program impacts. A longer-duration program focused on basic skills is likely to have a bigger interruption effect, and to yield positive impacts only after a relatively long time period. A short-duration program focused on job readiness skills,

on the other hand, will have a smaller interruption effect and is more likely to show impacts within a short period.

4.2. Contextual and Site Effects

As noted in Section 2, a common finding in the ALMP literature is the presence of site effects: differences in the apparent effectiveness of the program at different sites. Several types of implementation data and contextual information can be useful in understanding these effects. Among these are information on the type of program and the quality of implementation at each site, and information on labor market conditions at each site (e.g., the local unemployment rate). The latter information is especially useful in a DD design, since even with a well-designed comparison group the impact of the program may be confounded by changing labor market conditions experienced by the participant group and the comparison group.

5. Ex Post Analysis

This section presents a brief "check list" of the analyses that are normally conducted in an ALMP evaluation once the data have been collected and assembled. In principle a complete evaluation design includes a full description of the intended analysis. In practice, such "pre-specified" analyses are usually supplemented with a variety of additional analyses, depending on the issues that arose in the actual implementation. A credible evaluation should always include the "pre-specified" analyses that were originally intended in the design, as well as any additional results that are warranted by the context. For example, if a smaller fraction of the participant group than the control group was interviewed in the follow-up survey, the analysis should include *both* simple comparisons of outcomes between the groups (as were originally specified in the evaluation design) and comparisons that address the differential follow-up rate using statistical adjustment techniques.

5.1. Check List for a Randomized Design

The main steps in the analysis of the data from a randomized ALMP evaluation are:

i) Check of randomization

ii) Check of response rates to follow-up survey

iii) Analysis of compliance /cross-over

iv) Estimation of intention to treat impact estimates, with and without adjustments

v) Adjustment of intention to treat estimates for non-compliance/cross-over

5.1.1 Check of Randomization

There are two simple checks of randomization. First, all of the *baseline* characteristics of the treatment and control groups should be compared. Since small differences will emerge randomly, and one or two characteristics may appear to differ significantly even by chance, a second useful test is to fit a logistic (or probit) model for treatment group status using all available baseline characteristics. The overall chi-squared statistic for this model –which evaluates the orthogonality between treatment status and the baseline covariates– should not exceed a relevant critical value. Very rarely, of course, even when randomization was correctly implemented this test will reject orthogonality. If orthogonality is rejected, it is extremely important to review the implementation of the experiment and carefully consider whether randomization actually failed,

or whether the rejection is simply an "unlucky event". In the ALMP context, randomization failures will often arise because of imperfect compliance with enrollment and/or follow-up protocols. Typically, such failures will vary by site: thus, when a failure is suspected it is useful to examine the data site-by-site and look for patterns that indicate a problem.

5.1.2 Check of Response Rates to Follow-Up Survey

Typically, despite the best efforts of the design team and the survey group, it will be impossible to obtain complete data in the follow-up survey for everyone originally assigned to the treatment and control groups. If non-response is entirely random this does not pose a particular problem. If non-response rates vary by group, however, there may be biases in the observed outcomes of those in the follow-up survey (relative to the overall population in the two groups) that differ by group. Such *differential response bias* poses a threat to the interpretation of a well designed evaluation.

The issue of response bias can be addressed by a simple comparison of mean response rates of the two groups.[27] Though equality of response rates does not *guarantee* that there is no differential response bias, conventional models of response bias have the property that two samples with the same distribution of characteristics and the same response rate have the same response bias (Lee and Lemieux, 2009). Since under random assignment the treatment and control groups have the same characteristics (other than treatment status), equality of the response rates of the two groups implies that both have similar response biases.

If this comparison reveals that response rates vary significantly by group, a secondary analysis will be necessary to probe the robustness of the estimated program impacts (see below).

[27] A multivariate analysis (using a logistic or probit model) can also be conducted, relating the probability of response to baseline characteristics and a dummy for assignment to the treatment group. Under random assignment the covariates should be independent of treatment status so this should not give results that are too different from a simple comparison.

5.*1.3 Analysis of Compliance /Cross-Over*

The issue of compliance is addressed by a simple cross-tabulation of assignment status (treatment versus control group) and actual program receipt. Note that this requires that information be available on the actual program participation of both groups. With full compliance, assignment status is the same as actual program receipt. With dropouts, no-shows, and control group cross-over, however, the two concepts differ. Under the plausible restriction that assignment to treatment only increases the likelihood of actual program receipt, the observed intention to treat (ITT) effect (which is just the difference in outcomes between the originally assigned groups) can be translated to a estimate of the effect of "treatment on the treated" (TOT) by dividing the former by the difference in participation rates of the treatment and control groups (D): TOT=ITT/D (Angrist, Imbens, and Rubin, 1996).

One issue that often arises is how to define "participation" in the presence of dropout behavior. Sometimes, participation is defined as having completed a minimum fraction of the typical length of a program (e.g., at least 8 weeks of a program with mean duration of 16 weeks for program completers). The appropriateness of a particular definition depends in part on the costs associated with non-completers. If the costs are about the same for all participants, regardless of completion, then a useful measure of participation would include dropouts, though perhaps not "no-shows".

5.1.4 Estimation of Intention to Treat Impacts, with and without Adjustments

The basic tool used for impact estimation in most randomized ALMP evaluations is linear regression ("ordinary least squares" or OLS). Most often, the analysis of a specific outcome begins with a specification that includes only a single indicator for treatment group status.[28] The estimated coefficient is the basic experimental impact estimate (ITT estimate). Estimates from OLS models that include baseline characteristics can then be included. These should not vary much from the unadjusted estimate but may be a little more (or less) precise than the simpler estimate from a univariate model (Freedman, 2008). If the treatment effect is thought to vary substantially by site, it may be useful to construct differences in outcomes site-by-site, and then take a weighted average of these differences, using as weights the fraction of the treatment group

[28] The construction of appropriate sampling errors will not be discussed in detail here. In some situations it may be appropriate to present standard errors that allow a site effect (i.e., clustered by site).

(or of the overall treatment and control groups) at each site. This may differ slightly from the basic OLS impact estimate, since OLS implicitly weights the site-specific estimates using a different weighting scheme (Angrist and Pischke, 2009). When the outcome of interest is a dichotomous variable (e.g., employment status) researchers sometimes supplement the OLS models with results from logistic regression models, though the functional form of a logistic model cannot be justified by randomization. Except in special cases where the outcome is relatively rare (e.g., probability < 5%) or very prevalent (e.g., probability > 95%) the marginal effects from these models will usually be very close to the OLS estimates.

A second set of models may be needed if there is evidence of differential response to the follow-up survey. A useful approach is the bounding procedure suggested by Lee (2009). If, for example, a larger fraction of the treatment group responded to the survey, then Lee's procedure is to selectively drop observations from the treatment group until the fraction retained is equal to the response rate of the control group. To obtain an upper bound on the effect of treatment Lee proposes to drop observations with the lowest values of the outcome variable. To obtain a lower bound on the effect of treatment Lee proposes dropping observations with the highest values of the outcome variable.[29] When the difference in response rates is not too large and the upper and lower bounds of the outcomes of interest are not too extreme, the bounds from Lee's procedure will be relatively tight. When the gap in response rates is relatively large or the outcome is highly disperse the bounds will be wider.

5.1.5 Adjustment of Intention to Treat Estimates for Non-Compliance/Cross-Over

When there is significant non-compliance it may be useful to construct estimates of the effect of treatment on the treated (TOT). A simple informal approach is to divide the estimate of the intention to treat effect by the difference in program participation rates (D) between the treatment group and the control group. If, for example, 75% of the treatment group received the program (i.e., one quarter dropped out early or never showed up), and 25% of the control group received something close to the program (by evading the embargo or enrolling at other similar programs) then the difference in participation rates is 50% and the adjustment amounts to multiplying the

[29] If the response rate of the control group is higher, then the procedure is to selectively drop observations from the control group: deleting observations with the highest value of the outcome variable to obtain a upper bound, and observations with the highest values of the outcome variable to obtain an lower bound.

ITT by a factor of 2. A more formal approach is to estimate a two-stage least squares (TSLS) model for the outcome(s) of interest, expressing the outcome as a function of actual program participation, and using treatment status as an instrumental variable for program participation. This will give the same estimate of the TOT (=ITT/D), but will provide an appropriate sampling error for the TOT.

5.2. Check List for a Difference in Differences Design

The main steps in the analysis of the data from a DD evaluation design of an ALMP are:

i) Comparison of pre-program outcomes for participant group and comparison group

ii) Check of response rates to follow-up survey

iii) Analysis of compliance /cross-over

iv) Estimation of intention to treat impact estimates, with and without adjustments

v) Adjustment of intention to treat estimates for non-compliance/cross-over

5.2.1 Comparison of Pre-Program Outcomes for Participant Group and Comparison Group

The most important check for a DD design – and the only way to evaluate the plausibility of the selected comparison group –is to compare the *pre-program* outcomes of the participant group and the comparison group. Whenever possible this should be conducted using the same outcome variables that are the main focus of the evaluation: e.g., earnings and or employment rates. The comparison should begin with a graph or simple table of mean outcomes for the two groups in each quarter or month prior to the point in time when the participant group entered the program.[30] The analysis can be extended by plotting (or tabulating) regression-adjusted or reweighted outcomes for the two groups.

Ideally the outcomes of the two groups follow a strictly parallel path (i.e., a constant differential between the groups) in the pre-program period. When this is not true, the design is potentially threatened, since it is no longer clear that the difference between the groups would have remained constant except for the impact of the program.

[30] Unlike a randomized evaluation there is no common "point of random assignment" for the program group and the comparison group. For the program group it is natural to treat the period just prior to entry into the program as a "base point". In many DD designs it is easy to designate a similar period for the comparison group. In other cases –such as a design that uses a comparison group drawn from an earlier or later time period – the designation of the base point for the comparison group is more difficult (or more arbitrary).

5.2.2 Check of Response Rates to Follow-Up Survey

Assuming that a specialized survey will be used to collect data on post-program outcomes of the participant and comparison groups, the same issue of selective non-response arises in a DD design as in an experimental design. As above, the issue of potentially differential response bias can be addressed by a simple comparison of mean response rates of the two groups.[31] If this comparison reveals that response rates vary significantly by group, a secondary analysis may be necessary, using Lee's bounding procedure.

5.2.3 Analysis of Compliance /Cross-Over

As in an experimental design, the issue of non-compliance in a DD design is addressed by a simple comparison of program participation rates between the designated program group and the designated comparison group. This requires that information is available on the actual program participation of *both* groups. If there is significant cross-over, the basic impact estimates from the evaluation will be interpreted as intention to treat impacts (ITT).

5.2.4 Estimation of Intention to Treat Impacts, with and without Adjustments

The basic tool used for impact estimation in most DD designs is ordinary least squares (OLS) regression. The regression model takes as the dependent variable the outcome of interest (e.g. quarterly or monthly earnings), measured for members of the program group and the comparison group in *both* the pre-program period (or periods) and the post-program period (or periods). The key independent variables are (i) an indicator for members of the program group; (ii) an indicator for the post-program period; (iii) the *interaction* of program group status and the post-program period. The interaction coefficient measures the difference in differences between the two groups in the post-program period relative to the pre-program period. The results from the regression model can be supplemented with a simple graph showing the mean outcomes for the two groups in each available pre- and post-program period.

This basic model can be extended by adding additional covariates (measured in the pre-program period) and by adding additional observations on pre-program (or post-program)

[31] A multivariate analysis (using a logistic or probit model) can also be conducted, relating the probability of response to baseline characteristics and a dummy for assignment to the treatment group. Under random assignment the covariates should be independent of treatment status so this should not give results that are too different from a simple comparison.

outcomes for the two groups. Provided that the pre-program differential between the two groups is constant, the estimated differences in differences will be (approximately) constant when additional periods of pre-program data are added to the estimation sample.

As in an experimental design, a second set of models using Lee's bounding approach may be needed if there is evidence of differential response to the follow-up survey.

5.2.5 Adjustment of Intention to Treat Estimates for Non-Compliance/Cross-Over

When there is significant non-compliance, the DD estimate of the ITT effect can be converted to an estimates of the effect of treatment on the treated (TOT) using the same approaches described for an experimental design. Specifically, the estimated intention to treat effect can be divided by the difference in program participation rates between the treatment group and the control group to obtain a point estimate of the TOT effect. Alternatively, the DD regression model can be estimated by TSLS, expressing the outcome of interest for both groups as a function of actual program participation, and using program group status as an instrumental variable for program participation.

6. Example: Dominican Republic's Programa Juventud y Empleo – PJE

6.1 The program and its first evaluation[32]

The PJE –Youth and Employment Program– is a Dominican Republic ALMP designed to improve the labor market insertion of youth between the ages of 16 and 29 who have not completed high school studies. PJE was one of the first training programs in LAC to incorporate a randomized evaluation design. The program offers a wide range of job training courses focused on occupations like bakers, stylists, clerks, car repairers, and bartenders. The Ministry of Labor outsources the provision of training services to private training institutions (COS, Centers Operating in the training System). Courses (with a maximum duration of 350 hours) are conducted at the COS's facilities and split into two components: basic skills training, and technical/vocational training. Basic skills training is meant to strengthen trainees' self- esteem and work habits, while vocational training is targeted to the needs of local employers.

Eligible youth are recruited and screened by the training providers according their preferred vocations and the availability of their desired courses. Once a COS has successfully recruited 35 potential participants, it sends their names to a centralized program management unit that randomly selects 20 treatment group members and 15 control group members. People assigned to the treatment group are notified by telephone of the starting date and location for their course assignment. Among those assigned to the control group, 5 individuals are randomly selected to be placed on an ordered waiting list, and are added to the treatment group (if needed) to take the place of trainees who fail to show up for their program assignment. This process allows the COS's to maintain a relatively high program utilization rate, while ensuring the integrity of the randomized design.

The original analysis of the JE program was based on a sample of applicants for the second cohort of trainees, who applied for training in early 2004 (see Card et al. 2011). Baseline data were collected from a sample of applicants prior to random assignment. A follow-up survey was administered in the period from May to July 2005; some 10 to 14 months after most trainees had finished their initial course work. Simple comparisons between trainees in the follow-up survey and members of the control group show little impact on employment, although there is some evidence of a modest impact (10%) on wages. Unfortunately, however, the randomized

[32] This section is based on Card et al. (2011)

design of the JE evaluation was potentially compromised by the failure to include in the follow-up survey people who were originally assigned to receive training but failed to show up (or attended for only a very short time).

As in most programs with voluntary participation, in JE there was imperfect compliance and, hence, some selected for treatment did not participate in the courses, while some of those initially assigned to the control group were re-assigned to the treatment group.

The non-compliance issue is illustrated by a comparison of initial program assignment status and final training program participation status for the second cohort of JE applicants. Of the 8,365 applicants, 2,564 (31%) were initially assigned to the control group while 5,801 (69%) were assigned to the treatment group. Among the treatment group, 1,011 (17%) failed to show up or dropped out during the first two weeks of the program for training, while 4,791 (83%) are recorded as receiving at least two weeks training. To fill the places of the no-shows, 941 members of the original control group were reassigned to the treatment group, leaving a "realized control group" of 1,623.

Assuming that the reassignment process followed the protocol described above, the realized control group is a random selection of the original applicants and their outcomes can be used to obtain valid estimates of labor market outcome for the applicant population in the absence of the training intervention. The analogous "realized treatment group" of 5,723 (=4,791+941) is more problematic, because it excludes the 1,011 "no shows" who made a conscious choice *not* to participate in the training program. If the realized treatment group were extended to include *all* the originally assigned trainees (as well as the re-assigned controls) there would be no problem, because this group is again a random sample of the original applicant population. In the initial JE evaluation however, no information was collected on the post-program outcomes of the no-shows. The realized treatment group without the no-shows is unfortunately *not* a random sample of the original applicant population. Nevertheless, the initial evaluation of the JE program had to rely on a sample based on this group to provide information on the outcomes of the applicant population after exposure to training.

The key outcome measures in the JE evaluation are an indicator for being employed at the date of the follow-up survey, and total labor market earnings in the month prior to the follow-

up survey (which are equal to zero for non-workers).[33] The realized treatment group had a slightly higher employment rate (57.4% versus 56.0%) and somewhat higher average earnings (3,133 Pesos/month versus 2,677). The 1.5 percentage point difference in employment rates is not statistically significant (t=0.5) while the 455 peso difference in monthly earnings is significant at conventional levels (t=2.13) (see Table 2). Interestingly, the outcomes for the treatment group excluding the reassigned controls are slightly less positive, and not statistically significantly different from the outcomes from the control group.

Table 2. Impacts of Assignment to Training on Employment and Earnings Outcomes for Sample of 2004 Applicants to JE Program

	Impact on Employment Rate:		Impact on Monthly Earnings:	
	All	Exclude Reassigned	All	Exclude Reassigned
	(1)	(2)	(3)	(4)
a. OLS Models fit to Observations in Follow-up Survey Only				
1. No covariates	1.5	0.4	455	284
	(2.7)	(2.9)	(213)	(211)
2. With covariates (individual covariates and 11 region effects)	1.1	0.7	415	294
	(2.7)	(2.8)	(201)	(200)
3. With covariates and ICAP effects	1.3	1.1	390	288
	(2.7)	(2.9)	(205)	(204)

Notes: standard errors in parentheses. Entries in panel a. are coefficients of assignment to treatment dummy in linear models for the probability of employment at follow-up survey (columns 1-2) and the monthly labor earnings at follow-up survey (columns 3-4). These models include individual covariates and 11 region effects. Joint models in columns 1 and 2 combine a probit model for attending training if assigned to treatment, and a probit model for employment at the time of follow-up. Joint model in columns 3 and 4 combine probit model for attending training if assigned and linear model (with normally distributed error) for income. Models in columns 1 and 2 include re-assigned control group members in the treatment group. Models in columns 3 and 4 exclude these individuals.

The estimated impacts on employment are all fairly close to zero and there are no significant differences by gender, age, education, or location. The estimated impacts on monthly

[33] Just under 3% of those who are recorded as working report zero earnings. All those with earnings are recorded as employed.

earnings are fairly similar for men and women, and for younger and older workers, but show interesting patterns by education and location. In particular, the overall impact on earnings seems to be generated by a large positive effect for better-educated workers (adjusted impact = 807, t=2.54) coupled with a minimal effect for the less-educated. Distinguishing by location there is also a relatively large positive effect for residents of Santo Domingo (adjusted impact = 804, t=2.71) coupled with a minimal effect for those outside the capital city. Comparisons between better-educated applicants in Santo Domingo and all others are even more striking: this subgroup accounts for virtually all of the observed positive impact on monthly earnings. While interesting, it is important to note that these findings must be interpreted cautiously, since the subsample of largest impact was determined after the fact, rather than based on an ex ante analysis plan.

The lack of data for the no-shows in the follow-up survey used for the initial JE evaluation means that the observed mean outcomes for the realized treatment group are potentially biased estimates of the means for everyone who was initially assigned to treatment.[34] In 2010-2011 a second round of data was collected to evaluate the cohort of trainees that completed the program in 2009. The follow-up survey for this evaluation was conducted some eighteen months after course completion. The new evaluation retained the randomized design of the first evaluation. However the implementation was strengthened using lessons learned from the first evaluation. In particular the sample frame for the follow-up survey was extended to include no-shows and dropouts, the size of the follow-up survey was substantially increased, and extra resources and effort were devoted to achieving high response rates for the follow-up survey.

[34] The paper by Card et al (2011) presents a variety of non-experimentally based program estimates that attempt to address the non-randomness issue. In the case of employment, a simple bound can be constructed that is completely agnostic about the behavior of the missing no-show group (Manski, 1989). Unfortunately, this bound is relatively wide and provides little information about the impact of the program.

6.2 Applying the Check List for a Randomized Design to the 2009 JE Cohort[35]

6.2.1 Check of Randomization

The 2009 cohort of JE applicants included 10,309 applicants who met the selection criteria for the program. Of these, 5,914 (57%) were assigned to the treatment group and 4,395 (43%) were initially assigned to the control group. As shown in Table 3, both groups have the same characteristics in terms of age, gender, marital status, education, labor force participation and place of residence. The table also shows the poverty index of the household, where there are also no significant differences. Appendix 1 shows some characteristic from the baseline survey and a t-statistic for equality of means between the initially assigned groups as an evidence of randomness. It also illustrates the characteristics of the subsets of the initially assigned treatment and control groups who received training and those who did not. Overall, the 2009 applicant cohort program was predominantly female (64%), about one-quarter were married, most had some high school education, 90% were from urban areas, and about one-quarter were from Santo Domingo, the capital city.

Only 4% of 2009 JE applicants reported they were employed at the time of the baseline survey, while 52% reported being unemployed and 44% were out of the labor force. This low rate of pre-program labor market activity is presumably a reflection of the self-selection process underlying the decision to participate in the JE program, though it may also reflect some under-reporting of employment status by applicants who were aware of the program eligibility rules (which required applicants to be out of work). The low rate of labor market activity of applicants is also reflected by the relatively low number of jobs they report having held prior to baseline survey (less than one on average). Overall it appears that many applicants had never worked before.

[35] While this data is currently being analyzed, we present the process and preliminary results here to illustrate the implementation of the evaluation design. For the most recent version of the 2011 data analysis, please contact pibarraran@iadb.org

Table 3. Check of Randomization, Treatment and Control Status Defined by Lottery

Characteristic	*Zi = 1* *(a)*	*Zi = 0* *(b)*	Difference (Zi) (a) - (b)	*t*
Age	22.03	21.99	0.04	0.59
Gender (male = 1)	0.37	0.38	-0.01	-1.35
Marital status (married = 1)	0.24	0.24	0.00	-0.03
Number of children	0.71	0.70	0.01	0.42
Attend school (currently)	0.23	0.23	0.01	0.76
Incomplete elementary	0.20	0.20	0.00	-0.30
Complete elementary	0.05	0.05	0.00	-0.20
Incomplete high school	0.55	0.58	-0.03	1.09
Complete high school	0.04	0.03	0.00	0.79
More than high school	0.00	0.00	0.00	0.02
Missing education	0.04	0.04	0.00	-0.65
No data on education	0.11	0.12	-0.01	-1.19
Fraction with prior work experience	0.20	0.22	-0.02	1.37
Currently employed	0.04	0.04	0.00	0.15
Currently salaried worker	0.02	0.02	0.00	-0.23
Currently unemployed	0.53	0.52	0.00	0.28
Urban areas	0.89	0.89	0.00	-0.47
Live in Santo Domingo	0.33	0.33	0.00	-0.39
Receives remittances	0.11	0.11	0.00	-0.31
Observations	5,914	4,395		

Source: PJE baseline and follow-up data 2011.
Note: Clustered standard errors at course level in parenthesis. ***: significant at 1%; **: significant at 5%; *: significant at 10%.

To further test randomization, a logit model was estimated to predict participation based on the variables included in the baseline registration form, and the overall chi-squared statistic did not reject orthogonality between treatment status and the baseline covariates (p-value=0.8372).

6.2.2 Check of Response Rates to Follow-Up Survey

A follow-up survey of a sample of the 2009 JE cohort was conducted between November 2010 and February 2011 (18 to 24 months after their initial application to the program).[36] The sample

[36] In spite of the fact that the survey was interrupted by the end of the year, the field work was randomized to

size for the follow-up survey was set at 5,000. This size was determined by specifying a minimum power of 0.8 to detect an 8% effect on the employment rate, and assuming a 70% success rate in obtaining completed surveys from the initially targeted sample.[37] The sample consisted of 3,250 from the treatment group and 1,750 from the control group. The actual completion rate for the survey was somewhat higher than expected (81%), with nearly identical rates for the treatment group (80.8%) and control groups (80.4%).

We used baseline data to verify the similarity of the interviewed treatment and control groups and inspected differences in the basic characteristics of those that were interviewed from to those that were not within treatment and control groups (see Table 4). In this way we verify that randomization holds in the interviewed sample, and that there was no differential attrition between treatment and controls.

mitigate the impact of the seasonality on part of the survey. In fact 55.6% and 55.2% of the control and treatment groups were surveyed in December 2010, respectively.

[37] Several simulations were taken into account with our budget constraint and considering the outdated contact information of the registration forms. We used the Stata command *sampsi*.

Table 4. Check of Randomization in Follow-Up Data in Original and Realized Groups

Characteristic	Treatment - Control Original Sample		Treatment - Control Realized Sample		Original versus realized Treatment		Control	
	Diff	*T*	*Diff*	*t*	*Diff.*	*t*	Diff	*t*
Age	0.14	1.41	0.14	1.22	0.22	0.51	0.21	0.24
Gender (male = 1)	-0.01	-0.92	-0.01	-0.74	0.00	0.08	0.00	0.18
Marital status (married = 1)	0.00	0.34	0.01	0.84	0.00	0.35	0.01	1.22
Number of children	0.02	0.59	0.01	0.30	0.00	0.36	-0.01	-0.91
Attend school (currently)	0.00	0.05	0.00	0.21	-0.01	2.21**	-0.01	1.08
Incomplete elementary	0.00	-0.08	0.00	-0.36	0.00	-1.33	0.00	-0.18
Complete elementary	0.00	0.20	0.00	0.06	0.00	0.20	0.00	0.48
Incomplete high school	0.01	0.81	0.02	0.92	-0.01	1.78*	0.00	0.74
Complete high school	0.00	0.52	0.00	-0.05	0.00	-0.65	0.00	1.07
More than high school	0.00	1.25	0.00	1.53	0.00	1.58	0.00	-0.25
Missing education	0.00	-0.49	0.00	-0.08	0.00	-1.93*	0.01	-2.14
No data on education	-0.01	-0.08	-0.01	-1.24	0.00	0.34	0.00	-0.08
Number of jobs prior to PJE	-0.01	-1.03	-0.01	0.54	0.00	0.24	0.00	-0.67
Currently employed	0.00	-0.20	-0.01	-0.74	0.00	0.93	0.00	0.95
Currently salaried worker	0.00	-1.01	0.00	-0.83	0.00	0.05	0.00	0.24
Currently unemployed	0.01	0.51	0.01	0.80	0.00	0.55	0.00	0.54
ICV*	-0.02	-0.05	0.14	0.43	0.13	1.60	0.29	2.87**
Urban areas	0.00	-0.25	0.00	-0.47	0.01	3.94***	0.01	1.91*
Live in Santo Domingo	-0.01	-0.54	-0.01	-0.39	-0.02	-0.77	-0.02	-0.81
Recieves remittances	0.00	0.10	0.00	0.41	0.00	0.62	0.00	0.45
Obs Treatment Group	3250		2639					
Obs Control Group	1750		1407					

Source: PJE baseline and follow-up data 2011.

Note: Clustered standard errors at course level in parenthesis. ***: significant at 1%; **: significant at 5%; *: significant at 10%. One Dominica Peso = 0.026 US Dollars.

The first two columns show that there were no differences in the original sample of

5,000, and balancing was maintained in the realized sample of 4,046 interviewees. The last four columns show the differences within treatment and controls in terms of the original and realized samples. While there are some small statistical differences, they are substantially very small and overall balance is maintained. Thus, the analyses were done using the random design of the evaluation.

6.2.3 Analysis of Compliance/Cross-Over

Having checked that randomization was well implemented and the follow-up rate was high and there is no evidence of differential attrition, the next step is to analyze compliance. This was done in Table 5 based on the information from the follow-up survey, where respondents were asked whether or not they participated in a training program and received the stipend irrespectively of their participant/non-participant status in the administrative records.

Table 5. Final Composition of Treatment Groups

		Lottery	
		Control	*Treatment*
Participation	*No*	884	380
		63%	14%
	Yes	523	2,249
		37%	86%
TOTAL		1,407	2,629
		100%	100%

Source: PJE baseline and follow-up data 2011 and administrative data.
Note: Participation is defined as those who accepted and began the courses.

6.2.4 Estimation of Intention to Treat Impacts, with and without Adjustments

One of the advantages of using a random assignment design is that, when well implemented, the analysis is straightforward. A simple comparison of those assigned and those not assigned to treatment provides an unbiased estimate of the "intention-to-treat" of the program. The table below shows this comparison that was computed through standard linear regression, clustering standard errors at course level.

Table 6 shows the results of the basic analysis, for the complete sample and for Santo

Domingo (where the first evaluation suggested that there were some impacts). For both samples, the results indicate that the program was ineffective in increasing labor market outcomes for participants. Although most of the coefficients are positive, they are small and statistically insignificant. The evaluation design would detect impact of 3.5 percentage points in overall employment (about 6% over the mean for controls) and 6.3 percentage points in Santo Domingo (about 11% over the mean for controls), but the coefficients are virtually zero for the complete sample and close to 1.5 percentage points in Santo Domingo.

Table 6. Estimation of Intention to Treat Impacts, with and without Adjustments

Intention to Treat Effect Outcomes	(1) All	(2) All (adjusted)	(3) Santo Domingo	(4) Santo Domingo (adjusted)
Employed	-0.0055	-0.0028	0.0163	0.0148
	(0.0174)	(0.0167)	(0.0332)	(0.0313)
mean controls	0.622	0.627	0.591	0.591
Employed with health insurance	0.0072	0.0088	0.0243	0.0234
	(0.0123)	(0.0122)	(0.0201)	(0.0195)
mean controls	0.174	0.161	0.172	0.172
Monthly earnings	75.63	93.5500	271.62	246.71
	(115.54)	(110.97)	(222.71)	(218.14)
mean controls	2464	2464	2436	2436
Hourly earnings	-0.11	-0.0730	1.71	1.19
	(0.95)	(0.92)	(1.82)	(1.70)
mean controls	18.38	18.38	18.22	18.22
Labor force participation	0.0173	0.0194	0.0277	0.0285
	(0.0128)	(0.0126)	(0.0232)	(0.0222)
mean controls	0.834	0.834	0.837	0.837
Observations	3,876	3,876	1,249	1,249

Source: PJE baseline and follow-up data 2011.

Note: Labor participation is defined as those youngsters that either work or seek for a job.
Clustered standard errors at course level in parenthesis. ***: significant at 1%; **: significant at 5%; *: significant at 10%.
Adjusted models are those controlled by gender, age and education.
One Dominica Peso = 0.026 US Dollars.

6.2.5 Adjustment of Intention to Treat Impacts for Non-Compliance/Cross-Over

Finally, Table 7 shows the analyses that takes into account non-compliance, the fact that some of the youngsters that won the lottery did not take the course, and that some that did not win ended up participating in the training. As explained above, these analyses yields the Treatment-on-the-Treated effect, and it is basically a scaling up of the Intention-to-Treat estimates based on the difference in participation rates of treatment and control groups, which as seen in Table 5 was

about 50%. Thus, the coefficients are larger but so are the standard errors and the results remain not-significant.

Table 7. Adjustment of Intention to Treat Estimates for Non-Compliance/Cross-Over

Treatment Effect on the Treated Outcomes	(1) All	(2) All (adjusted)	(3) Santo Domingo	(4) Santo Domingo (adjusted)
Employed	-0.0112	-0.0056	0.0308	0.0279
	(0.0352)	(0.0339)	(0.0615)	(0.0585)
mean controls	0.627	0.627	0.584	0.627
Employed with health insurance	0.0147	0.0179	0.0458	0.0441
	(0.0250)	(0.0247)	(0.0370)	(0.0361)
mean controls	0.161	0.161	0.174	0.161
Monthly earnings	152.38	188.18	511.9061	462.89
	(232.83)	(223.29)	(415.1745)	(407.84)
mean controls	2440	2440	2468	2440
Hourly earnings	-0.21	-0.15	3.3124	2.29
	(1.93)	(1.86)	(3.4526)	(3.23)
mean controls	17.85	17.85	17.16	17.85
Labor force participation	0.0352	0.0394	0.0523	0.0537
	(0.0260)	(0.0256)	(0.0431)	(0.0416)
mean controls	0.817	0.817	0.829	0.817
Observations	3,876	3,876	1,249	1,249

Source: PJE baseline and follow-up data 2011.

Note: Labor participation is defined as those youngsters that either work or seek for a job. Clustered standard errors at course level in parenthesis. ***: significant at 1%; **: significant at 5%; *: significant at 10%. Adjusted models are those controlled by gender, age and education.
One Dominica Peso = 0.026 US Dollars.

7. Conclusions

The purpose of this guideline is to provide program managers with the necessary elements to incorporate a rigorous evaluation design in an ALMP. This is done by discussing the methods available emphasizing that the most credible and straightforward evaluation method is a *randomized* design, in which a group of potential participants is randomly divided into a treatment and a control group. Random assignment ensures that the two groups would have had similar experiences in the post-program period in the absence of the program intervention. The observed post-program difference therefore yields a reliable estimate of the program impact.

A second approach discussed in this guideline is the *difference in differences* design that compares the *change* in outcomes between the participant group and a selected comparison group from before to after the completion of the program. In general the outcomes of the comparison group may differ from the outcomes of the participant group, even in the absence of the program intervention. If the difference observed prior to the program would have persisted in the absence of the program, however, then the *change* in the outcome gap between the two groups yields a reliable estimate of the program impact.

Most importantly, this guideline reviews the various steps in the design and implementation of ALMP's, and in subsequent analysis of the program data, that will ensure a rigorous and informative impact evaluation using either of these two techniques. These practical steps will enable program managers to implement successful evaluations that will allow to measure the impact of the program and to get evidence to make policy decisions.

References

Abadie, Alberto. (2005). "Semiparametric Difference-in-Differences Estimators," *Review of Economic Studies*, 72(1): 1-19, 01

Angrist, Joshua and G. Imbens. (1994) "Identification and Estimation of Local Average Treatment Effects". *Econometrica*, 62(2): 467-475

Angrist, Joshua and J. Pischke. (2009). Mostly Harmless Econometrics: An Empiricist's Companion. Princeton NJ: Princeton University Press.

Angrist, Joshua D., G. W. Imbens and D. B. Rubin. 1996. "Identification of Causal Effects Using Instrumental Variables." *Journal of the American Statistical Association* (91): 444-55.

Ashenfelter, Orley. (1978). "Estimating the Effects of Training Programs on Earnings." *Review of Economics and Statistics* (60): 47-57.

Ashenfelter, Orley and D. Card (1985) Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs. *Review of Economics and Statistics*, 67(4): 648-660.

Betcherman, G., A. Dar, A. Luinstra and M. Ogawa (2000) "Active Labor Market Programs: Policy Issues for East Asia." SP, Discussion Papers, 0005, The World Bank

Betcherman, G. K. Olivas, and A. Dar (2004) "Impacts of Active Labour Market Programs: New Evidence from Evaluations with Particular Attention to Developing and Transition Countries."World Bank Social Protection, Discussion Paper 0402. The World Bank

Card, David, P. Ibarrarán, F. Regalia, D. Rosas-Shady and Y. Soares (2011) "The Labor Market Impacts of Youth Training in the Dominican Republic." *Journal of Labor Economics* 29(2): 267-300.

Card, David; J. Kluve, and A. Weber. (2009). Active Labor Market Policy Evaluations: A Meta-Analysis, NRN Working Papers 2009-02.

Card, David and D. Sullivan. (1988). "Measuring the Effects of Subsidized Training Programs on Movements In and Out of Employment". *Econometrica*, 56(3): 97-530.

DiNardo, John and D. Lee. (2010). "Program Evaluation and Research Designs," NBER Working Papers 16016, National Bureau of Economic Research.

Duflo, E., R. Glennerster, and M. Kremer. (2006). Using Randomization in Development Economics Research: A Toolkit (December 12, 2006). MIT Department of Economics Working Paper No. 06-36. Available at SSRN: http://ssrn.com/abstract=951841

Freedman, David A. (2008). "On Regression Adjustments to Experimental Data". *Advances in Applied Mathematics*, (40): 180-193.

Heckman, James J., R. J. Lalonde, and J. Smith. (1999). "The Economics and Econometrics of Active Labor Market Programs." In Orley Ashenfelter and David Card (editors), *Handbook of Labor Economics*, Volume 3A: 1865-2097. New York: Elsevier

Heinrich, Carolyn, A. Maffioli, and G. Vázquez. (2010) A Primer for Applying Propensity-Score Matching. SPD Impact-Evaluation Guidelines.Technical Notes. IDB-TN-161, Inter-American Development Bank

Imbens, Guido W. and T. Lemieux, (2008). Regression Discontinuity Designs: A Guide to Practice, *Journal of Econometrics*, 142(2): 615-635.

Ibarrarán, Pablo and David Rosas (2009) "Evaluating the impact of Job Training Programs in Latin America: evidence from IDB funded operations," *Journal of Development Effectiveness*, 2(1): 195-216.

Hotz, V. Joseph, G.W. Imbens and J. A. Klerman. (2006) "Evaluating the Differential Effects of Alternative Welfare-to-Work Training Components: A Re-Analysis of the California GAIN Program." *Journal of Labor Economics* (24): 521-566.

Lechner, Michaek. (2002). Microeconometric Evaluation of Active Labour Market Policies. University of St. Gallen, Department of Economics, Discussion Paper No. 2002-20

Lee, David. (2009). "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects." *Review of Economic Studies*, 6(3): 1071-1102.

Lee, D. and T. Lemieux, (2009). Regression Discontinuity Designs in Economics. *NBER* Working Paper No. 14723.

Manski, Charles F. (1989). "Anatomy of the Selection Problem," *Journal of Human Resources*, (24): 343-60.

Sianesi, Barbara. (2004). “An Evaluation of the Swedish System of Active Labor Market Programs in the 1990s”. *The Review of Economics and Statistics*, 86(1): 133-155, 09.

Appendix 1: Basic characteristics at baseline

Characteristic	*Originally Assigned To Treatment*	*Training Participation Of Original Treatment Group)*		*Originally Assigned To Control*	*Training Participation Of Original Control Group*		Difference: Original Treatment/Control	
	(a)	*Received Training*	*Dropout/ No-show*	*(b)*	*Reassigned*	*Not Reassigned*	Difference	*t-statistic*
Age	22.03	22.00	22.18	21.99	21.75	22.06	0.04	0.59
Gender (male = 1)	0.37	0.36	0.40	0.38	0.41	0.37	-0.01	-1.35
Marital status (married = 1)	0.24	0.24	0.26	0.24	0.22	0.25	0.00	-0.03
Number of children	0.71	0.70	0.73	0.70	0.66	0.71	0.01	0.42
Attend school (currently)	0.23	0.24	0.19	0.23	0.22	0.23	0.01	0.76
Incomplete elementary	0.20	0.20	0.22	0.20	0.20	0.21	0.00	-0.3
Complete elementary	0.05	0.05	0.04	0.05	0.06	0.05	0.00	-0.2
Incomplete high school	0.59	0.60	0.57	0.58	0.58	0.58	0.01	1.39
Complete high school	0.04	0.04	0.03	0.03	0.03	0.03	0.00	0.79
Number of jobs prior to PJE	0.25	0.24	0.29	0.23	0.25	0.23	0.01	1.37
Currently employed	0.04	0.04	0.06	0.04	0.04	0.04	0.00	0.15
Currently salaried worker	0.02	0.01	0.02	0.02	0.01	0.02	0.00	-0.23
Currently unemployed	0.53	0.53	0.53	0.52	0.55	0.52	0.00	0.28
ICV Score (0 to 100)*	62.81	62.78	62.97	62.93	62.91	62.94	-0.12	-0.59
- ICV (Type I)	0.02	0.03	0.01	0.02	0.02	0.02	0.00	0.95
- ICV (Type II)	0.03	0.03	0.02	0.03	0.03	0.03	0.00	-0.74
- ICV (Type III)	0.26	0.26	0.27	0.25	0.25	0.25	0.01	1.2
- ICV (Type IV)	0.08	0.07	0.08	0.08	0.07	0.08	0.00	0.12
- ICV (Type V)	0.53	0.53	0.54	0.56	0.55	0.56	-0.03	-2.33**
- ICV (Type VI)	0.08	0.08	0.08	0.07	0.07	0.07	0.01	2.26**
Urban areas	0.89	0.89	0.89	0.89	0.88	0.90	0.00	-0.47
Live in Santo Domingo	0.24	0.24	0.25	0.24	0.24	0.24	0.00	0.27
Recieves remittances	0.11	0.10	0.11	0.11	0.10	0.11	0.00	-0.31
Rosenberg's test score	23.95	23.96	23.85	23.80	23.82	23.79	0.15	1.94*
N	5,914	4,937	977	4,395	977	3,418		

Source: PJE baseline data and administrative records.

[a]: Stands for *Indice de Calidad de Vida* (Living Quality Index). It is divided into 6 levels from the poorest to the richest.

Note: Means, differences and t-statistics are calculated by linear regression with robust standard errors. ***: significant at 1%; **: significant at 5%; *: significant at 10%.