WPS3625

Evaluating Anti-Poverty Programs

Martin Ravallion[1]

Development Research Group, World Bank

Abstract: The paper critically reviews the methods available for the *ex-post* counterfactual analysis of programs that are assigned exclusively to individuals, households or locations. The discussion covers both experimental and non-experimental methods (including propensity-score matching, discontinuity designs, double and triple differences and instrumental variables). Two main lessons emerge: Firstly, despite the claims of advocates, no single method dominates; rigorous, policy-relevant evaluations should be open-minded about methodology. Secondly, future efforts to draw more useful lessons from evaluations will call for more policy-relevant measures and deeper explanations of measured impacts than are possible from the classic ("black box") assessment of mean impact.

World Bank Policy Research Working Paper 3625, June 2005

[1] For their comments the author is grateful to Pedro Carneiro. Aline Coudouel, Jishnu Das, Jed Friedman, Emanuela Galasso, Markus Goldstein, Jose Garcia-Montalvo, David McKenzie, Alice Mesnard, Norbert Schady, Paul Schultz, Emmanuel Skoufias, Petra Todd, Dominique van de Walle and participants at a workshop at the Rockefeller Foundation Center at Bellagio, Italy, May 2005.

1. Introduction

Governments, aid donors and the development community at large are increasingly asking for hard evidence on the impacts of public programs claiming to reduce poverty. Do we know if such interventions really work? How much impact do they have? Past "evaluations" that only provide qualitative insights into processes and do not assess outcomes against explicit and policy-relevant counterfactuals are now widely seen as unsatisfactory.

This paper critically reviews the main methods available for the counterfactual analysis of programs that are assigned exclusively to certain observational units. These may be people, households, villages or larger geographic areas. The key characteristic is that some units get the program and others do not. For example, a social fund (providing financial support to community-based projects) might ask for proposals from communities, with preference for those from poor areas; some areas do not apply, and some do, but are rejected. Or a workfare program (that requires welfare recipients to work for their benefits) entails extra earnings for participating workers, and gains to the residents of the areas in which the work is done; but others receive nothing. Or cash transfers are targeted exclusively to eligible households by certain criteria.

After an overview of the classic formulation of the evaluation problem and the generic problems it encounters, the bulk of the paper examines the main methods found in practice. The discussion reviews the assumptions each method makes for identifying a program's impact, how the methods compare with each other and what is known about their performance. Examples are drawn mainly from evaluations in developing countries. The penultimate section attempts to look forward — to see how future evaluations might be made more useful for knowledge building and policy making. The concluding section suggests two key lessons.

2. The archetypal evaluation problem

Impact evaluation (or "counterfactual analysis") assesses outcomes for a specific program relative to one or more explicit counterfactuals. This definition immediately distances the present discussion from many past "evaluations." For example, Kapoor (2002) reviews 78 evaluations of World Bank-funded projects since 1979; all the evaluations were done by the Bank's Operations Evaluation Department (OED).[2] Kapoor found that counterfactual analysis was only used in 21 evaluations; for over two-thirds of OED's evaluations there was no way to assess whether the observed outcomes were in fact attributed to the project being evaluated. This is now widely seen to be unsatisfactory. As we will see in this survey, there are now many examples available of how one can do better in attributing outcomes to interventions.

It will be assumed that the program is already in place — making the task *ex-post* impact evaluation. That includes the evaluation of a pilot project, as an input to the *ex-ante* assessment of whether the project should be scaled up. However, doing *ex-post* evaluations does not mean that the evaluation should start after the program finishes, or even after it begins. Indeed, the best *ex-post* evaluations are designed *ex-ante* — often side-by-side with the program itself.

To assess impact we need data on one or more outcome indicators. The choice of indicator will depend on the aims of the intervention. For example, in the case of a scheme that makes transfers targeted to poor families conditional on human resource investments in their children (such as *Cash-for-Education* in Bangladesh or Mexico's *PROGRESA*[3]) the relevant indicators will be a measure of current income poverty and measures of child schooling and health status (interpretable as indicators of future poverty). We will also need some way of

[2] Many other units in the Bank that do evaluations besides OED, including in the research department, which invariably uses counterfactual analysis.

[3] *PROGRESA* stands for *Program for Education, Health and Nutrition.*

inferring the counterfactual. This is inherently unobserved, since it is physically impossible to observe someone in two states of nature at the same time (participating in a program and not participating). Thus evaluation is essentially a problem of missing data. As we will see, there are many ways of filling in the missing data.

The archetypal formulation of the evaluation problem: Data are collected on an outcome indicator Y_i for unit *i* in a sample of size *n*. For example, Y_i might be the income of household *i* normalized by a household-specific poverty line (reflecting differences in the prices faced in different locations and differences in household size and composition). Some of the sampled units receive the program and some do not; a dummy variable takes the value $D_i = 1$ for units that receive the program and $D_i = 0$ for those that do not. The value of Y_i if unit *i* receives the program is Y_i^T (*T* for "treated") and it is Y_i^C (*C* for "counterfactual") if the program is not received.[4] The individual's gain from the program is $G_i \equiv Y_i^T - Y_i^C$.

We also collect data on a vector of covariates for outcomes ("control variables"), X_i, which includes unity as one element. The most common method of controlling for covariates assumes that outcomes are linear in the control parameters, giving:

$$Y_i^T = X_i\beta^T + \mu_i^T \quad (i=1,..,n) \tag{1.1}$$

$$Y_i^C = X_i\beta^C + \mu_i^C \quad (i=1,..,n) \tag{1.2}$$

The error terms are assumed to satisfy $E(\mu_{0i}|X_i) = E(\mu_{1i}|X_i) = 0$.

Two widely used impact parameters are the "average treatment effect" (ATE), $E(G_i)$, and the "average treatment effect on the treated" (ATET), $E(G_i|D_i = 1)$. The conditional ATE is:

---

[4] In the literature, Y_1 or $Y(1)$ and Y_0 or $Y(0)$ are more commonly used for Y^T and Y^C (respectively). My notation will make it easier to follow which group is which.

$$E(G_i|X_i) = X_i(\beta^T - \beta^C) \tag{2}$$

while the conditional ATET is:

$$E(G_i|X_i, D_i = 1) = X_i(\beta^T - \beta^C) + E(\mu_i^T - \mu_i^C|X_i, D_i = 1) \tag{3}$$

As is well recognized in the literature, the essential problem is estimating (2) or (3) is that (1.1) and (1.2) are not estimable, since we cannot know participants' outcomes in the counterfactual (Y_i^C when $D_i = 1$) and counterfactual outcomes under treatment (Y_i^T when $D_i = 0$). To try to get around this problem, suppose that we estimate (1.1) on the sub-sample for which $D_i = 1$ while (1.2) is estimated on the rest of the sample. The estimable model is then:

$$Y_i^T = X_i\beta^T + \mu_i^T \text{ if } D_i = 1 \tag{4.1}$$

$$Y_i^C = X_i\beta^C + \mu_i^C \text{ if } D_i = 0 \tag{4.2}$$

Equivalently, one can follow the more common practice in applied work of estimating a single ("switching") regression for the observed outcome measure on the pooled sample:

$$Y_i = D_iY_i^T + (1 - D_i)Y_i^C = X_i\beta^C + X_i(\beta^T - \beta^C)D_i + \varepsilon_i \quad (i=1,\ldots,n) \tag{5}$$

where the error term has the form:

$$\varepsilon_i = D_i(\mu_i^T - \mu_i^C) + \mu_i^C \tag{6}$$

Impacts are then given by the coefficients on D_i in (5). A special case that is popular in practice is the "common effect" specification in which all except the intercepts in the parameter vectors β^T and β^C are assumed to be invariant to treatment and hence the same in (1.1) and (1.2). (This assumption is rarely made with any obvious justification beyond the fact that one can immediately read off the mean impact from the standard regression output.) Then (5) collapses to a regression of outcomes on participation and the control variables:

$$Y_i = (\beta_0^T - \beta_0^C)D_i + X_i\beta^C + \varepsilon_i \qquad (7)$$

where β_0^T and β_0^C are the intercepts in (1.1) and (1.2).

How can the impact parameters be estimated? Ordinary Least Squares (OLS) applied to (4.1) and (4.2) or (5) will give consistent estimates of the impact parameters in (2) if there is no selection bias in placement conditional on *X*, i.e., that $E(\mu_i^T - \mu_i^C | X_i, D_i = 1) = 0$, or (equivalently) that the conditional mean outcomes do not depend on treatment, $E[Y_i^C | X_i, D_i = 1] = E[Y_i^C | X_i, D_i = 0]$. Then we say that program placement — the assignment of units between (4.1) and (4.2) — is exogenous. (A somewhat more demanding requirement is that outcomes are independent of treatment conditional on *X*. This is variously termed "exogeneity," "unconfoundedness" and "strong ignorability" in the literature.) Then the error term defined by (6) vanishes in expectation given the regressors — assuring that OLS gives consistent estimates under standard conditions. Also, *ATE* and *ATET* become identical. More generally, endogenous program placement will stem from purposive placement of the program, generating a bias, $BIAS \equiv E[Y_i^C | X_i, D_i = 1] - E[Y_i^C | X_i, D_i = 0]$, such that the mean difference in outcomes between the participants and non-participants is given by the identity:

$$E[Y_i^T | X_i, D_i = 1] - E[Y_i^C | X_i, D_i = 0] = ATET + BIAS$$

where $ATET \equiv E[Y_i^T | X_i, D_i = 1] = E[Y_i^C | X_i, D_i = 1]$.

One way to assure exogeneity is to randomize placement, in which case we are dealing with an <u>experimental evaluation</u>, to be considered in detail in the next section. By contrast, in a <u>non-experimental evaluation</u> (also called an "observational study" or "quasi-experimental evaluation") the program is taken to be purposively (non-randomly) placed. This is almost always the case with anti-poverty programs, which are typically targeted on the basis of

characteristics, such as households with many dependents living in poor areas. Additionally, there may be a selection process on the part of the participants, such that some choose not to participate even if they are assigned the program; indeed, this process is essential to a class of "self-targeted" anti-poverty programs (such as workfare schemes discussed later).

So the bulk of this paper will focus on non-experimental methods. These differ in the assumptions they make in identifying impacts and their related data requirements. The methods fall into two main groups, depending on which of two (non-nested) conditional independence assumptions is made. One group assumes exogeneity of program placement or changes in placement, given observable covariates. Methods that draw on some form of this exogeneity assumption are discussed in sections 5-8. Sections 5 and 6 look at "single difference" methods that compare outcome indicators between a sample of participants and one of non-participants, where those samples are chosen purposively to reduce selection bias. Sections 7 and 8 turn to "double or triple difference" methods that exploit data on outcomes in the absence of the program, such as using a "baseline survey" done prior to the intervention. The alternative conditional independence assumption is that there exists an "instrumental variable" that does not alter outcomes conditional on participation (and other covariates of outcomes) but nonetheless does influence participation; this is the method discussed in section 9.

Some evaluators clearly prefer to make one of these conditional independence assumptions over the other. However, there is no *a priori* reason for having a fixed preference in this choice, which should be made on a case-by-case basis, depending on what we know about the program and setting, and what data are available.

3. Generic issues in evaluating anti-poverty programs

This section reviews the generic issues that we will return to often, as the rest of the paper reviews the main methods found in practice. There are essentially two classes of problems that confound efforts to identify impact. The first is selection bias, which can arise from either observables or un-observables. The second is the existence of spillover effects, confounding efforts to attribute a program's impacts to only its direct participants. After examining these issues in turn, the section reviews some generic data issues.

Have we dealt adequately with selection on observables? A common concern in non-experimental evaluations is whether the selection process for the program being evaluated is captured adequately by the control variables X. This concern cannot be separated from the problem of non-random placement conditional on observables. One cannot judge whether exogeneity of placement is a plausible assumption without first establishing whether one has dealt adequately with the observable heterogeneity.

Equations (4) and (5) deal with selection on observables in a rather special way, in that the controls enter in a linear-in-parameters form. This *ad hoc* assumption is rarely justified by anything more than computational convenience (which is rather lame these days). Section 5 will consider formulations of the impact estimation problem under exogeneity that attempt to deal with this source of bias in a more general way.

In non-experimental evaluations of anti-poverty programs is it can be difficult to assure that observables are balanced between the two sets of observations. To see the problem clearly, suppose that placement is determined by a "proxy-means test," as often used for targeting anti-poverty programs in developing countries. This assigns a score to all potential participants as a function of observed characteristics. When strictly applied, the program is assigned if and only if a unit's score is below some critical level, as determined by the budget allocation to the

scheme — the pass-score is non-decreasing in the budget under plausible conditions — and the distribution of the population across the scores. With 100% take-up, there is no interval of the score for which we can observe both participants and non-participants in a sample of any size. This is an example of what is called "failure of common support" in the evaluation literature. The problem is plain enough: how can we infer the counterfactual for participants on the basis of non-participants who do not share the same characteristics, as summarized by their score on the proxy means test? If we want to infer mean impact for those receiving the program then we must have a serious concern about the validity of any comparison group design in this case. Thankfully, in practice, there is invariably some degree of fuzziness in the application of the proxy-means test and there is typically incomplete coverage of those who pass the test. Also, we may not need to know impact for the treatment group as whole. For example, the policy choice may be whether to increase the program's budget allocation (by raising the pass mark in the proxy-means test), in which case we should focus on impacts in a neighborhood that point; section 6 discusses "discontinuity designs" for such cases.

Typically, we will have to truncate the sample of non-participants to assure a valid comparison group; beyond the inefficiency of collecting unnecessary data, this is not a concern. More worrying is that a non-random sub-sample of participants may have to be dropped for lack of sufficiently similar comparators. This points to a trade-off between two sources of bias. On the one hand, there is the need to assure comparability in terms of initial characteristics. On the other hand, this creates a possible sampling bias in inferences about impact, to the extent that we find that we have to drop treatment units to achieve comparability.

Is there a latent selection process? Like most public programs, participation in direct interventions against poverty is almost never random. This is a problem if some of the variables

that jointly influence outcomes and program placement are unobserved to the evaluator. If this is the case then we cannot attribute to the program the observed differences in measured outcomes between units that receive the program and those who do not (conditional on the control variables). The differences in conditional means that we see in the data could just be due to the fact that the program participants were purposely selected by a process that we do not fully observe. When program take-up is a matter of individual choice, there must be a reasonable presumption that selection into the program depends on the gains from participation.

In terms of the classic formulation of the evaluation problem above, suppose that participants have latent attributes that yield higher outcomes than non-participants (at given *X*). Then the error terms in the equation for participants (4.1) will be centered to the right relative to those for non-participants (4.2). The error term in (5) cannot vanish in expectation and OLS will give biased and inconsistent estimates. Recall that the bias in the conditional mean impact estimator is the difference in the counterfactual means, $E[Y_i^C | X_i, D_i = 1] - E[Y_i^C | X_i, D_i = 0]$. Again it should be emphasized that the extent of concern about this form of selection bias in practice cannot be separated from the prior question above as to how well we have controlled for observable heterogeneity.

There are examples pointing to bias in non-experimental impact estimates in specific cases. A widely-cited study by Lalonde (1986) found large biases in non-experimental methods when compared to a randomized evaluation of a training program. Different non-experimental methods also gave quite different results, consistent with the presence of a latent selection process. Similarly, Glewwe et al. (2004) find that non-experimental methods give a larger estimated impact of "flip charts" on the test scores of Kenyan school children than implied by an experiment; they argue that biases in their non-experimental methods account for the difference.

Of course, one cannot reject non-experimental methods in other applications on the basis of such studies; arguably the lesson is that better data and methods are needed, informed by past knowledge of how such programs work. In a critique of the Lalonde study, Heckman and Smith (1995) point out that (amongst other things) the data used contained too little information relevant to eligibility for the program studied and that the methods used had limited power for addressing selection bias and did not include adequate specification tests.[5]

Nonetheless, some of the most popular non-experimental methods — such as simple cross-sectional comparisons of participants and non-participants or reflexive comparisons of participant's outcomes over time — can give severely biased results when the available data provide inadequate controls for heterogeneity. Using a different approach to testing non-experimental methods, van de Walle (2002) gives an example for rural road evaluation in which a naïve comparison of the incomes of villages that have a rural road with those that do not indicates large income gains when in fact there are none. Van de Walle used simulation methods in which the data were constructed from a model in which the true benefits were known with certainty and the roads were placed in part as a function of the average incomes of different villages. Only a seemingly small weight on village income in determining road placement was enough to severely bias the mean impact estimate. In practice, controls for observable correlates of income would almost certainly reduce the bias suggested by van de Walle's simulations.

Are there spillover effects on non-participants? Eliminating selection bias does not assure that impacts can be identified. The classic formulation of the evaluation problem outlined at the beginning of this section assumes that the treatment of unit *i* can only affect outcomes for

[5] Also see the discussion in Heckman et al., (1999).

that unit.[6] Then we can observe a group of non-participants who are in no way affected by the program in question. We may do this at the time the program is in place (giving a "single-difference" design, as discussed further in sections 4 and 5) or we may do it for participants before the program is in place (giving a "reflexive comparison" discussed in section 7); when we do both we have a "double difference" as discussed in section 7. Under certain conditions, we can also infer impacts by comparing those who leave a program with those who stay (section 8). However, all these cases assume that one can observe the non-participation state in a way that is uncontaminated by the program in question.

That can be a problematic assumption for anti-poverty programs. For example, suppose that we are evaluating a workfare program whereby the government commits to give work to anyone who wants it at a stipulated wage rate; this was the aim of the famous Employment Guarantee Scheme (EGS) in the state of Maharashtra in India and at the time of writing the Government of India is planning to expand the idea to the country as a whole. The attractions of an EGS as a safety net stem from the fact that access to the scheme is universal (anyone who wants help can get it) but that all participants must work to obtain benefits and at a wage rate that is considered low in the specific context. The universality of access means that the scheme can provide effective insurance against risk. The work requirement at a low wage rate is taken by proponents to imply that the scheme will be self-targeting to the income poor.

This can be thought of as an assigned program, in that there are well-defined "participants" and "non-participants." And at first glance it might seem appropriate to collect survey data on both groups and compare outcome indicators between the two, as a means of identifying impact (possibly after cleaning out any observable heterogeneity).

[6] This is sometimes called the "stable unit treatment assumption" in the evaluation literature; see for example, Angrist et al. (1996).

However, this classic evaluation design could give a severely biased result. The gains from such a program must spillover into the private labor market and ignoring the (likely positive) spillover effects will entail an underestimation of the benefits. Indeed, if the employment guarantee is effective then the scheme will establish a firm lower bound to the entire wage distribution, for no able-bodied worker would accept non-EGS work at any wage rate below the EGS wage. So even if one picks the observationally perfect comparison group of non-participants, one will conclude that the scheme has no impact, since wages will be the same for participants and non-participants. But that would entirely miss the impact.

Spillover effects can also arise from the behavior of governments. Whether the resources transferred to participants actually financed the identified project is often unclear. To some degree, all external aid is fungible. Yes, it could be verified in supervision that the proposed sub-project was actually completed. But one cannot rule out the possibility that it would have been done otherwise. Participants and local leaders would naturally have put forward the best development option they saw, even if it was something they planned to do anyway with the resources already available. Then there is some other (infra-marginal) expenditure that was really being financed by the aid. Similarly, there is no way of ruling out the possibility that non-project villages benefited by a re-assignment of public spending by local authorities, thus lowering the measured impact of program participation.

This problem is studied by van de Walle and Cratty (2005) in the context of a rural-roads project in Vietnam. The authors find no impact on comparing kilometers of roads rehabilitated by the (aid-financed) project with a comparison group of non-participating communes. This is interpreted as reflecting in part the fungibility of aid, though in this example it turns out that

selection bias is also at work, such that the degree of fungibility is overstated unless one controls adequately for the purposive geographic targeting of the development project.

In the above examples, spillover effects lead one to under-estimate impact. However, the bias could also go in the other direction. Suppose, for example, that the domestic government is keen to show impact of an aid-financed anti-poverty project and so it "tops up" the external (financial or other) resources targeted to the project locations. The difference between outcomes for participants and the comparison group will then overstate the impact of the external aid.

What data are required? As is clear from the above discussion, concerns about inadequate or imperfect data lie at the heart of the evaluation problem. When embarking on any impact evaluation, it is important to first know a lot about the administrative/institutional details of the program; that information typically comes from the program administration. For non-experimental evaluations, such information is key to designing a survey that collects the right data to control for the selection process. Knowledge of the program's context and design features can also help in dealing with selection on unobservables, since it can sometimes generate plausible identifying restrictions, as discussed further in section 9.

Non-experimental evaluations of anti-poverty programs can be very demanding in their data requirements. The precise sources of data used in an evaluation can embrace both informal, unstructured, interviews with participants in the program as well as quantitative data from representative samples. However, it is extremely difficult to ask counter-factual questions in interviews or focus groups; try asking someone who is currently participating in a public program: "what would you be doing now if this program did not exist?" Talking to program participants can be valuable, but it is unlikely to provide a credible evaluation on its own. One also needs data on the outcome indicators and relevant explanatory variables.

The data on outcomes and their determinants, including program participation, typically come from surveys. The observation unit could be the individual, household, geographic area or facility (school or health clinic) depending on the type of program. Survey data can often be supplemented with useful other data on the program (such as from the project monitoring data base) or setting (such as from geographic data bases).

A potentially serious concern is the comparability/consistency of different data sources, particularly those used for the observations on participants and non-participants. Differences in the design of the survey instruments can entail non-negligible differences in the outcome measures, thus confounding impact assessments. Heckman et al. (1999, Section 5.33) show how differences in data sources and data processing assumptions can make large differences in the results obtained for evaluating US training programs.[7]

4. Single difference comparisons with randomized assignment

A social experiment aims to randomize program placement, such that all units (within some well-defined set) have the same chance *ex-ante* of receiving the program. All (observed or unobserved) attributes prior to the intervention are then identically distributed between units receiving the program and those not. By implication, the observed *ex-post* difference in mean outcomes between the two groups is attributable to the program. In terms of the formulation of the evaluation problem in the previous section, randomization guarantees that there is no sample selection bias in estimating (4.1) and (4.2) or (equivalently) that the error term in equation (5) is orthogonal to the regressors. The non-participants are then a valid control group for identifying

[7] Also see the example in Diaz and Handa (2004).

the counterfactual,[8] and ATE is consistently estimated (nonparametrically) by the difference between the sample means of Y_i^T and Y_i^C (including for sub-samples with given values of X_i)

Examples: A number of evaluations of active labor market programs in the US have used social experiments, often applied to a pilot scheme. Much has been learnt about welfare policy reform from such trials (Moffitt, 2003). Two examples are the Job Training Partnership Act (JTPA) (see, for example, Heckman et al., 1997b), and the US National Supported Work Demonstration (studied by Lalonde, 1986, and Dehejia and Wahba, 1999). For wage subsidy programs, randomized evaluations have been done by Burtless (1985), Woodbury and Spiegelman (1987) and Dubin and Rivers (1993) — all for targeted wage subsidies in the US.

There have been a number of social experiments for anti-poverty programs in developing countries. A well-known example is Mexico's *PROGRESA* program, which provided cash transfers targeted to poor families conditional on their children attending school and obtaining health care and nutrition supplementation. The (considerable) influence that this program has had in the development community clearly stems in no small measure from the substantial, and public, effort that went into its evaluation.[9] Half of the original 500 communities chosen to receive *PROGRESA* were randomly retained as a control group for an initial period during which the rest received the program. Public access to the evaluation data has facilitated a number of valuable studies, indicating significant gains to health (Gertler, 2004) schooling (Schultz, 2004; Behrman et al., 2002) and food consumption (Hoddinott and Skoufias, 2004).

In another example in a developing country, Newman et al. (2002) were able to randomize eligibility to a World Bank supported social fund within one region of Bolivia. The

[8] The term "control group" is often confined to social experiments, with the term "comparison group" used in non-experimental evaluations.

[9] The evaluation was designed and implemented by the International Food Policy Research Institute (IFPRI).

fund-supported investments in education were found to have had significant impacts on school infrastructure but not education outcomes within the evaluation period.

Randomization was also used by Angrist et al. (2002) to evaluate a program in Colombia that allocated vouchers for schooling by a lottery. Three years later, the lottery winners had significantly better school attainments, with lower grade repition and higher test scores.

Another example is the *Proempleo* experiment in Argentina (Galasso et al., 2004). This was a randomized evaluation of a pilot wage subsidy and training program for assisting workfare participants in Argentina to find regular, private-sector jobs. Eighteen months later, recipients of the voucher for a wage subsidy had a higher probability of employment than the control group. (We will return later in this paper to examine some lessons from this evaluation more closely.)

While the World Bank has supported a number of social experiments (including all of the examples for developing countries above), that is not so of the Bank's Operations Evaluation Department (the semi-independent unit for the *ex-post* evaluation of its own lending operations). In the 78 evaluations by OED surveyed by Kapoor (2002), none used randomization.[10] OED has been criticized by proponents of social experiments for not doing more of them; see, for example, Cook (2001) and Duflo and Kremer (2005) (both published in OED volumes).

Issues with social experiments: While randomized designs can claim to be the theoretical ideal for identifying mean impact, problems arise in practice.[11] Ethical objections and political sensitivities stem from the perception that social experiments treat people like "guinea pigs," deliberately denying the program to those who need it (to form the control group) in favor of some who don't. In the case of anti-poverty programs, one can end up assessing impacts for

[10] Kapoor classifies one evaluation as using a control group to establish the counterfactual, namely for a rural water-supply project in Paraguay. However, from Kapoor's description, this was not a genuine social experiment which would have required that the intervention was randomly assigned *ex ante*.

[11] On social experiments see Heckman and Smith (1995), Burtless (1995) and Moffitt (2003).

types of people for whom the program is not intended and/or denying the program to poor people who need it — in both cases running counter to the very aims of the program.

One defense of social experiments points out that it is often the case that there are too few resources to go around. There are poor people who can't get the program. Deliberately excluding some people from a program for evaluation purposes — with bearing on prospects of fighting poverty more effectively — can hardly be a more serious ethical issue than the routine exclusions going on all the time because there are too few resources to go around. But two wrongs don't make a right. A possibly more persuasive defense of experiments in resource-poor settings says that, given that the program cannot cover all the eligible poor, the fairest solution is to assign it randomly, so that everyone has an equal chance of getting the limited resources available.[12] However, that argument runs into a problem: it is hard to appreciate the "fairness" of an anti-poverty program that deliberately ignores information that is readily available on differences in the extent of deprivation.

Other concerns have been raised about social experiments. Internal validity can be questionable when there is selective compliance with the theoretical randomized assignment. People are (typically) free agents. They do not have to comply with the evaluator's assignment. The fact that people can select out of the randomized assignment goes some way toward alleviating the aforementioned ethical concerns about social experiments. People who know they do not need the program will presumably decline participation. But selective compliance clearly invalidates inferences about impact. The extent of this problem depends of course on the specific program; selective compliance is more likely for a training program (say) than a cash transfer program. Sections 7 and 9 will return to this issue and discuss how non-experimental

[12] From the description of the Newman et al. (2003) study it appears that this is how randomization was defended in their case.

methods can help address the problem, and how partially randomized designs can help identify impacts using non-experimental methods.

Spillover effects are an important source of internal validity concerns about evaluations in practice, including social experiments. It is recognized in the literature that the choice of observational units should reflect likely spillover effects. For example, Miguel and Kremer (2004) study the evaluation of treatments for intestinal worms in children and argue that a randomized design in which some children are treated and some are retained as controls would seriously underestimate the gains from treatment by ignoring the externalities between treated and "control" children. The randomized design for the authors' experiment avoided this problem by using mass treatment at the school level instead of individual treatment (using control schools at sufficient distance from treatment schools).

The behavioral responses of third parties can also generate spillover effects. Recall the example in section 3 of how a higher level of government might adjust its own spending, counteracting the assignment (randomized or not). This may well be an even bigger problem for randomized evaluations. The higher level of government may not feel the need to compensate units that did not get the program when this was based on credible and observable factors that are agreed to be relevant. On the other hand, the authorities may feel obliged to compensate for the "bad luck" of units being assigned randomly to a control group. Randomization can induce spillovers that do not happen with selection on observables.

This is an instance of a more general and fundamental problem with randomized designs for anti-poverty programs, namely that the very process of randomization can alter the way a program works in practice. There may well be systematic differences between the characteristics of people normally attracted to a program and those randomly assigned the program from the

same population. (This is sometimes called "randomization bias.") Heckman and Smith (1995) discuss an example from the evaluation of the JTPA, whereby substantial changes in the program's recruiting procedures were required to form the control group. The evaluated pilot program is not then the same as the program that gets implemented — casting doubt on the validity of the inferences drawn from the evaluation.

The JTPA illustrates a further problem in practice, namely that institutional or political factors may delay the randomized assignment. This promotes selective attrition and adds to the cost, as more is spent on applicants who end up in the control group (Heckman and Smith, 1995).

A further critique of social experiments argues that they have not been informative about the economic and social processes influencing outcomes; see, for example, Heckman and Smith (1995). Even with randomized assignment we only know mean outcomes for the counterfactual, so we cannot infer the joint distribution of outcomes as would be required to say something about (for example) the proportion of gainers versus losers amongst those receiving a program.

The strength of experiments is in dealing with the problem of purposive placement based on unobserved factors; their weakness is in throwing light on the determinants of impacts and other policy-relevant parameters, though this weakness is shared by many non-experimental methods in practice. Section 10 returns to this issue.

What can be done to assess impact when a program was not randomly placed? The rest of this paper provides a critical overview of the main non-experimental methods.

5. Single difference matched comparisons

As section 3 emphasized, selection bias is to be expected in comparing a random sample of participants with a random sample of non-participants. There must be a general presumption that such comparisons misinform policy. How much so is an empirical question. On *a priori*

grounds it is worrying that many non-experimental evaluations of anti-poverty programs in practice provide too little information to properly assess whether the "comparison group" of non-participants is similar to the participants in the absence of the intervention.[13]

Some of the selection bias in single difference comparisons can be cleaned out by matching the two groups on observables. In trying to find a comparison group for assessing the counterfactual it is natural to search for non-participants with similar pre-intervention characteristics to the participants. However, there are potentially many characteristics one might look for to match on; how should they be weighted in choosing the comparison group?

Propensity-Score Matching: Rosenbaum and Rubin (1983) offer a solution to this problem.[14] The method selects comparators according to their predicted probabilities of participation (called their "propensity scores"). The key to PSM is understanding and modeling how the program is assigned. Participants are matched to non-participants on the basis of the propensity score, $P(Z_i) = E(D_i|Z_i)$ ($0< P(Z_i)<1$), where Z_i is a vector of pre-exposure control variables (which can include pre-treatment values of the outcome indicator).[15] PSM uses $P(Z_i)$ (or a monotone function of $P(Z_i)$) to select comparison units. It is known from Rosenbaum and Rubin (1983) that if (i) the D_i's are independent over all i, and (ii) outcomes are independent of participation given Z_i, then outcomes are also independent of participation given $P(Z_i)$.[16] Assumption (ii) is essentially a more general version of the exogeneity-of-placement assumption discussed in sections 2 and 3. Under these conditions, exact matching on $P(Z_i)$ eliminates

[13] See, for example, Kapoor's (2002) comments on OED's evaluations.

[14] The Rosenbaum-Rubin paper built on a series of papers in the statistical literature by Rubin and others. For a thorough recent review of the theory of propensity score matching see Imbens (2004).

[15] The present discussion is confined to the standard case of binary treatment. In generalizing to the case of multi-valued or continuous treatments one defines the generalized propensity score given by the conditional probability of a specific level of treatment (Imbens, 2000; also see Hirano and Imbens, 2004).

[16] For a clear recent statement and proof of the Rosenbaum-Rubin theorem see Imbens (2004).

selection bias.[17] As in a social experiment, ATE is non-parametrically identified by the difference between the sample means of Y_i^T and Y_i^C for the matched comparison group.

Intuitively, what PSM is doing is creating the observational analogue of a social experiment in which everyone has the same probability of participation. The difference is that in PSM it is the conditional probability (conditional on *Z*) that is uniform between participants and matched comparators, while randomization assures that the participant and comparison groups are identical in terms of the distribution of all characteristics whether observed or not. PSM essentially assumes away the problem of endogenous placement, leaving only the need to balance the conditional probability, i.e., the propensity score. An implication of this difference is that (unlike randomized evaluation) the impact estimates obtained by PSM must always depend on the variables used for matching and the quantity and quality of data.

The control variables in Z_i may well differ from the covariates of outcomes (the vector X_i in section 2); this distinction plays an important role in the impact estimates discussed in section 9 . But what should be included in Z_i? The theory of PSM, as developed by Rubin and colleagues, does not say much about that question, yet the choice must matter to the results obtained. Intuitively, one expects that the choice of variables should be based on theory and/or facts about the program and setting, as relevant to understanding the economic, social or political factors influencing program assignment. Qualitative work can help here; for example, the specification choices made in Jalan and Ravallion (2003b) reflected interviews with participants in Argentina's *Trabajar* program (a combination of workfare and social fund) and local program administrators. Similarly Godtland et al. (2004) validated their choice of covariates for participation in an agricultural extension program in Peru by interviews with farmers. Clearly if

[17] On the efficiency of PSM relative to covariate matching see Angist and Hahn (2004).

the available data do not include important determinants of participation then the presence of these unobserved characteristics will mean that PSM will not be able to reproduce (to a reasonable approximation) the results of a social experiment.

Common practice is to use the predicted values from a standard logit or probit regression to estimate the propensity score for each observation in the participant and the non-participant samples (though non-parametric binary response models can also be used; see Heckman et al., 1997). The participation regression is of interest in its own right as it can provide useful insights into the targeting performance of an anti-poverty program (see, for example, the discussion in Jalan and Ravallion, 2003b). The comparison group is then formed by picking the "nearest neighbor" for each participant, defined as the non-participant that minimizes $\left|\hat{P}(Z_i) - \hat{P}(Z_j)\right|$ as long as this does not exceed some caliper bound. Given measurement errors, more robust estimates are likely by taking the mean of the nearest (say) five neighbors, though this does not necessarily reduce bias.[18] It is a good idea to also to test for systematic differences in the covariates between the treatment and comparison groups constructed by PSM; Smith and Todd (2005a) describe a useful "balancing test" for this purpose.

More generally, the estimator for mean impact is $\sum_{j=1}^{NT}(Y_j^T - \sum_{i=1}^{NC} W_{ij} Y_{ij}^C)$ where NT is the number receiving the program, NC is the number of non-participant households and the W_{ij}'s are the weights applied in calculating the average outcome of the matched non-participants. (Sampling weights may also be needed, depending on the survey design.) There are several weighting schemes that have been used, ranging from nearest-neighbor weights to non-parametric weights based on kernel functions of the differences in scores whereby all the

[18] Rubin and Thomas (2000) use simulations to compare the bias in using the nearest five neighbors to just the nearest neighbor; no clear pattern emerges.

comparison units are used in forming the counterfactual for each participating unit, but with a weight that reaches its maximum for the nearest neighbor but declines as the absolute difference in propensity scores increases; Heckman et al. (1997b) discuss this weighting scheme.[19]

Mean impacts can be calculated conditional on observed characteristics. For anti-poverty programs one is interested in comparing the conditional mean impact across different pre-intervention incomes. For each sampled participant, one estimates the income gain from the program by comparing that participant's income with the income for matched non-participants. Subtracting the estimated gain from observed post-intervention income, it is then possible to estimate where each participant would have been in the distribution of income without the program. Thus one can construct the empirical and counter-factual cumulative distribution functions or their empirical integrals, and test for dominance over a relevant range of poverty lines and measures. Further discussion of how the results of an impact assessment by PSM can be used to assess impacts on poverty measures robustly to the choice of those measures and the poverty line can be found in Ravallion (2003), with a detailed worked example for an actual anti-poverty program (Argentina's *Trabajar* program).

How does PSM differ from other methods? Unlike a social experiment (at least in its pure form), PSM will naturally face concerns about selection bias whenever one can postulate the existence of a latent variable that jointly influences placement and outcomes (thus invalidating the key conditional independence assumption made by PSM). This must be judged for the application in hand.

Nor can it be assumed that eliminating selection bias based on observables will reduce the aggregate bias; that will only be the case if the two sources of bias — that associated with

[19] Frölich (2004) compares the finite-sample properties of various estimators and finds that a local linear ridge regression method is more efficient and robust than alternatives.

observables and that due to unobserved factors — go in the same direction, which cannot be assured on *a priori* grounds. If the selection bias based on unobservables counteracts that based on observables then eliminating only the latter bias will increase aggregate bias. While this is possible in theory, I do not know of any plausible example from practice.

A natural comparison is between PSM and an OLS regression of the outcome indicators on dummy variables for program placement, allowing for the observable covariates entering as linear controls (as in equations 4 and 6). OLS requires essentially the same conditional independence (exogeneity) assumption as PSM, but also imposes arbitrary functional form assumptions concerning the treatment effects and the control variables. By contrast, PSM (in common with experimental methods) does not require a parametric model linking outcomes to program participation. Thus PSM allows estimation of mean impacts without arbitrary assumptions about functional forms and error distributions. This can also facilitate testing for the presence of potentially complex interaction effects. For example, Jalan and Ravallion (2002a) use PSM to study how the interaction effects between income and education influence the child-health gains from access to piped water in rural India. The authors find a complex pattern of interaction effects; for example, poverty attenuates the child-health gains from piped water, but less so the higher the level of maternal education.

A variation on the standard OLS estimator is to add the estimated propensity score $\hat{P}(Z)$ to a regression of the outcome variable on the treatment dummy variable, *D*. Under assumptions of PSM this will clearly eliminate any omitted variable bias in having excluded *Z* from that regression, given that *Z* is independent of treatment given $P(Z)$.[20] Yet another variation is to include an interaction effect between $\hat{P}(Z_i)$ and D_i. These variations on the standard OLS

[20] This provides a further intuition as to how PSM works; see the discussion in Imbens (2004).

estimate of the common effects model (7) may or may not be an improvement or more convenient (certainly the OLS standard errors will need correction given that $\hat{P}(Z_i)$ requires previously estimated parameters), and they do not have the non-parametric flexibility of PSM.

PSM also differs from standard regression methods with respect to the sample. In PSM one confines attention to matched sub-samples; unmatched comparison units are dropped. Some of the non-participants may be excluded because they have a score that is outside the range found for the participant sample. The range of scores estimated for the participants should correspond closely to that for the retained sub-sample of non-participants. In other words, matching is confined to the region of common support, as illustrated in Figure 1. One may also want to restrict potential matches in other ways, depending on the setting. For example, one may want to only allow matches within the same geographic area to help assure that the comparison units come from the same economic environment. By contrast, the regression methods commonly found in the literature use the full sample. The simulations in Rubin and Thomas (2000) indicate that impact estimates based on full (unmatched) samples are generally more biased, and less robust to miss-specification of the regression function, than those based on matched samples.

A further difference relates to the choice of control variables. In the standard regression method one looks for predictors of outcomes, and preference is given to variables that one can argue are exogenous to outcomes. In PSM one is looking instead for covariates of participation, possibly including variables that are poor predictors of outcomes. Indeed, analytic results and simulations indicate that variables with weak predictive ability for outcomes can still help reduce bias in estimating causal effects using PSM (Rubin and Thomas, 2000).

It is an empirical question as to how much difference it would make to mean-impact estimates by using PSM rather than OLS. Comparative methodological studies have been rare.

In one exception, Godtland et al. (2004) use both an outcome regression and PSM for assessing the impacts of field schools on farmers' knowledge of good practices for pest management in potato cultivation. They report that their results were robust to changing the method used.

How well does PSM perform? Returning to the same data set used by the Lalonde (1986) study (described in section 3), an influential paper by Dehejia and Wahba (1999) found that propensity-score matching achieved a good approximation — much better than the non-experimental methods studied by Lalonde. However, the robustness of the Dehejia-Wahba findings to sample selection and the specification chosen for calculating the propensity scores has been questioned by Smith and Todd (2005a), who argue that PSM does not solve the selection problem in the program studied by Lalonde.[21]

Similar attempts to test PSM against randomized evaluations have shown mixed results. Agodini and Dynarski (2004) find no consistent evidence that PSM can replicate experimental results from evaluations of school dropout programs in the US. Using the *PROGRESA* data base, Diaz and Handa (2004) find that PSM performs well when the same survey instrument is used for measuring outcomes for the treatment and comparison groups, but not when the survey instruments differ. The importance of using the same survey instrument in PSM is also emphasized by Heckman et al. (1997a, 1998) in the context of their evaluation of a US training program. The latter study also points to the importance of both participants and non-participants coming from the same local labor markets, and of being able to control for employment history.

In summary: PSM is an important addition to the menu of tools available for counterfactual analysis. The emphasis on understanding the assignment mechanism — the observables determinants of program placement — is welcome, and is certainly a more

[21] Dehejia (2005) replies to Smith and Todd (2005a), who offer a rejoinder in Smith and Todd (2005b). Also see Smith and Todd (2001).

convincing starting point than the presumption that one has found a "natural experiment," which seems unlikely in general. Single-difference comparisons using PSM also have the advantage in evaluating anti-poverty programs in developing countries that the method does not require either randomization or baseline (pre-intervention) data. While this is a huge advantage, it comes at a cost. To accept the exogeneity assumption one must be confident that one has controlled for the factors that jointly influence program placement and outcomes. So PSM requires good data; Section 8 will give an example of how far wrong the method can go if data are inadequate.

6. Exploiting program design in single difference comparisons

Single difference non-experimental estimators can sometimes usefully exploit features of program design for identification. Discontinuities generated by program eligibility criteria can help identify impacts in a neighborhood of the cut-off points for eligibility. Delays in the implementation of a program can also facilitate forming comparison groups, which can also help pick up some sources of latent heterogeneity.

Discontinuity designs: The idea here is to infer impacts from the differences in outcomes under treatment between units on either side of a critical cut-off point used for determining program eligibility. Examples include a proxy-means test that sets a maximum score for eligibility (as discussed in section 3) and programs that confine eligibility to within certain geographic boundaries.

To see more clearly what this method involves, let M_i denote the score received by unit *i* in a proxy-means test (say) and let *m* denote the cut-off point for eligibility, such that $D_i = 1$ for $M_i \leq m$ and $D_i = 0$ otherwise. Impact is then $E(Y_i^T \left| M_i = m - \varepsilon \right.) - E(Y_i^C \left| M_i = m + \varepsilon \right.)$ for some arbitrarily small $\varepsilon > 0$. In practice, there is inevitably a degree of fuzziness in the application of

eligibility tests. So instead of assuming strict enforcement and compliance, one can follow Hahn et al. (2001) in postulating a probability of program participation, $P(M_i) = E(D_i|M_i)$, which is an increasing function of M_i with a discontinuity at *m*. The essential idea remains the same, in that impacts are measured by the difference in mean outcomes in a neighborhood of *m*.

The key identifying assumption is that there is no discontinuity in counterfactual outcomes at *m*.[22] However, the fact that a program has more-or-less strict eligibility rules does not (of itself) mean that continuity is a plausible assumption. For example, the geographic boundaries for program eligibility will often coincide with local political jurisdictions, entailing current or past geographic differences in (say) local fiscal policies and institutions that cloud identification. The plausibility of the continuity assumption for counterfactual outcomes must be judged in each application. A baseline survey can help clean out any pre-intervention differences in outcomes either side of the discontinuity, in which case one is combining the discontinuity design with the double difference method discussed in the next section; for an example see Jacob and Lefgren (2004).

In a test of how well discontinuity designs perform, Buddelmeyer and Skoufias (2004) use the cut-offs in *PROGRESA*'s eligibility rules to measure impacts and compare the results to those obtained by exploiting the program's randomized design. The authors find that the discontinuity design gives good approximations for almost all outcome indicators.

How does this method compare to PSM? A discontinuity design gives mean impact for a selected sample of the participants, while PSM aims to give mean impact for the treatment group as a whole. However, the aforementioned common support problem that is sometimes generated by eligibility criteria can mean that PSM is also confined to a selected sub-sample; the question

[22] Hahn et al. (2001) provide a formal analysis of identification and estimation of impacts for discontinuity designs under this assumption.

is then whether that is an interesting sub-sample. The truncation of treatment group samples under PSM will most likely tend to exclude those with the highest probability of participating (for which non-participating comparators are hardest to find), while discontinuity designs will tend to only include those with the lowest probability. The latter sub-sample can, nonetheless, be relevant for deciding about program expansion; section 10 returns to this point.

Although mean impacts are non-parametrically identified for discontinuity designs, the literature in economics has more often used an alternative parametric method in which the discontinuity in the eligibility criterion is used as an instrumental variable for program placement; we will we will return to give examples in section 9.

Pipeline comparisons: There is often a delay between successful application to a program and receiving it. Those who have applied but not yet received the program, have been used as a comparison group.[23] PROGRESA is an example; one third of eligible participants did not receive the program for 18 months, during which they formed the control group.

In the case of PRORESA, the pipeline comparison was randomized. Non-experimental pipeline comparisons have also been used in developing countries. An example can be found in Chase (2002) who used communities who had applied for a social fund (in Armenia) as the source of the comparison group in estimating the fund's impacts on communities that received its support. In another example, Galasso and Ravallion (2004) evaluated a large social protection program in Argentina, namely the Government's *Plan Jefes y Jefas*, which was the main social policy response to the severe economic crisis of 2002. To form a comparison group for participants they used those individuals who had successfully applied for the program, but had not yet received it, as the comparison group. Notice that this method does to some extent address

[23] This is sometimes called "pipeline matching" though this term is less than ideal given that no matching is actually done.

the problem of latent heterogeneity in other single-difference estimators, such as PSM; the successful applicants will tend to have similar unobserved characteristics, whether they have yet received the program or not.

The key assumption here is that the timing of treatment is random given application. In practice, one must anticipate a potential bias arising from selective treatment amongst the applicants or behavioral responses by applicants awaiting treatment. This is a greater concern in some settings than others. For example, Galasso and Ravallion argued that it was not a serious concern in their case given that they assessed the program during a period of rapid scaling up, during the 2002 crisis in Argentina. The authors also tested for observable differences between the two sub-sets of applicants, and found that observables (including idiosyncratic income shocks during the crisis) were well balanced between the two groups, alleviating concerns about bias. Using longitudinal observations also helped; we return to this method in section 7.

When feasible, pipeline comparisons offer a single-difference impact estimator that is likely to be more robust to latent heterogeneity. The results should, however, be tested for selection bias based on observables — possibly with PSM brought in to clean out the observable heterogeneity in pipeline comparisons.

7. Double difference methods

A popular approach to addressing concerns about the exogeneity assumption in single-difference cross-sectional comparisons is the double difference (or "difference-in-difference") (DD) method. This compares participant and comparison groups in terms of outcome changes over time relative to the outcomes observed for a pre-intervention baseline. In essence, the method allows for endogeneity in the initial placement of the program, but treats the changes in placement as exogenous to the changes in outcomes.

The DD method compares samples of participants and non-participants before and after the intervention. Typically one has a baseline survey before the intervention, covering both non-participants and participants. Often one does not know who will participate, and must make an informed guess in designing the sampling for the baseline survey. Knowledge of the program design and setting can help. One then does one or more follow-up surveys. These should be highly comparable to the baseline surveys (in terms of the questionnaire, the interviewing, etc). Finally one calculates the mean difference between the "after" and "before" values of the outcome indicator for each of the treatment and comparison groups. The difference between these two mean differences is the estimate of the impact of the program.

To see what is involved in more formal terms, suppose that we have collected data on an outcome measure *Y* for a set Ψ of participants and a comparison group. We can write the outcome measure, Y_{it}, for the *i*'th treatment unit ($D_i = 1$) observed at two dates *t*=1,2 as $(Y_{it} | D_i = 1) = Y_{it}^C + G_{it} + \varepsilon_{it}$ where Y_{it}^C is the counter-factual outcome measure for treatment unit *i* if the program had not existed, G_{it} is the gain attributable to the program and ε_{it} is a zero-mean innovation error term uncorrelated with program participation, to allow for measurement error in Y_{it}. An indicator of the counter-factual is available from a comparison group and is given by $\hat{Y}_{it}^C$. This may be a noisy indicator due to selection bias.

The two key assumptions of a DD estimator are: (i) the selection bias is separable and time invariant, and so it is swept away by taking differences over time, and (ii) period 1 outcomes are not contaminated by the expectation of the program's future placement, i.e., $G_{i1} = 0$. Under these assumptions, on taking the expectation over all participants, the DD

estimator of mean impact is the single-difference difference impact estimate for the second period *less* the single difference in the baseline:

$$DD = E[(Y_{i2}^{T} - \hat{Y}_{i2}^{C}) - (Y_{i1}^{T} - \hat{Y}_{i1}^{C}) \big| D_i = 1, i \in \Psi] = E(G_{i2} \big| D_i = 1, i \in \Psi) \qquad (8)$$

Equivalently, DD is the outcome gain observed over time for the treatment group less that for the comparison group. This can be readily generalized to multiple time periods and estimated by DD by the common effect regression of Y_{it} on the (individual and date-specific) participation dummy variable D_{it}, with individual and time fixed effects.[24]

Notice that when mean outcomes for the comparison group are time-invariant ($E[\hat{Y}_{i2}^{C} - \hat{Y}_{i1}^{C} \big| D_i = 1, i \in \Psi] = 0$), equation (8) collapses to a reflexive comparison in which one only monitors outcomes for treatment units. Unchanging mean outcomes for the counterfactual would appear to be an implausible assumption in most applications. However, with enough observations over time, methods of testing for structural breaks in the times series of outcomes for participants can offer some hope of identifying impacts; see for example Piehl et al. (2003).

Examples: Duflo (2001) estimated the impact on schooling and earnings in Indonesia of building schools. A key feature of the assignment mechanism was known, namely that more schools were built in locations with low enrolment rates. Also, the age cohorts that participated in the program could be easily identified. The fact that the gains in schooling attainments of the first cohorts exposed to the program were greater in areas that received more schools was taken to indicate that building schools promoted better education. Frankenberg et al. (2005) use a similar method to assess the impacts of providing basic health care serives through midwives on children's nutritional status (height-for-age), also in Indonesia.

[24] As is well-known in econometrics, when the error term is serially correlated one must take account of this in calculating the standard errors of the DD estimator; Bertrand et al., 2004, demonstrate the possibility for large biases in the uncorrected (OLS) standard errors for DD estimators.

In another example, Galiani et al. (2005) used a DD design to estimate the impact of the privatization of water services on child mortality in Argentina. The authors exploited the joint geographic (across municipalities) and inter-temporal variation in both child mortality and ownership of water services to identify impacts. Their results suggest that privatization of water services reduced child mortality.

A DD design can also be used to address possible biases in a social experiment, whereby there is some form of selective compliance or other distortion to the randomized assignment (as discussed in section 3). An example can be found in Thomas et al. (2003) who randomized assignment of iron-supplementation pills in Indonesia, with a randomize-out group receiving a placebo. By also collecting pre-intervention baseline data on both groups, the authors were able to address concerns about compliance bias.

While the classic DD design tracks the differences over time between participants and non-participants, that is not the only possibility. Jacoby (2002) used a DD design to test whether intra-household resource allocation shifted in response to a school-feeding program, to neutralize the latter's effect on child nutrition. Some schools had the feeding program and some did not, and some children attended school and some did not. The author's DD estimate of impact was then the difference between the mean food-energy intake of children who attended a school (on the previous day) that had a feeding program and the mean for those who did not attend such schools, less the corresponding difference between attending and non-attending children found in schools that did not have the program.

Another example can be found in Pitt and Khandker (1998) who assessed the impact of participation in Bangladesh's Grameen Bank (GB) on various indicators relevant to current and future living standards. GB credit is targeted to landless households in poor villages. Some of

their sampled villages were not eligible for the program and within the eligible villages, some households were not eligible, namely those with land (though it is not clear how well this was enforced). The authors implicitly use an unusual DD design to estimate impact.[25] Naturally, the returns to having land are higher in villages that do not have access to GB credit (given that access to GB raises the returns to being landless). Comparing the returns to having land between two otherwise identical sets of villages — one eligible for GB and one not — will thus reveal the impact of GB credit. So the Pitt-Khandker estimate of the impact of GB is actually the impact on the returns to land of taking away village-level access to the GB.[26] By interpretation, the "pre-intervention baseline" in the Pitt-Khandker study is provided by the villages that have the GB, and the "program" being evaluated is having land and hence becoming ineligible for GB.

Concerns about DD designs: DD designs can be particularly vulnerable to measurement errors in poor quality data. When the changes over time are measured with greater error than the levels, a trade-off emerges between (on the one hand) the ability of a DD design to estimate impacts robustly to time-invariant selection bias and (on the other) the attenuation bias and imprecision that arising from identifying impacts off the poorly measured changes over time. This is an instance of a well-known problem in estimating panel data models in the presence of measurement error (see, for example, the discussion in Deaton, 1995).

Even with good data, the DD assumption of time-invariant and additive selection bias is implausible for some anti-poverty programs in developing countries. DD will give a biased impact estimate if the subsequent outcome changes are a function of initial conditions that also

---

[25] This is my interpretation; Pitt and Khandker (1998) do not mention the DD interpretation of their design given here. However, it is readily verified that the impact estimator implied by solving equations (4a-d) in their paper is the DD estimator described here. (Note that the resulting DD must be normalized by the proportion of landless household in eligible villages to obtain the impact parameter for GB.)

[26] Equivalently, they measure impact by the mean gain amongst households who are landless from living in a village that is eligible for GB, less the corresponding gain amongst those with land.

influenced the assignment of the sample between the two groups. Anti-poverty programs are often targeted to poor sub-groups, such as poor areas. And these same targeting criteria could well influence subsequent growth rates.[27]

Figure 2 illustrates the point. Mean outcomes are plotted over time, before and after the intervention. The lightly-shaded circles represent the observed means for the treatment units, while the hatched circle is the counterfactual at date t=1. Panel (a) shows the initial selection bias, arising from the fact that the program targeted poor areas than the comparison units (dark-shaded). This is not a problem as long as the bias is time invariant, as in panel (b). However, when the attributes on which targeting is based also influence subsequent growth prospects we get a potentially large bias in the DD estimate, as in panel (c).

Two examples illustrate this point. Jalan and Ravallion (1998) show that poor-area development projects in rural China have been targeted to areas with poor infrastructure and that these same characteristics resulted in lower growth rates. They show that there is a large bias in DD estimators in this case, since the changes over time are a function of initial conditions (through an endogeneous growth model) that also influence program placement. On correcting for this bias by controlling for the area characteristics that initially attracted the development projects, the authors found significant longer-term impacts while none had been evident in the standard DD estimator.

The second example draws on the Pitt and Khandker (1998) study of Grameen Bank. Following my interpretation of the Pitt-Khandker method of assessing the impacts of GB credit, it is clear that the authors' key assumption is that the returns to having land are independent of village-level GB eligibility. A bias will arise if GB tends to select villages that have either

[27] There is also the well-known bias in using DD for inferring long-term impacts of training programs that can arise when there is a pre-program earnings "Ashenfelter's dip" (Ashenfelter, 1978).

unusually high or low returns to land. It seems plausible that the returns to land are lower in villages selected for GB, which may well be why they are poor in the first place, and low returns to land would also suggest to GB that such villages have a comparative advantage in the non-farm activities facilitated by GB credit. Then the Pitt-Khandker method will overestimate the impacts of the Grameen Bank.

Controlling for initial heterogeneity is thus crucial to the credibility of DD estimates. Combining DD with PSM — as the most flexible method of cleaning out initial heterogeneity prior to differencing — can go some way toward addressing this concern. Combining PSM for selecting the comparison group with DD can reduce (though probably not eliminate) the bias found in other evaluation methods, including single-difference matching. In an example in the context of poor-area development programs, Ravallion and Chen (2005) first used PSM to clean out the initial heterogeneity between targeted villages and comparison villages, before applying DD using longitudinal observations for both sets of villages. When relevant, pipeline comparison groups can also help to reduce bias in DD studies (Galasso and Ravallion, 2004).

These observations point to important synergies between better data and methods for making single difference comparisons (on the one hand) and double-difference (on the other). Longitudinal observations can help reduce bias in single difference comparisons (eliminating the additive time-invariant component of selection bias). And successful efforts to clean out the heterogeneity in baseline data such as by PSM can reduce the bias in DD estimators.

Panel data are not necessary for calculating the DD impact estimator. All one needs is the set of four means that make up DD, and the means for each of the participant and non-participants groups (possibly after matching) need not be calculated for the same sample. (For

example, recall the above interpretation of the Pitt-Khandker evaluation of Grameen Bank; in this design one does not need longitudinal observations of the villages with and without the GB.)

However, when available, household-level panel data open up further options for counterfactual analysis of the joint distribution of outcomes over time for the purpose of understanding the impacts on poverty dynamics. This approach is developed in Ravallion et al. (1995) for the purpose of measuring the impacts of changes in social spending on the inter-temporal joint distribution of income. So instead of only measuring the impact on poverty (the marginal distribution of income) the authors distinguish impacts on the number of people who escape poverty over time (the "promotion" role of a safety net) from impacts on the number who fall into poverty (the "protection" role). Ravallion et al., apply this approach to an assessment of the impact on poverty transitions of reforms in Hungary's social safety net. Other examples can be found in Lokshin and Ravallion (2000) (on the impacts of changes in Russia's safety net during an economy-wide financial crisis), Gaiha and Imai (2002) (on the Employment Guarantee Scheme in the Indian state of Maharashtra) and van de Walle (2004) (on assessing the performance of Vietnam's safety net in dealing with income shocks).

8. Higher-order differencing: following up ex-participants

Pre-intervention baseline data are sometimes unavailable. Safety-net interventions such as workfare programs and social funds often have to be set up quickly in response to a macroeconomic or agro-climatic crisis, and it is not feasible to delay the operation in order to do a baseline survey. (Nor is randomization feasible in such settings.) Nonetheless, under certain conditions, impacts can still be identified by observing participants' outcomes in the absence of the program after the program rather than before it.

To see what is involved in more formal terms, write the gain for unit *i* from the program at date *t* as $G_{it} = Y_{it}^{T*} - Y_{it}^{C*}$ where Y_{it}^{T*} is the true value of the outcome variable and Y_{it}^{C*} is the true value of the counter-factual outcome for the participant. The observed values are (dropping the *i* subscripts):

$$Y_t^T = Y_t^{T*} + \eta^T + \varepsilon_t^T \tag{9.1}$$

$$Y_t^C = Y_t^{C*} + \eta^C + \varepsilon_t^C \tag{9.2}$$

where η^i (*i*=*T*,*C*) are time invariant error components (such as due to selection bias) and ε_t^i (*i*=*T*,*C*) are zero-mean time-varying error terms. I assume that an estimate of Y_t^C is available for an observationally similar comparison group and I focus on the case of two time periods.

Recall that the key identifying assumption in all double-difference studies is that the selection bias into the program is both additively separable from outcomes and time invariant:

$$E[(Y_2^C - Y_1^C) \mid D_2 = 1] = E[(Y_2^C - Y_1^C) \mid D_2 = 0] \tag{10}$$

Under this assumption, the overall difference-in-difference can be written as:

$$DD = E[(Y_2^T - Y_1^T) \mid D_2 = 1] - E[(Y_2^C - Y_1^C) \mid D_2 = 0] = E[G_2 - G_1 \mid D_2 = 1] \tag{11}$$

In the standard DD set-up with two time periods, period 1 precedes the intervention and $G_1 = 0$. Then *DD* gives the mean current gain to participants at time 2, $DD = E(G_2 \mid D_2 = 1)$. However, in this case, the program is in operation in period 1. The scope for identification arises from the fact that some participants at date 1 subsequently drop out of the program. The triple-difference estimator proposed by Ravallion et al. (2005) is the difference between the double difference for stayers and leavers:

$$DDD = E[(Y_2^T - Y_2^C) - (Y_1^T - Y_1^C) \mid D_2 = 1, D_1 = 1] - E[(Y_2^T - Y_2^C) - (Y_1^T - Y_1^C) \mid D_2 = 0, D_1 = 1] \tag{12}$$

On re-arranging terms, this can also be written as:

$$DDD = [E(G_2 | D_2 = 1, D_1 = 1) - E(G_2 | D_2 = 0, D_1 = 1)] - [E(G_1 | D_2 = 1, D_1 = 1) - E(G_1 | D_2 = 0, D_1 = 1)] \quad (13)$$

The first term in square brackets on the RHS of equation (13) is the net gain to continued participation in the program, given by the difference between the gain to participants in period 2 and the gain to those who dropped out. If we are only interested in the marginal gains from longer participation in the program amongst participants then this first term is the one of interest; selection into the program at the outset is not then an issue. Notice also that there may be some gain to leavers from past participation ($E(G_2 | D_2 = 0, D_1 = 1) \neq 0$). For example, participants may have learnt a skill that raises their post-program earnings. The loss to those who leave the program is $E(G_2 | D_2 = 1, D_1 = 1) - E(G_2 | D_2 = 0, D_1 = 1)$, allowing for the possibility that leavers may benefit from past participation. The second term on the RHS of (13) is the selection bias arising from any effect of the gains at date 1 on participation at date 2.

It is readily verified from (13) that *DDD* consistently identifies the mean gain to participants at period 2, $E(G_2 | D_2 = 1, D_1 = 1)$, if two conditions hold: (i) there is no selection bias in terms of who leaves the program i.e., $E(G_1 | D_2 = 1, D_1 = 1) = E(G_1 | D_2 = 0, D_1 = 1)$; and (ii) there are no current gains to non-participants, i.e., $E(G_2 | D_2 = 0, D_1 = 1) = 0$. A third survey round allows a joint test of these two conditions. If these conditions hold and there is no selection bias in period 3, then there should be no difference in the estimate of gains to participants in period 2 according to whether or not they drop out in period 3.

In applying the above approach, Ravallion et al. (2005) examine what happens to participants' incomes when they leave Argentina's *Trabajar* program as compared to the incomes of continuing participants, after netting out economy-wide changes, as revealed by a

matched comparison group of non-participants. The authors find partial income replacement, amounting to one-quarter of the *Trabajar* wage within six months of leaving the program, though rising to one half in 12 months. Thus there is evidence of a post-program "Ashenfelter's dip," namely when earnings drop sharply at retrenchment, but then recover.

As an aside, suppose that we do not have a comparison group of nonparticipants; instead, we just calculate the double difference for stayers versus leavers (that is, the gain over time for stayers less that for leavers). It is evident that this will only deliver an estimate of the current gain to participants if the counter-factual changes over time are the same for leavers as for stayers. More plausibly, one might expect stayers to be people who tend to have lower prospects for gains over time than leavers in the absence of the program. Then the simple *DD* for stayers versus leavers will underestimate the impact of the program. In the particular case studied by Ravallion et al., the DD for stayers relative to leavers (ignoring those who never participated) turned out to give a quite good approximation to the DDD estimator. However, this need not hold in other applications.

This study also illustrates the potential pitfalls of PSM when data are weak. As compared to the study by Jalan and Ravallion (2002b) on the same program, Ravallion et al., had no choice but to use a lighter survey instrument, with far fewer questions on relevant characteristics of participants and non-participants. This did not deliver plausible single-difference estimates using PSM when compared to the Jalan and Ravallion estimates using single-difference PSM for the same program on richer data. The likely explanation is that using the lighter survey instrument meant that there were many unobservable differences; in other words the conditional independence assumption of PSM was not valid. Given the sequence of the two evaluations, the key omitted variables in the later study were known — they mainly related to local level

connections (as evident in memberships of various neighborhood associations and length of time living in the same barrio). It would appear that Ravallion et al., were able to satisfactorily address this problem by tracking households over time, even using their lighter survey instrument; the follow-up evaluation design was able to difference out the miss-matching errors arising from incomplete data.

From the point of view of evaluation design, this suggests that a trade-off exists between the resources devoted to cross-sectional data collection for the purpose of single-difference matching, versus collecting longitudinal data with a lighter survey instrument.

9. Instrumental variables

We now turn to a method that relaxes the exogeneity assumption of OLS or PSM, and is also robust to time-varying selection bias, unlike DD. The method makes a different identifying assumption to the previous methods — though an assumption that can also be questioned.

Returning to the discussion in section 2, let us now assume that program placement depends on an instrumental variable (IV), *Z*, as:

$$D_i = \gamma Z_i + \nu_i \tag{14}$$

To simplify the exposition, I focus on the common effects specification (equation 7), for which the reduced form equation for outcomes is simply:

$$Y_i = \pi Z_i + X_i \beta^C + \mu_i \tag{15}$$

where $\pi = (\beta_0^T - \beta_0^C)\gamma$ and $\mu_i = (\beta_0^T - \beta_0^C)\nu_i + \varepsilon_i$. When it exists, the Instrumental Variables Estimator (IVE) for the impact parameter is $(\hat{\beta}_0^T - \hat{\beta}_0^C)_{IVE} = \hat{\pi}_{OLS} / \hat{\gamma}_{OLS}$ (in obvious notation). A variation on this is to estimate (14) as a nonlinear binary response model (a probit or logit) and

use the predicted values as the IV for program placement in equation (7). (The first stage regression can include *X* as well.)[28]

How does IVE compare to other methods? The key difference is that the IVE method does not require the exogeneity assumption of the previous methods. We need not assume that $\text{cov}(D_i, \varepsilon_i) = 0$ in an OLS estimate of equation (7) or that changes in participation are exogeneous (as in the DD estimator). Instead, it is assumed that Z_i is exogenous (justifying estimating (14) by OLS), that Z_i matters to placement ($\gamma \neq 0$, assuring existence of the IVE) and that Z_i is not an element of the vector of controls, X_i (allowing us to identify π in (15) separately from β^C). The latter condition is called the "exclusion restriction" (in that Z_i is excluded from (7)). If these assumptions hold then IVE identifies the mean impact of the program that is attributable to the instrument robustly to selection bias stemming from unobserved heterogeneity.

Like all the preceding non-experimental methods, the IVE requires an untestable conditional independence assumption, though it is a different assumption to PSM or OLS; in the case of IVE this is the exclusion restriction.[29] Also note that this is not strictly required when a nonlinear binary response regression is used for the first stage. Then the model is identified off the nonlinearity of the first stage regression. However, it is widely considered preferable to have an identification strategy that is robust to using a linear first stage regression. That requires a justification for excluding Z_i from (7). We shall return to this issue.

---

[28] A good discussion of this estimator can be found in Wooldridge (2002, Paper 18).

[29] If *Z* is a vector (with more than one variable) then the model is over-identified and one can test whether all but one of the IVs is significant when added to the main equation of interest (7). However, one must still leave one IV and so the exclusion restriction is un-testable.

There are some similarities too. As with OLS, the validity of causal inferences for (parametric) IVE rests on mostly *ad hoc* functional form assumptions. Note also that the first stage equation (14) echoes the first stage of PSM method. However, IVE is arguably less demanding on our ability to model the program's assignment than PSM; the instrumental variable *Z* needs to have explanatory power for predicting participation but the IVE is almost certainly more forgiving of insufficient data on the covariates of participation than is PSM.

Notice also that IVE only identifies the effect for a specific population sub-group, namely those induced to take up the program by the instrument; naturally, it is only for that sub-group that the IV can reveal the exogenous variation in program placement. The outcome gain for the sub-group induced to switch by the IV is called the "local average treatment effect" (LATE) in the evaluation literature (Imbens and Angrist, 1994). This sub-group is typically not identified explicitly, so it remains unclear in practice for whom exactly one has identified the mean impact.

The exclusion restriction: This is the Achilles heel of IVE in practice. Until quite recently, the assumption was barely commented on in papers using IVE (possibly even relegated to a footnote on a table of IVE results). However, these days the validity of the exclusion restriction is now routinely questioned in assessments of IVE evaluations in practice. This questioning typically takes the form of proposing some alternative theoretical model for outcomes conditional on placement for which the assumption would not hold.

It can be argued that working in developing country settings makes it even harder than usual to justify the exclusion restrictions found in much applied work in the IVE tradition. This will be the case if (as is widely believed — though not always with well-documented justification) markets tend to work less well in developing country settings. Incomplete and imperfect markets tend to generate non-separabilities and spillover effects that make it harder to

justify excluding certain variables from outcome regressions. For example, in a perfect markets model we would happily exclude household characteristics from a farm-profit regression, and reserve these as potential instruments for the placement of an anti-poverty program that attempts to raise farm yields but affects consumption-leisure choices. However, with imperfect factor markets, the separability between production and consumption decisions breaks down, invalidating this identification strategy.

To give another example, consider the problem of identifying the impact of an individually assigned training program on wages. Following past literature in labor economics one might use characteristics of the household to which each individual belongs as IVs for program participation. The usual argument justifying the exclusion restriction is that these characteristics are not observable to employers and so should not affect wages conditional on program participation (and other observable control variables, such as age and education of the individual worker). However, for at least some of these potential IVs the exclusion restriction is questionable in developing country settings in which the presence of a literate person in the household can exercise a strong productivity effect on an illiterate worker's productivity; this is argued in theory and with supporting evidence (for rural Bangladesh) in Basu et al. (2002).

The validity of widely used exclusion restrictions can be particularly questionable with only a single cross-sectional data set; while one can imagine many variables that are correlated with placement, such as geographic characteristics of an area, it is questionable on *a priori* grounds that those variables are uncorrelated with outcomes given placement. There is more potential for identification with longitudinal (panel) data, on the assumption that (backward and forward) lagged effects die out sufficiently rapidly to justify the exclusion restrictions required

by IVE. Examples of this approach include Rosenzweig and Wolpin (1986), Pitt et al., (1995) and Jalan and Ravallion (2002).

Where do we find an IV? There are essentially two sources, experimental design features and theoretical arguments about the determinants of program placement and outcomes. The following discussion considers these in turn.

Partially randomized designs as a source of instrumental variables: As noted in section 4, it is often the case in social experiments that some of those randomly selected for the program do not want to participate. While finding a valid IV is often difficult, the randomized assignment is a natural choice in this case. Here the exclusion restriction is that being randomly assigned to the program only affects outcomes via actual participation. Looking more closely at how this works is instructive for understanding IVE in this context.[30]

The IVE is now a dummy variable for assignment to the program (=1 if assigned to treatment, 0 if control). (To simplify the exposition I shall drop the control variables from (15).) The parameter γ in (14) is simply the treatment take-up rate while $\pi = E(Y|Z=1) - E(Y|Z=0)$, which is sometimes referred to as the "intention-to-treat" (ITT) effect in the evaluation literature, namely the mean impact for those who are offered the opportunity to be treated. For a pure randomization, *Z* is exogenous. So (14) and (15) are consistently estimated by OLS, giving $\hat{\gamma}$ and $\hat{\pi}$. The ratio:

$$\frac{\hat{\pi}}{\hat{\gamma}} = \frac{\sum (Y_i - \overline{Y})(Z_i - \overline{Z})}{\sum (D_i - \overline{D})(Z_i - \overline{Z})} \qquad (16)$$

[30] For a more general characterization of the theoretical condions under which an IVE delivers the mean impact of a program see Angrist et al. (1996). Also see the dusicussion in Dubin and Rivers (1997).

is then recognized as the IVE regression coefficient of *Y* on *D* with *Z* as the IV. This is the same as simply taking the ITT effect and deflating by the compliance rate, which has been used in past work to correct for selective compliance in randomized evaluations, following Bloom (1984).

Under what conditions does π / γ give the mean impact, as would be obtained by simply comparing the outcome means for treatment and control observations in a social experiment?[31] By construction, π is the weighted mean of $E(Y|Z=1,D=1) - E(Y|Z=0,D=1)$ and $E(Y|Z=1,D=0) - E(Y|Z=0,D=0)$ with weights γ and $1-\gamma$ respectively. This can be re-written as:

$$\begin{aligned} \pi = \gamma\, ATE - \gamma[E(Y|Z=0,D=1) - E(Y|Z=0,D=0)] \\ + E(Y|Z=1,D=0) - E(Y|Z=0,D=0) \end{aligned} \tag{17}$$

where $ATE \equiv E[(Y|Z=1,D=1) - E(Y|Z=1,D=0)]$. Sufficient conditions for $ATE = \pi / \gamma$ are that (i) assignment to the program is a necessary condition for treatment to have an impact, i.e., $E(Y|Z=0,D=1) = E(Y|Z=0,D=0)]$; and (ii) assignment can have no effect if one is not treated, i.e., $E(Y|Z=1,D=0) = E(Y|Z=0,D=0)$. Condition (i) is the more questionable condition. This will fail to hold of non-assigned units become (in effect) treated as a result of some spillover effect (as discussed in section 3).

An example of the above approach to correcting for bias in randomized designs can be found in the aforementioned *Proempleo* experiment. Recall that this included a training component that was assigned randomly. Under the assumption of perfect take-up or random non-compliance, neither the employment nor incomes of those receiving the training were

[31] This is essentially the same question addressed by Angrist et al. (1996) who provide a more complete discussion of the conditions under which causal effects can be identified using instrumental variables. There is a similarity to the analysis in Dubin and Rivers (1993).

significantly different to those of the control group 18 months after the experiment began.[32] However, some of those assigned the training component did not want it, and this selection process was correlated with the outcomes from training. An impact of training was revealed for those with secondary schooling – when the authors corrected for compliance bias using assignment as the IV for treatment (Galasso et al., 2004).

The above discussion has focused on the use of randomized assignment as an IV for treatment, given selective compliance. This idea can be generalized to the use of randomization in identifying economic models of outcomes, or of behaviors instrumental to determining outcomes. We return to this topic in section 10.

Non-experimental sources of instrumental variables: While the existence of a randomized assignment offers an easily defended IV for dealing with endogeneous compliance, in the vast majority of applications one is not so fortunate. Here I point to three popular sources of instrumental variables, geography, politics and discontinuities created by program design.

The geography of program placement has been used for identification in a number of studies using IVE. A recent example of this approach can be found in Attanasio and Vera-Hernandez (2004) who study the impacts of a large nutrition program in rural Colombia that provided food and child care through local community centers. Some people used these facilities while some did not, and there must be a strong presumption that usage is endogenous to outcomes in this setting. To deal with this problem, Attanasio and Vera-Hernandez used the distance of a household to the community center as the IV for attending the community center. The authors address the objections that can be raised against the exclusion restriction. (The other requirement of a valid IV, namely that it is correlated with treatment, is more easily satisfied in

[32] The wage subsidy included in the *Proempleo* experiment did have a significant impact on employment, but not current incomes, though it is plausible that expected future incomes were higher; see Galasso et al., (2004) for further discussion.

this case.) Distance could itself be endogenous through the location choices made by either households or the community centers. Amongst the justifications they give for their choice of IV, the authors note that survey respondents who have moved recently never identified the desire to move closer to a community center as one of the reasons for choosing their location (even though this was one of the options). Tellingly, they also note that if their results were in fact driven by endogeneity of their IV then they would find (spurious) effects on variables that should not have any effect on a priori grounds, such as child birth weight. However, they do not find such effects, supporting the choice of IV.

Political characteristics of geographic areas have been another source of instruments. Understanding the political economy of program placement can aid in identifying impacts. For example, Besley and Case (2000) use the presence of women in state parliaments (in the US) as the IV for workers' compensation insurance when estimating the impacts of compensation on wages and employment. They argue that female law makers favor workers' compensation but that this is unlikely to have an independent effect on the labor market.

To give another example, in evaluating a Bank-supported social fund in Peru, Paxson and Schady (2002) used the extent to which recent elections had seen a switch against the government as the IV for the geographic allocation of program spending in explaining schooling outcomes; their idea was that the geographic allocation of social fund spending would be used in part to "buy back" voters that had switched against the government in the last election. (Their first stage regression was consistent with this hypothesis.) The exogenous variation in spending identified in this way was found to significantly increase school attendance rates.

The third set of examples exploit discontinuities in program design, as discussed in section 6. Here the LATE is in the neighborhood of a cut-off for program eligibility. An

example of this approach can be found in Angrist and Lavy (1999) who assessed the impact on school attainments in Israel of class size. For identification they exploited the fact that an extra teacher (in Israel) was assigned when the class size went above 40. Yet there is no plausible reason why this cut-off point in class size would have an independent effect on attainments, thus justifying the exclusion restriction. The authors find sizeable gains from smaller class sizes, which were not evident using OLS.

Another example is found in Duflo's (2003) study of the impacts of old-age pensions in South Africa on child anthropometric indicators. Women only become eligible for a pension at age 60, while for men it is 65. It is implausible that there would be a discontinuity in outcomes (conditional on treatment) at these critical ages. Following Case and Deaton (1998), Duflo uses eligibility as the IV for receipt of a pension in regressions for anthropometric outcome variables. The author finds that pensions going to women improve girls' nutritional status but not boys', while pensions going to men have no effect on outcomes for either boys or girls.

Two remarks can made about how these methods relate to the discontinuity designs discussed in section 6, whereby one makes a single difference comparison of means either side of the cut-off point. Firstly, and similarly to the aforementioned problem of selective compliance in a randomized design, the use of the discontinuity in the eligibility rule as an IV for actual program placement can address any concerns about selective compliance with those rules; this is discussed further in Battistin and Rettore (2002). Secondly, these IV methods will not in general give the same results as the discontinuity designs discussed in section 6. Specific conditions for equivalence of the two methods are derived in Hahn et al. (2001); the main conditions for equivalence are that the means used in the single-difference comparison are calculated using

appropriate kernel weights and that the IVE estimator is applied to a specific sub-sample, in a neighborhood of the eligibility cut-off point.

As these examples illustrate, the justification of an IVE must ultimately rest on sources of information outside the confines of the quantitative analysis. Those sources might include theoretical arguments, common sense, or empirical arguments based on different types of data, including qualitative data, such as based on knowledge of how the program operates in practice.

10. Learning more from evaluations

So far we have focused on the "internal validity" of an evaluation: does the evaluation design plausibly allow us to obtain a reliable estimate of impact in the specific context? This has been the primary focus of the literature to date. However, there are other concerns related to what can be learnt from an evaluation — to apply the results from the evaluation in other settings and to draw lessons for development knowledge and future policy making. As we will see, these concerns feedback in turn to both data collection and methodology.

Can the lessons from an evaluation be scaled up? The context of a specific intervention often matters to its outcomes, thus confounding inferences for "scaling up" from an impact evaluation. These "external validity" concerns relate to both experimental and non-experimental evaluations.

The essential problem is that if you allow properly for contextual factors it can be hard to make meaningful generalizations for scaling up and replication from trials. The same program works well in one village but fails hopelessly in another. This point is illustrated by the results of Galassso and Ravallion (2005) studying Bangladesh's Food for Education Program. The program worked well in reaching the poor in some villages but not others, even in relatively close proximity.

The key point here is that the institutional context of an intervention may well be hugely important to its impact. External validity concerns about impact evaluations can arise when certain institutions need to be present to even facilitate the experiments. For example, when randomized trails are tied to the activities of specific Non-Governmental Organizations (NGOs) as the facilitators (as in the cases cited by Duflo and Kremer, 2005), there is a concern that the same intervention at national scale may have a very different impact in places without the NGO. Making sure that the control group areas also have the NGO can help, but even then we cannot rule out interaction effects between the NGO's activities and the intervention. In other words, the effect of the NGO may not be "additive" but "multiplicative," such that the difference between measured outcomes for the treatment and control groups does not reveal the impact in the absence of the NGO.

A further concern about external validity is that while partial equilibrium assumptions may be fine for a pilot, that can cease to be so when the program is scaled up nationally, and general equilibrium effects become important (sometimes called "feedback" or "macro" effects in the evaluation literature). For example, an estimate of the impact on schooling of a tuition subsidy based on a randomized trial may be deceptive when scaled up, given that the structure of returns to schooling will alter; Heckman et al., (1998) demonstrate that the partial equilibrium analysis can greatly overestimate the impact of a tuition subsidy once relative wages adjust. To give another example, a small pilot wage subsidy program such as implemented in the *Proempleo* experiment may be unlikely to have much impact on the market wage rate, but that will change when the program is scaled up. Here again the external validity concern stems from the context-specificity of trials; outcomes in the context of the trial may differ appreciably (in either direction) once the intervention is scaled up and prices and wages respond.

Contextual factors are clearly crucial to policy and program performance; at the risk of overstating the point, in certain contexts anything will work, and in others everything will fail. A key factor in program success is often adapting properly to the institutional and socio-economic context in which you have to work. That is what good project staff do all the time. They might draw on the body of knowledge from past evaluations, but these can almost never be conclusive and may even be highly deceptive if used mechanically.

The realized impacts on scaling up can also differ from the trial results (whether randomized or not) because the socio-economic composition of program participation varies with scale. Ravallion (2004a) discusses how this can happen, and presents results from a series of country case studies, all of which suggest that the incidence of program benefits improves with scaling up. Trial results may well underestimate how pro-poor a program is likely to be after scaling up because initial benefits tend to be captured more by the non-poor.

What determines impact? These external validity concerns point to the need to supplement the evaluation tools described above by other sources of information that can throw light on the processes that influence the measured outcomes.

One approach is to repeat the evaluation in different contexts, as proposed by Duflo and Kremer (2005). An example can be found in the aforementioned study by Galasso and Ravallion in which the impact of Bangladesh's Food-for-Education program was assessed across each of 100 villages in Bangladesh and the results were correlated with characteristics of those villages. The authors found that the revealed differences in program performance were partly explicable in terms of observable village characteristics, such as the extent of intra-village inequality (with more unequal villages being less effective in reaching their poor through the program). Repeating evaluations across different settings and at different scales can clearly help address

these concerns. The practical feasibility of being able to do a sufficient number of trials (to span the relevant domain of variation found in reality) remains a moot point. The scale of a randomized trial needed to test a large national program could well be prohibitive. Nonetheless, varying contexts for trials is clearly a good idea, subject to feasibility.

An alternative approach is to probe more deeply into why a program has (or does not have) impact in a specific context, as a basis for inferring whether it would work in a different context. Here better use can often be made of intermediate indicators. The most common evaluation design identifies a relatively small number of "final outcome" indicators, and aim to assess the program's impact on those indicators. However, instead of using only final outcome indicators, one may choose to also study impacts on certain intermediate indicators of behavior.

For example, the inter-temporal behavioral responses of participants in anti-poverty programs are of obvious relevance to understanding their impacts. An impact evaluation of a program of compensatory cash transfers to Mexican farmers found that the transfers partly invested, with second-round effects on future incomes (Sadoulet et al., 2001). Similarly, Ravallion and Chen (2005) found that participants in a World Bank poor-area development program in China saved a large share of the income gains from the program (as estimated using the matched double-difference method described in section 7). Identifying responses through savings and investment provides a clue to understanding current impacts on living standards and the possible future welfare gains beyond the project's current life span. Instead of focusing solely on the agreed welfare indicator, one collects and analyzes data on a potentially wide range of intermediate indicators relevant to understanding the processes determining impacts.

As an aside, this also illustrates a common concern in evaluation studies, given behavioral responses. The study period is rarely much longer than the period of the program's

disbursements. However, a share of the impact on peoples' living standards may occur beyond the life of the project. This does not necessarily mean that credible evaluations will need to track welfare impacts over much longer periods than is typically the case — raising concerns about feasibility. But it does suggest that evaluations need to look carefully at impacts on partial intermediate indicators of longer-term impacts even when good measures of the welfare objective are available within the project cycle. The choice of such indicators will need to be informed by an understanding of participants' behavioral responses to the program.

In learning from an evaluation, one often needs to draw on other sources of information external to the evaluation. There are many possible sources, including qualitative research (intensive interviews with participants and administrators).[33] The essential idea is to "test" the assumptions made by an intervention. This is sometimes called "theory-based evaluation," though that is hardly an ideal term given that non-experimental identification strategies for mean impacts are often theory-based (as discussed in the last section). Weiss (2001) illustrates this approach in the abstract in the context of evaluating the impacts of community-based anti-poverty programs. An example is found in an evaluation of social funds (SFs) by the World Bank's Operations Evaluation Department, as summarized in Carvalho and White (2004). While the overall aim of a SF is typically to reduce poverty, the OED study was interested in seeing whether SFs worked the way that was intended by their designers. For example, did local communities participate? Who participated? Was there "capture" of the SF by local elites (as some critics have argued)? Building on Weiss (2001), the OED evaluation identified a series of key hypothesized links connecting the intervention to outcomes and tested whether each one worked. For example, in one of the country studies for the OED evaluation of SFs, Rao and Ibanez (2003) tested the assumption that a SF works by local communities collectively proposing

[33] See the discussion on "mixed methods" in Rao and Woolcock (2003).

the sub-projects that they want; for a SF in Jamaica, the authors found that the process was often dominated by local elites.

In practice, it is very unlikely that all the relevant assumptions are testable (including alternative assumptions made by different theories that might yield similar impacts). Nor is it clear that the process determining the impact of a program can always be decomposed into a neat series of testable links within a unique causal chain; there may be more complex forms of interaction and simultaneity that do not lend themselves to this type of analysis. So the "theory-based evaluation" approach cannot be considered a serious substitute for assessing impacts on final outcomes by credible (experimental or non-experimental) methods. However, it can be a useful complement to such evaluations, to better understanding measured impacts.

Project monitoring data bases are an important, under-utilized, source of information. Too often the project monitoring data and the information system have negligible evaluative content. This is not inevitably the case. For example, the idea of combining spending maps with poverty maps for rapid assessments of the targeting performance of a decentralized anti-poverty program is a promising illustration of how, at modest cost, standard monitoring data can be made more useful for providing information on how the program is working and in a way that provides sufficiently rapid feedback to a project to allow corrections along the way (Ravallion, 2001).

The *Proempleo* experiment provides an example of how information external to the evaluation can carry important lessons for scaling up. Recall that *Proempleo* randomly assigned vouchers for a wage subsidy across (typically poor) people currently in a workfare program and tracked their subsequent success in getting regular work. A randomized control group located the counterfactual. The results did indicate a significant impact of the wage-subsidy voucher on employment. But when cross-checks were made against central administrative data,

supplemented by informal interviews with the hiring firms, it was found that there was very low take-up of the wage subsidy by firms (Galasso et al., 2004). The scheme was highly cost effective; the government saved 5% of its workfare wage bill for an outlay on subsidies that represented only 10% of that saving.

However, the supplementary cross-checks against other data revealed that *Proempleo* did not work the way its design had intended. The bulk of the gain in employment for participants was not through higher demand for their labor induced by the wage subsidy. Rather the impact arose from supply side effects; the voucher had credential value to workers – it acted like a "letter of introduction" that few people had (and how it was allocated was a secret). This could not be revealed by the (randomized) evaluation, but required supplementary data. The extra insight obtained about how *Proempleo* actually worked in the context of its trial setting also carried implications for scaling up, which put emphasis on providing better information for poor workers about how to get a job rather than providing wage subsidies.

Spillover effects also point to the importance of a deeper understanding of how a program operates. Indirect (or "second-round") impacts on non-participants are common. A workfare program may lead to higher earnings for non-participants. Or a road improvement project in one area might improve accessibility elsewhere. Depending on how important these indirect effects are thought to be in the specific application, the "program" may need to be redefined to embrace the spillover effects. Or one might need to combine the type of evaluation discussed here with other tools, such as a model of the labor market to pick up other benefits.

The extreme form of a spillover effect is an economy-wide program. The evaluation tools discussed in this paper are for assigned programs, but have little obvious role for economy-wide programs in which no explicit assignment process is evident, or if it is, the spillover effects

are likely to be pervasive. When some countries get the economy-wide program but some do not, cross-country comparative work (such as growth regressions) can reveal impacts. That identification task is often difficult, notably because there are typically latent factors at country level that simultaneously influence outcomes and whether a country adopts the policy in question. And even when the identification strategy is accepted, carrying the generalized lessons from cross-country regressions to inform policy-making in any one country can be highly problematic. There are also a number of promising examples of how simulation tools for economy wide policies such as Computable General Equilibrium models can be combined with household-level survey data to assess impacts on poverty and inequality.[34] These simulation methods make it far easier to attribute impacts to the policy change, though this advantage comes at the cost of the need to make many more assumptions about how the economy works.

Is the evaluation answering the relevant policy questions? Arguably the most important things we want to learn from any evaluation relate to its lessons for future policies. Here standard evaluation practices can start to look disappointingly uninformative on closer inspection.

One issue is the choice of counterfactual. The classic formulation of the evaluation problem assesses mean impacts on those who receive the program, relative to counterfactual outcomes in the absence of the program. However, this may fall well short of addressing the concerns of policy makers. While common practice is to use outcomes in the absence of the program as the counterfactual, the alternative of interest to policy makers is often to spend the same resources on some other program (possibly a different version of the same program), rather than to do nothing. The evaluation problem is formally unchanged if we think of some alternative program as the counterfactual. Or, in principle, we might repeat the analysis relative to the "do nothing counterfactual" for each possible alternative and compare them. This is rare in

[34] See, for example, Bourguignon et al. (2003) and Chen and Ravallion (2004).

practice, however. A specific program may appear to perform well against the option of doing nothing, but poorly against some feasible alternative.

For example, drawing on their impact evaluation of a workfare program in India, Ravallion and Datt (1995) show that the program substantially reduced poverty amongst the participants relative to the counterfactual of no program. Yet, once the costs of the program were factored in (including the foregone income of workfare participants), the authors found that the alternative counterfactual of a uniform (un-targeted) allocation of the same budget outlay would have had more impact on poverty.[35]

A further issue, with greater bearing on the methods used for evaluation, is whether we have identified the most relevant impact parameters from the point of view of the policy question at hand. The classic formulation of the evaluation problem focuses on mean outcomes, such as mean income or consumption. This is hardly appropriate for programs that have as their (more-or-less) explicit objective to reduce poverty, rather than to promote economic growth *per se*. There is nothing to stop us re-interpreting the outcome measure as a variable taking the value Y_i=1 if unit *i* is deemed to be poor and Y_i=0 otherwise. Then equation (2) gives the program's impact on the headcount index of poverty (% below the poverty line).[36] By repeating the impact calculation for multiple "poverty lines" one can then trace out the impact on the cumulative distribution of income. Higher order poverty measures (that penalize inequality amongst the poor) can also be accommodated as long as they are members of the (broad) class of additive measures, by which the aggregate poverty measure can be written as the population-weighed mean of all individual poverty measures in that population.[37] We may have multiple indicators,

[35] For another example of the same result see Murgai and Rravallion (2005).
[36] I leave aside the econometric issues that arise with linear binary response models.
[37] See Atkinson (1987) on the general form of these measures and examples in the literature.

such as for different poverty lines or other non-income dimensions of poverty, such as non-income indicators of future poverty (such as schooling). This is all feasible with the same tools, though evaluation practice has been rather narrow in its focus.

Poverty measures are still mean impacts, albeit for a summary statistic about the marginal distribution of outcomes. They are also conditional means, in that the impacts vary across the sample of participants, but only as functions of the observables used as control variables. More generally, we may want to know about the joint distribution of Y^T and Y^C. We cannot know this from a social experiment, which only reveals net counterfactual mean outcomes for those treated; *ATET* gives the gives the mean gain net of losses amongst participants. However, there are anti-poverty programs in practice that can generate losses amongst participants; for example, an agricultural development project invariably imposes costs on participants (such as their own time) yet with uncertain future gains (such as due to the risk that prices will fall).

This points to the need for estimates of a wider range of impact parameters than the simple mean impact (*ATET*) or the marginal distribution of impacts. (Recall that this came up in section 7, in the discussion of the use of panel data in studying impacts on poverty dynamics.) Instead of focusing solely on the net gains to the poor (say) we may ask how many losers there are amongst the poor, and how many gainers. Some interventions may yield losers even though mean impact is positive and policy makers will understandably want to know about those losers, as well as the gainers. (This can be true at any given poverty line.) Thus one can relax the "anonymity" or "veil of ignorance" assumption of traditional welfare analysis, whereby outcomes are judged solely by changes in the marginal distribution (Carneiro et al., 2001). This is clearly of greater relevance to some applications than others. For example, trade reforms are very likely to generate both losers and gainers at any given level of living, given general

equilibrium effects and the heterogeneity in net trading positions in relevant markets (Ravallion, 2004b). But even for policies with only non-negative impacts, horizontal impacts at given levels of living can be expected. Methods for estimating the joint distribution of Y^T and Y^C are developed in Heckman et al. (1997a). The analogous estimator to OLS applied to the classic "common effects" specification (equation 7) is a random effects estimator in which the coefficient on treatment dummy variable contains a stochastic components.

When the policy issue is whether to expand or contract a given program at the margin, the classic estimator of mean-impact on the treated (by experimental or non-experimental methods) is actually of little or no interest. The problem of estimating the marginal impact of a greater duration of exposure to the program on those treated was considered in section 8, using the example of comparing "leavers" and "stayers" in a workfare program (Ravallion et al., 2005). Another example can be found in the study by Behrman et al. (2004) of the impacts on children's cognitive skills and health status of longer exposure to a preschool program in Bolivia. The authors provide an estimate of the marginal impact of higher program duration by comparing the cumulative effects of different durations using a matching estimator. In such cases, selection into the program is not an issue, and we do not even need data on units who never participated. The discontinuity design method discussed in section 5 (in its non-parametric form) and section 9 (in its parametric IV form) is also delivering an estimate of the marginal gain from a program, namely the gain when the program is expanded (or contracted) by a small change in the eligibility cut-off point.

A deeper understanding of the factors determining outcomes in *ex post* evaluations can also help in simulating the likely impacts of changes in program or policy design *ex ante*.

Naturally, *ex ante* simulations require many more assumptions about how an economy works.[38] As far as possible one would like to see those assumptions anchored to past knowledge built up from rigorous *ex post* evaluations. For example, by combining a randomized evaluation design with a structural model of education choices, Attanasio et al. (2004) are able to greatly expand the set of policy-relevant questions about the design of *PROGRESA* that a conventional evaluation can answer. The authors replicate past work indicating that the program significantly increased secondary school attendance from poor families. However, by also modeling the determinants of school choice — exploiting the randomized design for identification — they also show that a budget-neutral switch of the enrolment subsidy from primary to secondary school would have delivered a net gain. *PROGRESA* had impact, but it could have had more impact.

11. Conclusions

Two main lessons for future evaluations of anti-poverty programs emerge from this survey. Firstly, no single evaluation tool can claim to be ideal in all circumstances. While randomization can be a powerful tool for assessing impact, it is neither necessary nor sufficient for a good evaluation. While economists have sometimes been too uncritical of their non-experimental identification strategies, credible means of isolating at least a share of the exogenous variation in an endogenously placed program can still be found in practice. Good evaluations draw pragmatically from the full range of tools available, often combining methods: randomizing some aspects and using econometric methods to deal with the non-random elements, using randomized elements of a program as a source of instrumental variables, or by combining score matching methods with longitudinal observations to try to eliminate matching errors with imperfect data. Good evaluations typically also require that the evaluator is involved

[38] For a useful overview of *ex ante* methods see Bourguignon and Ferreira (2003).

from the programs' inception and is very well informed about how the program works on the ground; the features of program design and implementation can sometimes provide important clues for assessing impact by non-experimental means.

Secondly, even putting internal validity concerns to one side, it is unlikely that the tools of counter-factual analysis for mean impacts on well-defined outcome variables are ever going to be sufficient for drawing reliable lessons for future development projects and policies. The context in which a program is placed can exercise a powerful influence on outcomes. This points to the need for a deeper understanding of *why* a program does or does not have impact. It also calls for an eclectic approach drawing on various sources, including replications across differing contexts when feasible, and testing the assumptions made in a program's design, such as by tracking intermediate variables of relevance or by drawing on supplementary theories or evidence external to the evaluation. In drawing useful lessons for anti-poverty policy, we need a richer set of impact parameters than has been traditional in evaluation practice, including distinguishing the impacts on gainers from losers at any given level of living. The choice of parameters to be estimated in an evaluation must ultimately depend on the policy question to be answered; for policy makers this is a mundane point, but for evaluators it seems to be ignored too often.

Figure 1: Region of common support

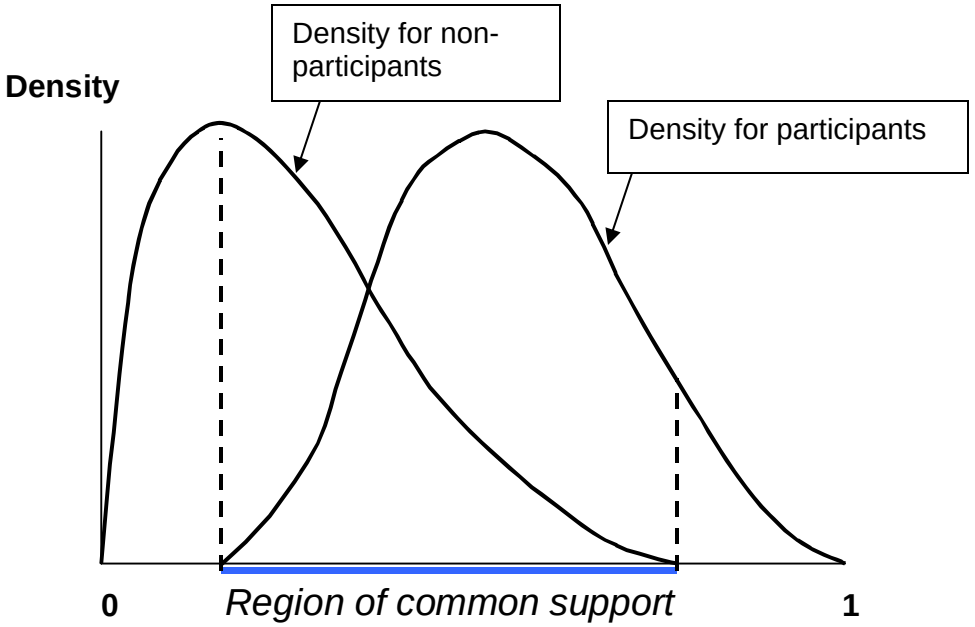
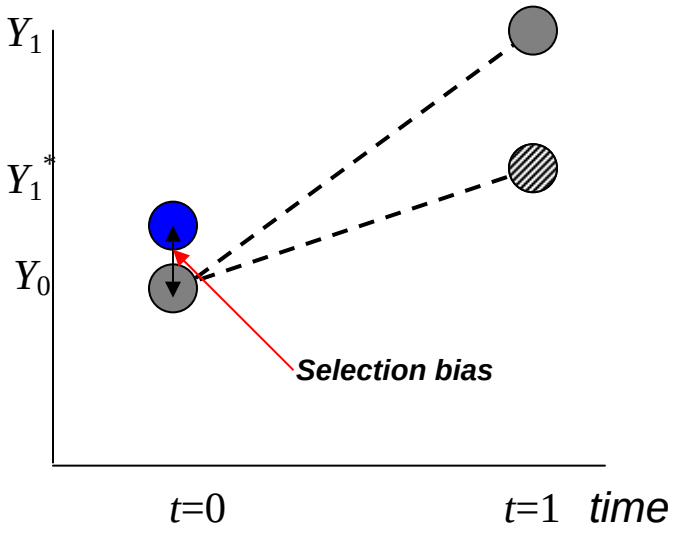
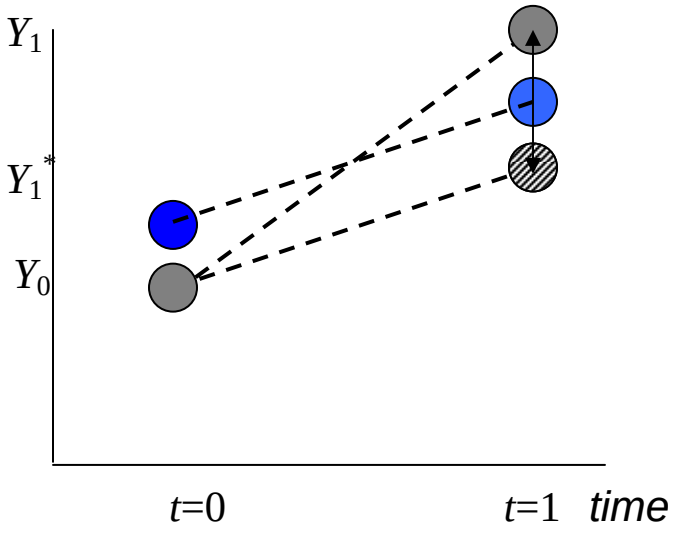


Figure 2: Bias in double-difference estimates for a targeted anti-poverty program

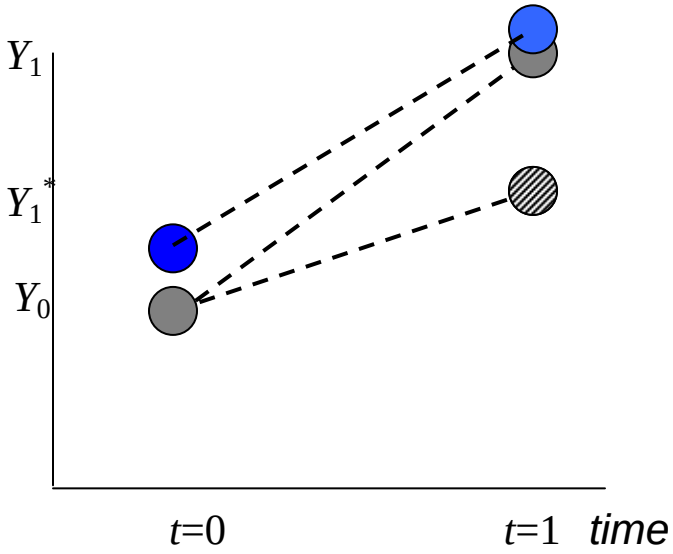
(a)



(b)



(c)



References

Agodini, Roberto and Mark Dynarski, 2004. "Are Experiments the Only Option? A Look at Dropout Prevention Programs," *Review of Economics and Statistics* 86(1): 180-194.

Angrist, Joshua, Eric Bettinger, Erik Bloom, Elizabeth King and Michael Kremer, 2002, "Vouchers for Private Schooling in Colombia: Evidence from a Randomized Natural Experiment," *American Economic Review*, 92(5): 1535-1558.

Angrist, Joshua and Jinyong Hahn, 2004, "When to Control for Covariates? Panel Asymptotics for Estimates of Treatment Effects," *Review of Economics and Statistics*, 86(1): 58-72.

Angrist, Joshua, Guido Imbens and Donald Rubin, 1996, "Identification of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association*, XCI: 444-455.

Angrist, Joshua and Alan Krueger, 2001, "Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments," *Journal of Economic Perspectives* 15(4): 69-85.

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

Ashenfelter, Orley, 1978, "Estimating the Effect of Training Programs on Earnings," *Review of Economic Studies* 60: 47-57.

Atkinson, Anthony, 1987, "On the Measurement of Poverty," *Econometrica*, 55: 749-64.

Attanasio, Orazio, Costas Meghir and Ana Santiago, 2004, "Education Choices in Mexico: Using a Structural Model and a Randomized Experiment to Evaluate PROGRESA," Working Paper EWP04/04, Centre for the Evaluation of Development Policies, Institute of Fiscal Studies London.

Attanasio, Orazio and A. Marcos Vera-Hernandez, 2004. "Medium and Long Run Effects of Nutrition and Child Care: Evaluation of a Community Nursery Programme in Rural Colombia," Working Paper EWP04/06, Centre for the Evaluation of Development Policies, Institute of Fiscal Studies London.

Basu, Kaushik, Ambar Narayan and Martin Ravallion, 2002, "Is Literacy Shared Within Households?" *Labor Economics* 8: 649-665.

Battistin, Erich and Enrico Rettore, 2002, "Testing for Programme Effects in a Regression Discontinuity Design with Imperfect Compliance," *Journal of the Royal Statistical*

Society A, 165(1): 39-57

Behrman, Jere, Yingmei Cheng and Petra Todd, 2004, "Evaluating Preschool Programs When Length of Exposure to the Program Varies: A Nonparametric Approach," *Review of Economics and Statistics*, 86(1): 108-32.

Behrman, Jere, Piyali Sengupta and Petra Todd, 2002, "Progressing through PROGESA: An Impact Assessment of a School Subsidy Experiment in Mexico," mimeo, University of Pennsylvania.

Bertrand, Marianne, Esther Duflo and Sendhil Mullainathan, 2004, "How Much Should we Trust Differences-in-Differences Estimates?" *Quarterly J ournal of Economics*, 119(1): 249-275.

Besley, Timothy and Anne Case, 2000, "Unnatural Experiments? Estimating the Incidence of Endogeneous Policies," *Economic Journal* 110(November): F672-F694.

Bloom, Howard S., 1984, "Accounting for No-shows in Experimental Evaluation Designs," *Evaluation Review* 8: 225-246.

Bourguignon, François and Francisco Ferreira, 2003, "Ex-ante Evaluation of Policy Reforms Using Behavioural Models," in Bourguignon, F. and L. Pereira da Silva (eds.) *The Impact of Economic Policies on Poverty and Income Distribution*, New York: Oxford University Press.

Bourguignon, Francois, Anne-Sophie Robilliard and Sherman Robinson, 2003. "Representative Versus Real Households in the Macro-Economic Modeling of Inequality," Working Paper 2003-05, DELTA, Paris.

Buddelmeyer, Hielke and Emmanuel Shoufias, 2004, "An Evaluation of the Performance of Regression Discountinuity Design on PROGRESA," Policy Research Working Paper 3386, World bank, Washington DC.

Burtless, Gary, 1985, "Are Targeted Wage Subsidies Harmful? Evidence from a Wage Voucher Experiment," *Industrial and Labor Relations Review*, Vol. 39, pp. 105-115.

_____, 1995, "The Case for Randomized Field Trials in Economic and Policy Research," *Journal of Economic Perspectives* 9(2): 63-84.

Carneiro, Pedro, Karsten Hansen and James Heckman, 2001, "Removing the Veil of Ignorance in Assessing the Distributional Impacts of Social Policies," *Swedish Economic Policy Review* 8: 273-301.

Carvalho, Soniya and Howard White, 2004, "Theory-Based Evaluation: The Case of Social Funds," *American Journal of Evaluation* 25(2): 141-160.

Case, Anne and Angus Deaton, 1998, "Large Cash Transfers to the Elderly in South Africa," *Economic Journal* 108:1330-61.

Chase, Robert, 2002, "Supporting Communities in Transition: The Impact of the Armenian Social Investment Fund," *World Bank Economic Review*, 16(2): 219-240.

Chen, Shaohua and Martin Ravallion, 2004, "Household Welfare Impacts of WTO Accession in China,", *World Bank Economic Review*, 18(1): 29-58.

Cook, Thomas, 2001. "Comments: Impact Evaluation: Concepts and Methods," in O. Feinstein and R. Piccioto (eds), *Evaluation and Poverty Reduction*, New Brunswick, NJ: Transaction Publications.

Deaton, Angus, 1995, "Data and Econometric Tools for Development Analysis," in Jere Behrman and T.N. Srinivasan (eds), *Handbook of Development Economics, Volume 3*, Amsterdam: North-Holland.

Dehejia, Rajeev, 2005, "Practical Propensity Score Matching: A Reply to Smith and Todd," *Journal of Econometrics* 125(1-2), 355-364.

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

Diaz, Juan Jose and Sudhanshu Handa, 2004, "An Assessment of Propensity Score Matching as a Nonexperimental Impact Estimator: Evidence from a Mexican Poverty Program," mimeo, University of North Carolina Chapel Hill.

Dubin, Jeffrey A., and Douglas Rivers, 1993, "Experimental Estimates of the Impact of Wage Subsidies," *Journal of Econometrics*, 56(1/2): 219-242.

Duflo, Esther, 2001, "Schooling and Labor Market Consequences of School Construction in Indonesia: Evidence from an Unusual Policy Experiment," *American Economic Review*, 91(4): 795-813.

_____, 2003, "Grandmothers and Granddaughters: Old Age Pension and Intrahousehold Allocation in South Africa," *World Bank Economic Review* 17(1): 1-26.

Duflo, Esther and Michael Kremer, 2005, "Use of Randomization in the Evaluation of Development Effectiveness," in George Pitman, George, Osvaldo Feinstein and Gregory

Ingram (eds.) *Evaluating Development Effectiveness*, New Brunwick, NJ: Transaction Publishers.

Dubin, Jeffrey A., and Douglas Rivers, 1993, "Experimental Estimates of the Impact of Wage Subsidies," J*ournal of Econometrics*, 56(1/2), 219-242.

Foster, James, J. Greer, and Erik Thorbecke, 1984, "A Class of Decomposable Poverty Measures," *Econometrica*, 52: 761-765.

Frankenberg, Elizabeth, Wayan Suriastini and Duncan Thomas, 2005, "Can Expanding Access to Basic Healthcare Improve Children's Health Status? Lessons from Indonesia's 'Midwife in the Village' Program," *Population Studies* 59(1): 5-19.

Frölich, Markus, 2004, "Finite-Sample Properties of Propensity-Score Matching and Weighting Estimators," *Review of Economics and Statistics*, 86(1): 77-90.

Gaiha, Raghav and Katushi Imai, 2002, "Rural Public Works and Poverty Alleviation: The Case of the Employment Guarantee Scheme in Maharashtra," *International Review of Applied Economics* 16(2): 131-151.

Galasso, Emanuela and Martin Ravallion, 2004, "Social Protection in a Crisis: Argentina's *Plan Jefes y Jefas*," *World Bank Economic Review*, 18(3): 367-399.

_____ and _____, 2005, "Decentralized Targeting of an Anti-Poverty Program," *Journal of Public Economics*, 85: 705-727.

Galasso, Emanuela, Martin Ravallion and Agustin Salvia, 2004, "Assisting the Transition from Workfare to Work: Argentina's Proempleo Experiment", *Industrial and Labor Relations Review*, 57(5):.128-142.

Galiani, Sebastian, Paul Gertler, and Ernesto Schargrodsky, 2005, "Water for Life: The Impact of the Privatization of Water Services on Child Mortality," *Journal of Political Economy*, 113(1): 83-119.

Gertler, Paul, 2004. "Do Conditional Cash Transfers Improve Child Health? Evidence from PROGRESA's Control Randomized Experiment" *American Economic Review, Papers\ and Proceedings* 94(2): 336-41.

Glewwe, Paul, Michael Kremer, Sylvie Moulin and Eric Zitzewitz, 2004, "Retrospective vs. Prospective Analysis of School Inputs: The Cse of Flip Charts in Kenya," *Journal of Development Economics* 74: 251-268.

Godtland, Erin, Elizabeth Sadoulet, Alain De Janvry, Rinku Murgai and Oscar Ortiz, 2004, "The

Impact of Farmer Field Schools on Knowledge and Productivity: A Study of Potato Farmers in the Peruvian Andes," *Economic Development and Cultural Change*, 53(1): 63-92.

Hahn, Jinyong, Petra Todd and Wilbert Van der Klaauw, 2001, "Identification and Estimation of Treatment Effects with a Regression-Discontinutiy Design," *Econometrica* 69(1): 201-209.

Heckman, James, Jeffrey Smith and N. Clements, 1997a, "Making the Most Out of Programme Evaluations and Social Experiments: Accounting for heterogeneity in Programme Impacts," *Review of Economic Studies* 64(4), 487-535.

Heckman, James, Hidehiko Ichimura, and Petra Todd, 1997b, "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme," *Review of Economic Studies* 64(4), 605-654.

Heckman, James, Hidehiko Ichimura, Jeffrey Smith, and Petra Todd, 1998, "Characterizing Selection Bias using Experimental Data," *Econometrica* 66, 1017-1099.

Heckman, James, Robert Lalonde and James Smith, 1999, "The Economics and Econometrics of Active Labor Market Programs," *Handbook of Labor Economics, Volume 3*, Ashenfelter, A. and D. Card, eds., Amsterdam: Elsevier Science.

Heckman, James, L. Lochner and C. Taber, 1998, "General Equilibrium Treatment Effects," *American Economic Review Papers and Proceedings* 88: 381-386.

Heckman, James and Richard Robb, 1985, "Alternative Methods of Evaluating the Impact of Interventions: An Overview", *Journal of Econometrics*, 30: 239-67.

Heckman, James and Jeffrey Smith, 1995, "Assessing the Case for Social Experiments," *Journal of Economic Perspectives* 9(2): 85-110.

Hirano, Keisuke and Guido Imbens, 2004, "The Propensity Score with Continuous Treatments," In *Missing Data and Bayesian Methods in Practice*, Wiley forthcoming.

Hoddinott, John and Emmanuel Skoufias, 2004, "The Impact of PROGRESA on Food Consumption," *Economic Development and Cultural Change* 53(1): 37-61.

Imbens, Guido, 2000, "The Role of the Propensity Score in Estimating Dose-Response Functions," *Biometrika* 83: 706-710.

_____, 2004, "Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Review," *Review of Economics and Statistics* 86(1): 4-29.

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

Jacob, Brian and Lars Lefgren, 2004, "Remedial Education and Student Achievement: A Regression-Discontinuity Analysis," *Review of Economics and Statistics* 86(1): 226-44

Jacoby, Hanan G., 2002, "Is There an Intrahousehold 'Flypaper Effect'? Evidence from a School Feeding Programme," *Economic Journal* 112(476): 196-221.

Jalan, Jyotsna and Martin Ravallion, 1998, "Are There Dynamic Gains from a Poor-Area Development Program?" *Journal of Public Economics*, 67(1), 65-86.

_____ and _____, 2002, "Geographic Poverty Traps? A Micro Model of Consumption Growth in Rural China", *Journal of Applied Econometrics* 17(4): 329-346.

_____ and _____, 2003a, "Does Piped Water Reduce Diarrhea for Children in Rural India?" *Journal of Econometrics* 112: 153-173.

_____ and _____, 2003b, "Estimating Benefit Incidence for an Anti-poverty Program using Propensity Score Matching," *Journal of Business and Economic Statistics*, 21(1): 19-30.

Kapoor, Anju Gupta, 2002, *Review of Impact Evaluation Methodologies Used by the Operations Evaluation Department over 25 Years*, Operations Evaluation Department, World Bank.

Lalonde, Robert, 1986, "Evaluating the Econometric Evaluations of Training Programs," *American Economic Review* 76: 604-620.

Lokshin, M., and M. Ravallion, 2000, "Welfare Impacts of Russia's 1998 Financial Crisis and the Response of the Public Safety Net." *Economics of Transition*, 8(2): 269-295.

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

Moffitt, Robert, 1991, "Program Evaluation with Nonexperimental Data," *Evaluation Review*, 15(3): 291-314.

_____, 2003, "The Role of Randomized Field Trials in Social Science Research: A Perspective from Evaluations of Reforms of Social Welfare Programs," Cemmap Working Paper, CWP23/02, Department of Economics, University College London.

Murgai, Rinku and Martin Ravallion, 2005, "Is a Guaranteed Living Wage a Good Anti-Poverty Policy?" Policy Research Working Paper, World Bank, Washington DC.

Newman, John, Menno Pradhan, Laura B. Rawlings, Geert Ridder, Ramiro Coa, and Jose Luis

Evia, 2002, “An Impact Evaluation of Education, Health, and Water Supply Investments by the Bolivian Social Investment Fund,” *World Bank Economic Review*, 16: 241-274.

Paxson, Christina and Norbert R. Schady, 2002, **“**The Allocation and Impact of Social Funds: Spending on School Infrastructure in Peru,” *World Bank Economic Review* 16: 297-319.

Piehl, Anne, Suzanne Cooper, Anthony Braga and David Kennedy, 2003, “Testing for Structural Breaks in the Evaluation of Programs,” *Review of Economics and Statistics* 85(3): 550-558.

Pitt, Mark and Shahidur Khandker, 1998, “The Impact of Group-Based Credit Programs on Poor Households in Bangladesh: Does the Gender of Participants Matter?” *Journal of Political Economy* 106: 958-998.

Pitt, Mark, Mark Rosenzweig, and Donna Gibbons, 1995, “The Determinants and Consequences of the Placement of Government Programs in Indonesia, in: D. van de Walle and K. Nead, eds., *Public spending and the poor: Theory and evidence* (Johns Hopkins University Press, Baltimore).

Rao, Vijayendra and Ana Maria Ibanez, 2003, “The Social Impact of Social Funds in Jamaica: A Mixed Methods Analysis of Participation, Targeting and Collective Action in Community Driven Development,” *Journal of Development Studies*, forthcoming. Policy Research Working Paper 2970, World Bank.

Rao, Vijayendra and Michael Woolcock, 2003. “Integrating Qualitative and Quantitative Approaches in Program Evaluation,” in F. Bourguignon and L. Pereira da Silva (eds.), *The Impact of Economic Policies on Poverty and Income Distribution*, New York: Oxford University Press.

Ravallion, Martin, 2000, “Monitoring Targeting Performance when Decentralized Allocations to the Poor are Unobserved,” *World Bank Economic Review* 14(2): 331-45.

_____, 2003, “Assessing the Poverty Impact of an Assigned Program,” in Bourguignon, F. and L. Pereira da Silva (eds.) *The Impact of Economic Policies on Poverty and Income Distribution*, New York: Oxford University Press.

_____, 2004a, “Who is Protected from Budget Cuts?” *Journal of Policy Reform*, 7(2): 109-22.

_____, 2004b, "Looking beyond Averages in the Trade and Poverty Debate," Policy Research Working Paper 3461, World Bank, Washington DC.

Ravallion, Martin and Shaohua Chen, 2005, "Hidden Impact: Household Saving in Response to a Poor-Area Development Project," *Journal of Public Economics*, forthcoming.

Ravallion, Martin and Gaurav Datt, 1995. "Is Targeting through a Work Requirement Efficient? Some Evidence for Rural India," in D. van de Walle and K. Nead (eds) *Public Spending and the Poor: Theory and Evidence*, Baltimore: Johns Hopkins University Press.

Ravallion, Martin, Emanuela Galasso, Teodoro Lazo and Ernesto Philipp, 2005, "What Can Ex-Participants Reveal About a Program's Impact?" *Journal of Human Resources*, 40(Winter): 208-230.

Ravallion, Martin, Dominique van de Walle and Madhur Gaurtam, 1995, "Testing a Social Safety Net," *Journal of Public Economics*, 57(2): 175-199.

Rosenbaum, Paul and Donald Rubin, 1983, "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70, 41-55.

Rosenzweig, Mark and Kenenth Wolpin, 1986, "Evaluating the Effects of Optimally Distributed Public Programs: Child Health and Family Planning Interventions," *American Economic Review* 76, 470-82.

Rubin, Donald B., 1979, "Using Multivariate Matched Sampling and Regression Adjustment to Control Bias in Observational Studies," *Journal of the American Statistical Association* 74: 318-328.

Rubin, Donald B., and N. Thomas, 2000, "Combining propensity score matching with additional adjustments for prognostic covariates," *Journal of the American Statistical Association* 95, 573-585.

Sadoulet, Elizabeth, Alian de Janvry and Benjamin Davis, 2001, "Cash Transfer Programs with Income Multipliers: PROCAMPO in Mexico," *World Development* 29(6): 1043-56.

Schultz, T. Paul, 2004, "School Subsidies for the Poor: Evaluating the Mexican PROGRESA Poverty Program," *Journal of Development Economics*, 74(1): 199-250.

Smith, Jeffrey and Petra Todd, 2001, "Reconciling Conflicting Evidence on the Performance of Propensity-Score Matching Methods," *American Economic Review*, 91(2), 112-118.

_____ and _____, 2005a, "Does Matching Overcome LaLonde's Critique of Nonexperimental Estimators?" *Journal of Econometrics*, 125(1-2): 305-353.

_____ and _____, 2005b, "Rejoinder," *Journal of Econometrics*, 125(1-2): 365-375.

Thomas, Duncan, Elizabeth Frankenberg, Jed Friedman et al., 2003, "Iron Deficiency and the Well-Being of Older Adults: Early Results from a Randomized Nutrition Intervention," Paper Presented at the Population Association of America Annual Meetings, Minneapolis.

Van de Walle, Dominique, 2002, "Choosing Rural Road Investments to Help Reduce Poverty," *World Development* 30(4).

_____, 2004, "Testing Vietnam's Safety Net," *Journal of Comparative Economics*, 32(4): 661-679.

Van de Walle, Dominique, and Dorothy-Jean Cratty, 2005. "Do Aid Donors Get What they Want? Microevidence on Fungibility," Policy Research Working Paper 3542, World Bank.

Watts, H.W., 1968, "An Economic Definition of Poverty," in D.P. Moynihan (ed.), *On Understanding Poverty*. New York, Basic Books.

Weiss, Carol, 2001, "Theory-Based Evaluation: Theories of Change for Poverty Reduction Programs," in O. Feinstein and R. Piccioto (eds), *Evaluation and Poverty Reduction*, New Brunswick, NJ: Transaction Publications.

Woodbury, Stephen and Robert Spiegelman, 1987, "Bonuses to Workers and Employers to Reduce Unemployment," *American Economic Review*, 77, 513-530.

Wooldridge, Jeffrey, 2002, *Econometric Analysis of Cross-Section and Panel Data*, Cambridge, Mass.: MIT Press.