WPS3625

Evaluating Anti-Poverty Programs

Martin Ravallion[1]
Development Research Group, World Bank

World Bank Policy Research Working Paper 3625, June 2005

[1] These are the views of the author, and should not be attributed to the World Bank or any affiliated organization. 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, Ren Mu, Norbert Schady, Paul Schultz, John Strauss, Emmanuel Skoufias, Petra Todd, Dominique van de Walle and participants at a number of presentations at the World Bank and at an authors' 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 chapter 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 might ask for proposals from communities, with preference for those from poor areas; some areas do not apply, and some do, but are rejected.[2] 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 households deemed eligible by certain criteria.

After an overview of the archetypal formulation of the evaluation problem for such assigned programs in the following section, the bulk of the chapter examines the strengths and weaknesses of the main methods found in practice; examples are given throughout, mainly from 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, including in "scaling-up" development initiatives. The concluding section suggests some lessons for evaluation practice.

2. The archetypal evaluation problem

An "impact evaluation" assesses a program's performance in attaining well-defined objectives against an explicit counterfactual, such as the absence of the program. An observable outcome indicator, Y, is identified as relevant to the program and time-period over which impacts are expected. "Impact" is the change in Y that can be causally attributed to the program. The

[2] Social funds provide financial support to a potentially wide range of community-based projects, with strong emphasis given to local participation in proposing and implementing the specific projects.

data include an observation of Y_i for each unit *i* in a sample of size *n*. Treatment status, T_i, is observed, with $T_i = 1$ when unit *i* receives the program (is "treated") and $T_i = 0$ when not.[3]

The archetypal formulation of the evaluation problem postulates two potential outcomes for each *i*: the value of Y_i under treatment is denoted Y_i^T while it is Y_i^C under the counterfactual of not receiving treatment.[4] Then unit *i* gains $G_i \equiv Y_i^T - Y_i^C$. In the literature, G_i is variously termed the "gain", "impact" or the "causal effect" of the program for unit *i*.

In keeping with the bulk of the literature, this chapter will be mainly concerned with estimating average impacts. The most widely-used measure of average impact is the <u>average treatment effect on the treated</u>: $TT \equiv E(G|T=1)$. In the context of an anti-poverty program, *TT* is the mean impact on poverty amongst those who actually receive the program. One might also be interested in the average treatment effect on the un-treated, $TU \equiv E(G|T=0)$ and the combined average treatment effect (*ATE*):

$$ATE \equiv E(G) = TT \Pr(T=1) + TU \Pr(T=0)$$

We often want to know the <u>conditional</u> mean impacts, $TT(X) \equiv E(G|X, T=1)$, $TU(X) \equiv E(G|X, T=0)$ and $ATE(X) \equiv E(G|X)$, for a vector of covariates *X* (including unity as one element). The most common method of introducing *X* assumes that outcomes are linear in its parameters and the error terms (μ^T and μ^C), giving:

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

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

---

[3] The bio-medical connotations of the word "treatment" are unfortunate in the context of social policy, but the near-universal usage of this term in the evaluation literature makes it hard to avoid.

[4] This formulation of the evaluation problem in terms of potential outcomes in two possible states was proposed by Rubin (1974) (although with an antecedent in Roy, 1951). In the literature, Y_1 or $Y(1)$ and Y_0 or $Y(0)$ are more commonly used for Y^T and Y^C. My notation (following Holland, 1986) makes it easier to recall which group is which, particularly when I introduce time subscripts later.

We define the parameters β^T and β^C such that X is exogenous ($E(\mu^T|X) = E(\mu^C|X) = 0$).[5] The conditional mean impacts are then:

$$TT(X) = ATE(X) + E(\mu^T - \mu^C|X, T = 1)$$

$$TU(X) = ATE(X) + E(\mu^T - \mu^C|X, T = 0)$$

$$ATE(X) = X(\beta^T - \beta^C)$$

How can we estimate these impact parameters from the available data? The literature has long recognized that impact evaluation is essentially a problem of <u>missing data</u>, given that it is physically impossible to measure outcomes for someone in two states of nature at the same time (participating in a program and not participating). It is assumed that we can observe T_i, Y_i^T for $T_i = 1$ and Y_i^C for $T_i = 0$. But then G_i is not directly observable for any *i* since we are missing the data on Y_i^T for $T_i = 0$ and Y_i^C for $T_i = 1$. Nor are the mean impacts identified without further assumptions; neither $E(Y^C|T = 1)$ (as required for calculating *TT* and *ATE*) nor $E(Y^T|T = 0)$ (as needed for *TU* and *ATE*) is directly estimable from the data. Nor do equations (1.1) and (1.2) constitute an estimable model, given the missing data.

With the data that are likely to be available, an obvious place to start is the <u>single difference</u> (*D*) in mean outcomes between the participants and non-participants:

$$D(X) \equiv E[Y^T|X, T = 1] - E[Y^C|X, T = 0] \tag{2}$$

This can be estimated by the difference in the sample means or (equivalently) the Ordinary Least Squares (OLS) regression of *Y* on *T*. For the parametric model with controls, one would estimate (1.1) on the sub-sample of participants and (1.2) on the rest of the sample, giving:

$$Y_i^T = X_i\beta^T + \mu_i^T \text{ if } T_i = 1 \tag{3.1}$$

$$Y_i^C = X_i\beta^C + \mu_i^C \text{ if } T_i = 0 \tag{3.2}$$

[5] This is possible since we do not need to isolate the direct effects of *X* from those operating through omitted variables correlated with *X*.

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, giving a "random coefficients" specification:[6]

$$Y_i = X_i(\beta^T - \beta^C)T_i + X_i\beta^C + \varepsilon_i \quad (i=1,\ldots,n) \tag{4}$$

where $\varepsilon_i = T_i(\mu_i^T - \mu_i^C) + \mu_i^C$. A popular special case in practice is the <u>common-impact model</u>, which assumes that $G_i = ATE = TT = TU$ for all *i*, so that (4) collapses to:[7]

$$Y_i = ATE.T_i + X_i\beta^C + \mu_i^C \tag{5}$$

A less restrictive version only imposes the condition that the latent effects are the same for the two group (i.e., $\mu_i^T = \mu_i^C$), so that interaction effects with *X* remain.

While these are all reasonable starting points for an evaluation, and of obvious descriptive interest, further assumptions are needed to assure unbiased estimates of the impact parameters. To see why, consider the difference in mean outcomes between participants and non-participants (equation 2). This can be written as:

$$D(X) = TT(X) + B^{TT}(X) \tag{6}$$

where:[8]

$$B^{TT}(X) \equiv E[Y^C | X, T=1] - E[Y^C | X, T=0] \tag{7}$$

is the bias in using *D*(*X*) to estimate *TT*(*X*); B^{TT} is termed <u>selection bias</u> in much of the evaluation literature. Plainly, the difference in means (or OLS regression coefficient on *T*) only delivers the average treatment effect on the treated if counterfactual mean outcomes do not vary with treatment, i.e., $B^{TT} = 0$. In terms of the above parametric model, this is equivalent to assuming that $E[\mu^C | X, T=1] = E[\mu^C | X, T=0] = 0$, which assures that OLS gives consistent estimates of (5). If this also holds for μ^T then OLS will give consistent estimates of (4). I shall refer to the assumption that $E(\mu^C | X, T=t) = E(\mu^T | X, T=t) = 0$ for *t*=0,1 as "conditional exogeneity of placement." In the evaluation literature, this is also variously called "selection on

---

6 Equation (4) is derived from (3.1) and (3.2) using the identity: $Y_i = T_iY_i^T + (1-T_i)Y_i^C$.

7 The justification for this specialization of (4) is rarely obvious and (as we will see) some popular estimation methods for equation (5) are not robust to allowing for heterogeneity in impacts.

8 Similarly $B^{TU}(X) \equiv E(Y^T | X, T=1) - E(Y^T | X, T=0)$ and $B^{ATE}(X) = B^{TT}(X)\Pr(T=1) + B^{TU}(X)\Pr(T=0)$ in obvious notation.

observables," "unconfounded assignment" or "ignorable assignment," although the latter two terms usually refer to the stronger assumption that Y^T and Y^C are independent of T given X.

The rest of this chapter examines the estimation methods found in practice. One way to assure that $B^{TT} = 0$ is to randomize placement. Then we are dealing with an experimental evaluation, to be considered in section 4. By contrast, in a nonexperimental (NX) evaluation (also called an "observational study" or "quasi-experimental evaluation") the program is non-randomly placed.[9] NX methods fall into two groups, depending on which of two (non-nested) identifying assumptions is made. The first group assumes conditional exogeneity of placement, or the weaker assumption of exogeneity of changes in placement with respect to changes in outcomes. Sections 5 and 6 look at single-difference methods while section 7 turns to double- or triple-difference methods, which exploit data on changes in outcomes and placement, such as when we observe outcomes for both groups before and after program commencement.

The second set of NX methods does not assume conditional exogeneity (either in single-difference or higher-order differences). To remove selection bias based on unobserved factors these methods require some potentially strong assumptions. The main assumption found in applied work is that there exists an instrumental variable that does not alter outcomes conditional on participation (and other covariates of outcomes) but is nonetheless a covariate of participation. The instrumental variable thus isolates a part of the variation in program placement that can be treated as exogeneous. This method is discussed in section 8.

Some evaluators prefer to make one of these two identifying assumptions over the other. However, there is no sound *a priori* basis 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 its setting, what one wants to know about its impacts and (crucially) what data are available.

3. Generic issues in practice

The first problem often encountered in practice is getting the key stakeholders to agree to doing an impact evaluation. There may be vested interests that feel threatened, possibly including project staff. And there may be ethical objections. The most commonly heard objection to an impact evaluation says that if one finds a valid comparison group then this must

[9] As we will see later, experimental and NX methods are sometimes combined in practice, although the distinction is still useful for expository purposes.

include equally needy people to the participants, in which case the only ethically acceptable option is to help them, rather than just observe them passively for the purposes of an evaluation. Versions of this argument have stalled many evaluations in practice. Often, some kind of "top-down" political or bureaucratic force is needed; for example, state-level randomized trials of welfare reforms in the US in the 1980s and '90s were mandated by the federal government.

The ethical objections to impact evaluations are clearly more persuasive if eligible people have been knowingly denied the program for the purpose of the evaluation and the knowledge from that evaluation does not benefit them. However, the main reason in practice why valid comparison groups are possible is typically that fiscal resources are inadequate to cover everyone in need. While one might object to that fact, it is not an objection to the evaluation *per se*. Furthermore, knowledge about impacts can have great bearing on the resources available for fighting poverty. Poor people benefit from good evaluations, which weed out defective anti-poverty programs and identify good programs.

Having (hopefully) secured agreement to do the evaluation, a number of problems must then be addressed at the design stage, which this section reviews.

Is there selection bias? The assignment of an anti-poverty program typically involves purposive placement, reflecting the choices made by those eligible and the administrative assignment of opportunities to participate. It is likely that many of the factors that influence placement also influence counterfactual outcomes. Thus there must be a general presumption of selection bias when comparing outcomes between participants and non-participants.

In addressing this issue, it is important to consider both observable and unobservable factors. If the X's in the data capture the "non-ignorable" determinants of placement (i.e., those correlated with outcomes) then it can be treated as exogenous conditional on X. To assess the validity of that assumption one must know a lot about the specific program; conditional exogeneity should not be accepted, or rejected, without knowing how the program works in practice and what data are available.

In equations (3) and (4), the X's enter in a linear-in-parameters form. This is commonly assumed in applied work, but it is an *ad hoc* assumption, which is rarely justified by anything more than computational convenience (which is rather lame these days). Under the conditional exogeneity assumption, removing the selection bias is a matter of assuring that the X's are adequately balanced between treatment and comparison observations. When program placement

is independent of outcomes given the observables (implying conditional exogeneity) then the relevant summary statistic to be balanced between the two groups is the conditional probability of participation, called the propensity score. The propensity score plays an important role in a number of NX methods, as we will see in section 5.

The region of the propensity scores for which a valid comparison group can be found is termed the region of common support, as in Figure 1. Plainly, when this region is small it will be hard to identify the average treatment effect. This is a potentially serious problem in evaluating certain anti-poverty programs. To see why, suppose that placement is determined by a "proxy-means test" (PMT), as often used for targeting programs in developing countries. The PMT 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 PMT pass-score is non-decreasing in the budget under plausible conditions.) With 100% take-up, there is no value 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 sometimes called "failure of common support" in the evaluation literature.

This example is a rather extreme. In practice, there is usually some degree of fuzziness in the application of the PMT and incomplete coverage of those who pass the test. There is at least some overlap, but whether it is sufficient to infer impacts must be judged in each case.

Typically, we will have to truncate the sample of non-participants to assure common support; 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.

All is not lost when there is too little common support to credibly infer impacts on those treated. Indeed, knowing only the local impact in a neighborhood of the cut-off point may well be sufficient. Consider the policy choice of whether to increase a program's budget by raising the "pass mark" in the PMT. In this case, we only need know the impacts in a neighborhood of the pass-mark. Section 6 further discusses "discontinuity designs" for such cases.

So far we have focused on selection bias due to observable heterogeneity. However, it is almost never the case that the evaluator knows and measures all the relevant variables. Even controlling optimally for the X's by nicely balancing their values between treatment and comparison units will leave latent non-ignorable factors — unobserved by the evaluator but known to those deciding participation. Then we cannot attribute to the program the observed $D(X)$ (equation 2). The differences in conditional means could just be due to the fact that the program participants were purposely selected by a process that we do not fully observe. The impact estimator is biased in the amount given by equation (7). For example, suppose that the latent selection process discriminates against the poor, i.e., $E[Y^C | X, T = 1] > E[Y^C | X, T = 0]$ where Y is income relative to the poverty line. Then $D(X)$ will overestimate the impact of the program. A latent selection process favoring the poor will have the opposite effect. In terms of the classic parametric formulation of the evaluation problem in section 2, if participants have latent attributes that yield higher outcomes than non-participants (at given X) then the error terms in the equation for participants (3.1) will be centered to the right relative to those for non-participants (3.2). The error term in (4) will not vanish in expectation and OLS will give biased and inconsistent estimates. (Again, concerns about this source of bias cannot be separated from the question as to how well we have controlled for <u>observable</u> heterogeneity.)

A worrying possibility for applied work is that the two types of selection biases discussed above (one due to observables, the other due to unobservables) need not have the same sign. So eliminating selection bias due to one source need not reduce the total bias, which is what we care about. I do not know of an example from practice, but this theoretical possibility does point to the need to think about the likely <u>directions of the biases</u> in specific contexts, drawing on other evidence or theoretical models of the choices underlying program placement.

Is selection bias a serious concern in practice? Lalonde (1986) and Fraker and Maynard (1987) found large biases in various NX methods when compared to randomized evaluations of a U.S. training program. (Different NX methods also gave quite different results, but that is hardly surprising since they make different assumptions.) Similarly, Glewwe et al. (2004) found that NX 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 NX methods account for the difference. In a meta-study, Glazerman et al. (2003) review 12 replication studies of the impacts of training and employment programs on earnings; each study compared NX estimates

of impacts with results from a social experiment on the same program. They found large discrepancies in some cases, which they interpreted as being due to biases in the NX estimates.

Using a different approach to testing NX 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 villages. Only a seemingly small weight on village income in determining road placement was enough to severely bias the mean impact estimate.

Of course, one cannot reject NX 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 the presence of severe data problems it cannot be too surprising that observational studies perform poorly in correcting for selection bias. For example, in a persuasive 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, that the methods used had limited power for addressing selection bias and did not include adequate specification tests.[10] Heckman and Hotz (1989) argue that suitable specification tests can reveal the problematic NX methods in the Lalonde study, and that the methods that survive their tests give results quite close to those of the social experiment.

The 12 studies used by Glazerman et al. (2003) provided them with over 1,100 observations of paired estimates of impacts — one experimental and one NX. The authors then regressed the estimated biases on regressors describing the NX methods. They found that NX methods performed better (meaning that they came closer to the experimental result) when comparison groups were chosen carefully on the basis of observable differences (using regression, matching or a combination of the two). However, they also found that standard econometric methods for addressing selection bias due to unobservables using a control function and/or instrumental variable tended to <u>increase</u> the divergence between the two estimates.

These findings warn against presuming that more ambitious and seemingly sophisticated NX methods will perform better in reducing the total bias. The literature also points to the

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

importance of specification tests and critical scrutiny of the assumptions made by each estimator. This chapter will return to this point in the context of specific estimators.

Are there hidden impacts for "non-participants"? The classic formulation of the evaluation problem outlined in section 2 assumes no interference with the comparison units, which allows us to locate a program's impacts amongst only its direct participants. We observe the outcomes under treatment (Y_i^T) for participants ($T_i = 1$) and the counterfactual outcome (Y_i^C) for non-participants ($T_i = 0$). The comparison group is unaffected by the program.[11]

This can be a strong assumption in practice. Spillover effects to the comparison group are have been a prominent concern in evaluating large public programs, for which contamination of the control group can be hard to avoid due to the responses of markets and governments, and in drawing lessons for scaling up ("external validity") based on randomized trials (Mofitt, 2003).

To give a rather striking example in the context of antipoverty programs in developing countries, 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 Indian state of Maharashtra and in 2006 the Government of India implemented a national version of this scheme. The attractions of an EGS as a safety net stem from the fact that access to the program 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 data on both groups and compare their outcomes (after cleaning out observable heterogeneity). However, this classic evaluation design could give a severely biased result. The gains from such a program are very likely to spillover into the private labor market. If the employment guarantee is effective then the scheme will establish a firm lower bound to the entire wage distribution — assuming that 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, one will

[11] Rubin (1980) dubbed it the stable unit treatment value assumption (SUTVA).

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, which could be large for both groups.

To give another example, in assessing treatments for intestinal worms in children, Miguel and Kremer (2004) 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 design for the authors' own experiment neatly avoided this problem by using mass treatment at the school level instead of individual treatment (using control schools at sufficient distance from treatment schools).

Spillover effects can also arise from the behavior of governments. Indeed, whether the resources made available actually financed the identified project is often unclear. To some degree, all external aid is fungible. Yes, it can 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 under the counterfactual and that there is some other (infra-marginal) expenditure that is actually being financed by the external aid. Similarly, there is no way of ruling out the possibility that non-project villages benefited through 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 Mu (2006) in the context of a World Bank financed rural-roads project in Vietnam. Relative to the original plans, the project had only modest impact on its immediate objective, namely to rehabilitate existing roads. This stemmed in part from the fungibility of aid, although it turns out that there was a "flypaper effect" in that the aid stuck to the roads sector as a whole. Chen et al. (2006) also find evidence of a "geographic flypaper effect" for a poor-area development project in China.

How are outcomes for the poor to be measured? For anti-poverty programs the objective is typically defined in terms of household income or expenditure normalized by a household-specific poverty line (reflecting differences in the prices faced and in household size and composition). If we want to know the program's impact on poverty then we can set Y=1 for the "poor" versus Y=0 for the "non-poor."[12] That assessment will typically be based on a set of poverty lines, which aim to give the minimum income necessary for unit i to achieve a given reference utility, interpretable as the minimum "standard of living" needed to be judged non-

[12] Collapsing the information on living standards into a binary variable need not be the most efficient approach to measuring impacts on poverty; we return to this point.

poor. The normative reference utility level is typically anchored to the ability to achieve certain functionings, such as being adequately nourished, clothed and housed for normal physical activity and participation in society.[13]

With this interpretation of the outcome variable, *ATE* and *TT* give the program's impacts on the headcount index of poverty (% below the poverty line). By repeating the impact calculations 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.[14]

However, focusing on poverty impacts does not imply that we should use the constructed binary variable as the dependent variable (in regression equations such as (4) or (5), or nonlinear specifications such as a probit model). That entails an unnecessary loss of information relevant to explaining why some people are poor and others are not. Rather than collapsing the continuous welfare indicator (as given by income or expenditure normalized by the poverty line) into a binary variable at the outset it is probably better to exploit all the information available on the continuous variable, drawing out implications for poverty after the main analysis.[15]

What data are required? When embarking on any impact evaluation, it is obviously important to know the programs' objectives. More than one outcome indicator will often be identified. Consider, for example, a scheme that makes transfers targeted to poor families conditional on parents making human resource investments in their children.[16] The relevant outcomes comprise a measure of current poverty and measures of child schooling and health status, interpretable as indicators of future poverty.

---

[13] Note that the poverty lines will (in general) vary by location and according to the size and demographic composition of the household, and possibly other factors. On the theory and methods of setting poverty lines see Ravallion (2006).

[14] See Atkinson (1987) on the general form of these measures and examples in the literature.

[15] I have heard it argued a number of times that transforming the outcome measure into the binary variable and then using a logit or probit allows for a different model determining the living standards of the poor versus non-poor. This is not correct, since the underlying model in terms of the latent continuous variable is the same. Logit and probit are only appropriate estimators for that model if the continuous variable is unobserved, which is not the case here.

[16] The earliest program of this sort in a developing country appears to have been the *Food-for-Education* program (now called *Cash-for-Education*) introduced by the Government of Bangladesh in 1993. A famous example of this type of program is the *Program for Education, Health and Nutrition* (*PROGRESA*) (now called *Opportunidadas*) introduced by the Government of Mexico in 1997.

It is also important to know the salient administrative/institutional details of the program. For NX 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 sections 6 and 8.

The data on outcomes and their determinants, including program participation, typically come from sample surveys. The observation unit could be the individual, household, geographic area or facility (school or health clinic) depending on the type of program. Clearly the data collection must span the time period over which impacts are expected. The sample design is invariably important to both the precision of the impact estimates and how much can be learnt from the survey data about the determinants of impacts. Section 9 returns to this point.

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).[17] Integrating multiple data sources (such by unified geographic codes) can be highly desirable.

An important concern is the comparability of the data on participants and non-participants. Differences in survey design can entail differences in outcome measures. 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. Diaz and Handa (2004) come to a similar conclusion with respect to Mexico's *PROGRESA* program; they find that differences in the survey instrument generate significant biases in a propensity-score matching estimator (discussed further in section 5), although good approximations to the experimental results are achieved using the same survey instrument.

There are concerns about how well surveys measure the outcomes typically used in evaluating anti-poverty programs. Survey-based consumption and income aggregates for nationally representative samples typically do not match the aggregates obtained from national accounts (NA). This is to be expected for GDP, which includes non-household sources of domestic absorption. Possibly more surprising are the discrepancies found with both the levels and growth rates of private consumption in the NA aggregates (Ravallion. 2003b).[18] Yet here

---

[17] For excellent overviews of the generic issues in the collection and analysis of household survey data in developing countries see Deaton (1997).

[18] The extent of the discrepancy depends crucially on the type of survey (notably whether it collects consumption expenditures or incomes) and the region; see Ravallion (2003b).

too it should be noted that (as measured in practice) private consumption in the NA includes sizeable and rapidly growing components that are typically missing from surveys (Deaton, 2005). However, aside from differences in what is being measured, surveys encounter problems of under-reporting (particularly for incomes; the problem appears to be less serious for consumptions) and selective non-response (whereby the rich are less likely to respond).[19]

Survey measurement errors can to some extent be dealt with by the same methods used for addressing selection bias. For example, if the measurement problem affects the outcomes for treatment and comparison units identically (and additively) and is uncorrelated with the control variables then it will not be a problem for estimating *ATE*. This again points to the importance of the controls. But even if there are obvious omitted variables correlated with the measurement error, more reliable estimates may be possible using the double-difference estimators discussed further in section 7. This still requires that the measurement problem can be treated as a common (additive) error component, affecting measured outcomes for treatment and comparison units identically. These may, however, be overly strong assumptions in some applications.

It is sometimes desirable to collect panel data (also called longitudinal data), in which both participants and non-participants are surveyed repeatedly over time, spanning a period of expansion in program coverage and over which impacts are expected. Panel data raise new problems, including respondent attrition (another form of selection bias). Some of the methods described in section 7 do not strictly require panel data, but only observations of both outcomes and treatment status over multiple time periods, but not necessarily for the same observation units; these methods are thus more robust to the problems in collecting panel data.

As the above comments suggest, NX evaluations can be data demanding as well as methodologically difficult. One might be tempted to rely instead on "short cuts" including less formal, unstructured, interviews with participants. The problem in practice is that it is quite difficult to ask counter-factual questions in interviews or focus groups; try asking someone participating in a program: "what would you be doing now if this program did not exist?" Talking to participants (and non-participants) can be a valuable complement to quantitative surveys data, but it is unlikely to provide a credible impact evaluation on its own.

[19] On the implications of such selective survey compliance for measures of poverty and inequality and some evidence (for the US) see Korinek et al. (2006).

Sometimes it is also possible to obtain sufficiently accurate information on the past outcomes and program participation using respondent recall, although this can become quite unreliable, particularly over relatively long periods, depending on the variable and whether there are important memory "markers." Chen et al. (2006) demonstrate that 10-year recall by survey respondents in an impact evaluation is heavily biased toward more recent events.

4. Social experiments

A social experiment aims to randomize placement, such that all units (within some well-defined set) have the same chance *ex-ante* of receiving the program. Unconditional randomization is virtually inconceivable for anti-poverty programs, which policy makers are generally keen to target on the basis of observed characteristics, such as households with many dependents living in poor areas. However, it is sometimes feasible a program assignment that is partially randomized, conditional on some observed variables, X. The key implication for the evaluation is that all other (observed or unobserved) attributes prior to the intervention are then independent of whether or not a unit actually receives the program. By implication, $B^{TT}(X) = 0$, and so the observed *ex-post* difference in mean outcomes between the treatment and control groups is attributable to the program.[20] In terms of the parametric formulation of the evaluation problem in section 2, randomization guarantees that there is no sample selection bias in estimating (3.1) and (3.2) or (equivalently) that the error term in equation (4) is orthogonal to all regressors. The non-participants are then a valid control group for identifying the counterfactual, and mean impact is consistently estimated (nonparametrically) by the difference between the sample means of the observed values of Y_i^T and Y_i^C at given values of X_i.

Issues with social experiments: There has been much debate about whether randomization is the ideal method in practice.[21] Social experiments have often raised ethical objections and generated political sensitivities, particularly for governmental programs. (It is easier to do social experiments with NGOs, though for small interventions.) There is a perception that social experiments treat people like "guinea pigs," deliberately denying program access for some of those who need it (to form the control group) in favor of some who don't

[20] However, the simple difference in means is not necessarily the most efficient estimator; see Hirano et al. (2003).

[21] On the arguments for and against social experiments see (*inter alia*) Heckman and Smith (1995), Burtless (1995), Moffitt (2003) and Keane (2006).

(since a random assignment undoubtedly picks up some people who would not normally participate). In the case of anti-poverty programs, one ends up assessing impacts for 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 aim of fighting poverty.

These ethical and political concerns have stalled experiments or undermined their continued implementation. This appears to be why randomized trials for welfare reforms went out of favor with state governments in the US after the mid-1990s (Mofitt, 2003) and why subsequent evaluations of Mexico's *PROGRESA* program have turned to NX methods.

Are these legitimate concerns? As noted in section 3, the evaluation itself is rarely the reason for incomplete coverage of the poor in an anti-poverty program; rather it is that too few resources are available. When there are poor people who can't get on the program given the resources available, it has been argued that the ethical concerns favor social experiments; by this view, the fairest solution to rationing is to assign the program randomly, so that everyone has an equal opportunity of getting the limited resources available.[22]

However, it is hard to appreciate the "fairness" of an anti-poverty program that ignores available information on differences in the extent of deprivation. A key, but poorly understood, issue is what constitutes the "available information." As already noted, social experiments typically assign participation conditional on certain observables. But the things that are observable to the evaluator are generally a subset of those available to key stakeholders. The ethical concerns with experiments persist when it is known to <u>at least some</u> observers that the program is being withheld from those who clearly need it, and given to those who do not.

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. And their choices will undoubtedly be influenced by latent factors determining the returns to participation.[23] The extent of this problem depends on the specific program and setting; selective compliance is more likely for a training program (say) than a cash transfer program.

---

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

[23] The fact that people can select out of the randomized assignment goes some way toward alleviating the aforementioned ethical concerns about social experiments. But selective compliance clouds inferences about impact.

Sections 7 and 8 will return to this issue and discuss how NX methods can help address the problem.

The generic point is that the identification of impacts using social experiments is rarely "assumption-free." It is important to make explicit all the assumptions that are required, including about behavioral responses to the experiment; see, for example, the interesting discussion in Keane (2006) comparing experiments with structural modeling.

Recall that the responses of third parties can generate confounding spillovers (section 3). A higher level of government might adjust its own spending, counteracting the assignment. This is a potential problem whether the program is randomized or not, but it may well be a 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 potential problem, 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 points out that, 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 (Heckman and Smith, 1995). Section 9 returns to this topic.

Examples: Randomized trials for welfare programs and reforms were common in the US in the 1980s and early 1990s and much has been learnt from such trials (Moffitt, 2003). In the case of active labor market programs, 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 targeted wage subsidy programs in the US, randomized evaluations have been studied by Burtless (1985), Woodbury and Spiegelman (1987) and Dubin and Rivers (1993).

Another (rather different) example is the Moving to Opportunity (MTO) experiment, in which randomly chosen public-housing occupants in poor inner-city areas of five US cities were offered vouchers for buying housing elsewhere (Katz et al., 2001; Moffitt, 2001). This was motivated by the hypothesis that attributes of the area of residence matter to individual prospects of escaping poverty. The randomized assignment of MTO vouchers helps address some long-standing concerns about past NX tests for neighborhood effects (Manski, 1993).[24]

There have also been a number of social experiments 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 longevity of this program (surviving changes of government) and its influence in the development community clearly stem in part from the substantial, and public, effort that went into its evaluation. One third of the sampled communities deemed eligible for the program were chosen randomly to form a control group that did not get the program for an initial period during which the other two-thirds 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). A comprehensive overview of the design, implementation and results of the *PROGRESA* evaluation can be found in Skoufias (2005).

In another example for a developing country, Newman et al. (2002) were able to randomize eligibility to a World Bank supported social fund for a region of Bolivia. The fund-supported investments in education were found to have had significant impacts on school infrastructure but not on education outcomes within the evaluation period.

---

[24] Note that the design of the MTO experiment does not identify neighborhood effects at the origin, given that attributes of the destination also matter to outcomes (Moffitt, 2001).

Randomization was also used by Angrist et al. (2002) to evaluate a Colombian program that allocated schooling vouchers by a lottery. Three years later, the lottery winners had significantly lower incidence of grade repetition and higher test scores.

Another example is Argentina's *Proempleo* experiment (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 chapter to examine some lessons from this evaluation more closely.)

It has been argued that the World Bank should make greater use of social experiments. While the Bank has supported a number of experiments (including most 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), only one used randomization;[25] indeed, only 21 used any form of counterfactual analysis. Cook (2001) and Duflo and Kremer (2005) have advocated that OED should do many more social experiments.[26] Before accepting that advice one should be aware of some of the concerns raised by experiments.

A well-crafted social experiment will eliminate selection bias, but that leaves many other concerns about both their internal and external validity. The rest of this chapter turns to the main nonexperimental methods found in practice.

5. Propensity-score methods

As section 3 emphasized, selection bias is to be expected in comparing a random sample from the population of participants with a random sample of non-participants (as in estimating $D(X)$ in equation 2). 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 NX evaluations in practice provide too little information to assess whether the "comparison group" is likely to be sufficiently similar to the participants in the absence of the intervention.

---

[25] From Kapoor's description it is not clear that even this evaluation was a genuine experiment.

[26] OED only assesses Bank projects after they are completed, which makes it hard to do proper impact evaluations. Note that other units in the Bank that do evaluations besides OED, including in the research department invariably use counterfactual analysis and sometimes randomization.

In trying to find a comparison group it is natural to search for non-participants with similar pre-intervention characteristics to the participants. However, there are potentially many characteristics that one might use to match. How should they be weighted in choosing the comparison group? This section begins by reviewing the theory and practice of matching using propensity scores and then turns to other uses of propensity scores in evaluation.

Propensity-score matching: This method aims to select comparators according to their propensity scores, as given by $P(Z) = \Pr(T = 1|Z)$ (0< $P(Z)$<1), where Z is a vector of pre-exposure control variables (which can include pre-treatment values of the outcome indicator).[27] The values taken by Z are assumed to be unaffected by whether unit i actually receives the program. PSM uses $P(Z)$ (or a monotone function of $P(Z)$) to select comparison units. An important paper by Rosenbaum and Rubin (1983) showed that if outcomes are independent of participation given Z, then outcomes are also independent of participation given $P(Z_i)$.[28] The independence condition implies conditional exogeneity of placement ($B^{TT}(X) = 0$), so that the (unobserved) $E(Y^C|X, T = 1)$ can be replaced by the (observed) $E(Y^C|X, T = 0)$. Thus, as in a social experiment, TT is non-parametrically identified by the difference in the sample mean outcomes between treated units and the matched comparison group ($D(X)$). Under the independence assumption, exact matching on $P(Z)$ eliminates selection bias, although it is not necessarily the most efficient impact estimator (Hahn, 1998; Angrist and Hahn, 2004).

Thus 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 a social experiment) the impact estimates obtained by PSM must always depend on the variables used for matching and (hence) the quantity and quality of available data.

There is an important (often implicit) assumption in PSM and other NX methods 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 observables and that due to unobserved factors — go in the same direction, which cannot be assured on *a priori* grounds (as noted in section 3). If the selection bias based on unobservables counteracts that based on observables

---

[27] The present discussion is confined to a 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; Lechner, 2001; Hirano and Imbens, 2004).

[28] The result also requires that the T_i's are independent over all i. For a clear exposition and proof of the Rosenbaum-Rubin theorem see Imbens (2004).

then eliminating only the latter bias will increase aggregate bias. While this is possible in theory, replication studies (comparing NX evaluations with experiments for the same programs) do not appear to have found an example in practice; I review lessons from replication studies below.

The variables in Z may well differ from the covariates of outcomes (X in section 2); this distinction plays an important role in the methods discussed in section 8. But what should be included in Z?[29] The choice should be based on knowledge about the program and setting, as relevant to understanding the economic, social or political factors influencing program assignment that are correlated with counterfactual outcomes. Qualitative field work can help; for example, the specification choices made in Jalan and Ravallion (2003b) reflected interviews with participants and local administrators in Argentina's *Trabajar* program (a combination of workfare and social fund). Similarly Godtland et al. (2004) validated their choice of covariates for participation in an agricultural extension program in Peru through interviews with farmers.

Clearly if the available data do not include a determinant of participation relevant to outcomes then PSM will not have removed the selection bias (in other words it will not be able to reproduce the results of a social experiment). Knowledge of how the specific program works and theoretical considerations on likely behavioral responses can often reveal likely candidates for such an omitted variable. Under certain conditions, bounds can be established to a matching estimator, allowing for an omitted covariate of program placement (Rosenbaum, 1995; for an example see Aakvik, 2001). Later in this chapter we will consider alternative estimators that can be more robust to such an omitted variable (although requiring further assumptions).

Common practice in implementing PSM is to use the predicted values from a logit or probit as the propensity score for each observation in the participant and the non-participant samples, although non-parametric binary-response models can be used; see Heckman et al., (1997).[30] The comparison group can be 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

[29] For guidance on this and the many other issues that arise when implementing PSM see the useful paper by Caliendo and Kopeinig (2005).

[30] The participation regression is of interest in its own right, as it provides insights into the targeting performance of the program; see, for example, the discussion in Jalan and Ravallion (2003b).

not exceed some reasonable bound.[31] Given measurement errors, more robust estimates take the mean of the nearest (say) five neighbors, although this does not necessarily reduce bias.[32]

It is a good idea 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.

The typical PSM estimator for mean impact takes the form $\sum_{j=1}^{NT}(Y_j^T - \sum_{i=1}^{NC} W_{ij} Y_{ij}^C) / NT$ where *NT* is the number receiving the program, *NC* is the number of non-participants and the W_{ij}'s are the weights. There are several weighting schemes that have been used in the literature (see the overview in Caliendo and Kopeinig, 2005). These range from nearest-neighbor weights to non-parametric weights based on kernel functions of the differences in scores whereby all the 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.[33]

The statistical properties of matching estimators (in particular their asymptotic properties) are not as yet well understood. In practice, standard errors are typically derived by a bootstrapping method, although the appropriateness of this method is not evident in all cases. Abadie and Imbens (2006) examine the formal properties in large samples of nearest-k neighbor matching estimators (for which the standard bootstrapping method does not give valid standard errors) and provide a consistent estimator for the asymptotic standard error.

Mean impacts can also 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. On averaging this across different strata defined by pre-intervention incomes one can

---

[31] When treated units have been over-sampled (giving a "choice-based sample") and the weights are unknown one should instead match on the odds ratio, $P(Z)/(1- P(Z))$ (Heckman and Todd, 1995).

[32] 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.

[33] 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.

assess the incidence of impacts. In doing so, it is a good idea to test if propensity-scores (and even the *Z*'s themselves) are adequately balanced within strata (as well as in the aggregate), since there is a risk that one may be confusing matching errors with real effects.

Similarly 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. This is illustrated in Figure 2, for Argentina's *Trabajar* program. The figure gives the cumulative distribution function (CDF) (or "poverty incidence curve") showing how the headcount index of poverty (% below the poverty line) varies across a wide range of possible poverty lines (when that range covers all incomes we have the standard cumulative distribution function). The vertical line is a widely-used poverty line for Argentina. The figure also gives the estimated counter-factual CDF, after deducting the imputed income gains from the observed (post-intervention) incomes of all the sampled participants. Using a poverty line of $100 per month (for which about 20% of the national population is deemed poor) we see a 15 percentage point drop in the incidence of poverty amongst participants due to the program; this rises to 30 percentage points using poverty lines nearer the bottom of the distribution. We can also see the gain at each percentile of the distribution (looking horizontally) or the impact on the incidence of poverty at any given poverty line (looking vertically).[34]

In evaluating anti-poverty programs in developing countries, single-difference comparisons using PSM have the advantage that they do not require either randomization or baseline (pre-intervention) data. While this can be a huge advantage in practice, 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. In practice, one must always consider the possibility that there is a latent variable that jointly influences placement and outcomes, such that the mean of the latent influences on outcomes is different between treated and untreated units. This invalidates the key conditional independence assumption made by PSM. Whether this is a concern or not must be judged in the context of the application at hand; how much one is concerned about unobservables must depend, of course, on what data one has on the relevant observables. Section 7 will give an example of how far wrong the method can go with inadequate data on the joint covariates of participation and outcomes.

[34] On 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 see Ravallion (2003b).

How does PSM differ from other methods? In a social experiment (at least in its pure form), the propensity score is a constant, since everyone has the same probability of receiving treatment. Intuitively, what PSM tries to do is create the observational analogue of such an experiment in which everyone has the same probability of participation. The difference is that in PSM it is the conditional probability ($P(Z)$) that is intended to be 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. Hence there are always concerns about remaining selection bias in PSM estimates.

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 5). 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 (2003a) 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.

PSM also differs from standard regression methods with respect to the sample. In PSM one confines attention to the region of common support (Figure 1). Non-participants with a score lower than any participant are excluded. One may also want to restrict potential matches in other ways, depending on the setting. For example, one may want to restrict matches to being 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 to be exogenous to outcomes. In PSM one is looking instead for covariates of participation. It is clearly important that these include those variables that also matter to outcomes. However, variables with seemingly 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. However, other studies have reported large differences between OLS with controls for Z and PSM based on $P(Z)$ (Jalan and Ravallion, 2003a; van de Walle and Mu, 2006).

How well does PSM perform? Returning to the same data set used by the Lalonde (1986) study (described in section 3), Dehejia and Wahba (1999) found that PSM achieved a fairly good approximation — much better than the NX methods studied by Lalonde. It appears that the poor performance of the NX methods used by Lalonde stemmed in large part from the use of observational units outside the region of common support. 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.[35]

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 as long as the same survey instrument is used for measuring outcomes for the treatment and comparison groups. 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,

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

and of being able to control for employment history. The meta-study by Glazerman et al. (2003) finds that PSM is one of the NX methods that can significantly reduce bias, particularly when used in combination with other methods.

Other uses of propensity scores in evaluation: There are other evaluation methods that make use of the propensity score. These methods can have advantages over PSM although there have as yet been very few applications in developing countries.

While matching on propensity scores eliminates bias (under the conditional exogeneity assumption) this need not be the most efficient estimation method (Hahn, 1998). Rather than matching by estimated propensity scores, an alternative impact estimator has been proposed by Hirano et al. (2003). This method weights observation units by the inverses of a nonparametric estimate of the propensity scores. Hirano et al. show that this practice yields a fully efficient estimator for average treatment effects. Chen et al. (2006) and van de Walle and Mu (2006) provide examples in the context of evaluating the impacts on poverty of development projects.

Propensity scores can also be used in the context of more standard regression-based estimators. Suppose one simply added the estimated propensity score $\hat{P}(Z)$ to an OLS regression of the outcome variable on the treatment dummy variable, *T*. (One can also include an interaction effect between $\hat{P}(Z_i)$ and T_i.) Under the assumptions of PSM this will eliminate any omitted variable bias in having excluded *Z* from that regression, given that *Z* is independent of treatment given $P(Z)$.[36] However, this method does not have the non-parametric flexibility of PSM. Adding a suitable function of $\hat{P}(Z)$ to the outcome regression is an example of the "control function" (CF) approach, whereby under standard conditions (including exogeneity of *X* and *Z*) the selection bias term can be written as a function of $\hat{P}(Z)$.[37] Identification rests either on the nonlinearity of the CF in *Z* or the existence of one or more covariates of participation (the vector *Z*) that only affect outcomes *via* participation. Subject to essentially the same identification conditions, another option is to use $\hat{P}(Z)$ as the instrumental variable for program placement, as discussed further in section 8.

---

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

[37] Heckman and Robb (1985) provide a thorough discussion of this approach; also see the discussion in Heckman and Hotz (1989). On the relationship between CF and PSM see Heckman and Navarro-Lozano (2004) and Todd (2007). On the relationship between CF approaches and instrumental variables estimators (discussed further in section 8) see Vella and Verbeek (1999).

6. Exploiting program design

Nonexperimental 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. Or delays in the implementation of a program can also facilitate forming comparison groups, which can also help pick up some sources of latent heterogeneity. This section discusses these methods and some examples.

Discontinuity designs: Under certain conditions one can infer impacts from the differences in mean outcomes between units on either side of a critical cut-off point determining program eligibility. 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 $T_i = 1$ for $M_i \leq m$ and $T_i = 0$ otherwise. Examples include a proxy-means test that sets a maximum score for eligibility (section 3) and programs that confine eligibility within geographic boundaries. The impact estimator is $E(Y^T | M = m - \varepsilon) - E(Y^C | M = m + \varepsilon)$ 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, $P(M) = E(T|M)$, which is an increasing function of M 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 the discontinuity at *m* is in outcomes under treatment <u>not</u> outcomes under the counterfactual.[38] The existence of strict eligibility rules does not mean that this 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.

In a test of how well discontinuity designs perform in reducing selection bias, 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

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

design. The authors find that the discontinuity design gives good approximations for almost all outcome indicators.

The method is not without its drawbacks. It is assumed that the evaluator knows M_i and (hence) eligibility for the program. That will not always be the case. Consider (again) a means-tested transfer whereby the income of the participants is supposed to be below some pre-determined cut-off point. In a single cross-section survey, we observe post-program incomes for participants and incomes for non-participants, but typically we do not know income at the time the means test was actually applied. And if we were to estimate eligibility by subtracting the transfer payment from the observed income then we would be assuming (implicitly) exactly what we want to test: whether there was a behavioral response to the program. Retrospective questions on income at the time of the means test will help (though recognizing the possible biases), as would a baseline survey at or near the time of the test. A baseline survey can also 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 further in section 7.

Note also that a discontinuity design gives mean impact for a selected sample of the participants, while most other methods (such as social experiments and PSM) aim 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 other evaluations are also confined to a highly selected sub-sample; the question is then whether that is an interesting sub-sample. The truncation of treatment group samples to assure common support 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 include only those with the lowest probability. The latter sub-sample can, nonetheless, be relevant for deciding about program expansion; section 9 returns to this point.

Although impacts in a neighborhood of the cut-off point are non-parametrically identified for discontinuity designs, the applied literature 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 return to give examples in section 8.

Pipeline comparisons: The idea here is to use as the comparison group people who have applied for a program but not yet received it.[39] *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 *PROGRESA*, the pipeline comparison was randomized. NX pipeline comparisons have also been used in developing countries. An example can be found in Chase (2002) who used communities that 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. Notice that this method does to some extent address the problem of latent heterogeneity in other single-difference estimators, such as PSM; the prior selection process will tend to mean that successful applicants will tend to have similar unobserved characteristics, whether or not they have actually received the treatment.

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 financial crisis in Argentina when it was physically impossible to immediately help everyone who needed help. 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 example in the next section.

When feasible, pipeline comparisons offer a single-difference impact estimator that is likely to be more robust to latent heterogeneity. The estimates should, however, be tested for bias due to poorly balanced observables and (if need be) a method such as PSM can be used to deal with this prior to making the pipeline comparison (Galasso and Ravallion, 2004).

---

[39] This is sometimes called "pipeline matching" in the literature, although this term is less than ideal given that no matching is actually done.

Pipeline comparisons might also be combined with discontinuity designs. Although I have not seen it used in practice, a possible identification strategy for projects that expand along a well defined route is to measure outcomes on either side of the project's current frontier. Examples might include projects that progressively connect houses to an existing water, sanitation, transport or communications network, as well as projects that expand that network in discrete increments. New facilities (such as electrification or telecommunications) often expand along pre-existing infrastructure networks (such as roads, to lay cables along their right-of-way). Clearly one would also want to allow for observable heterogeneity and time effects. There may also be concerns about spillover effects; the behavior of non-participants may change, in anticipation of being hooked up to the expanding network.

7. Higher-order differences

So far the discussion has focused on single-difference estimators that only require a single survey. More can be learnt if we track outcomes for both participants and non-participants over time. A pre-intervention "baseline survey" in which one knows who eventually participates and who does not, can reveal specification problems in a single-difference estimator. If the outcome regression (such as equations 4 or 5) is correctly specified then running that regression on the baseline data should indicate an estimate of mean impact that is not significantly different from zero (Heckman and Hotz, 1989).

With baseline data one can go a step further and allow some of the latent determinants of outcomes to be correlated with program placement given the observables. This section begins with the double-difference (*DD*) method, which relaxes the conditional exogeneity assumption of single-difference NX estimators by exploiting a baseline and at least one follow-up survey post-intervention. The discussion then turns to situations — common for safety-net programs set-up to address a crisis — in which a baseline survey is impossible, but we can track ex-participants; this illustrates the triple-difference estimator.

The double-difference estimator: The essential idea is to compare samples of participants and non-participants before and after the intervention. After the initial baseline survey of both non-participants and (subsequent) participants, one does a follow-up survey of both groups after the intervention. Finally one calculates the difference between the "after" and "before" values of the mean outcomes for each of the treatment and comparison groups. The

difference between these two mean differences (hence the label "double difference" or "difference-in-difference") is the impact estimate.

To see what is involved, let Y_{it} denote the outcome measure for the *i*'th observation unit observed at two dates, *t*=0,1. By definition $Y_{it} = Y_{it}^C + T_{it}G_{it}$ and (as in the archetypal evaluation problem described in section 2), it is assumed that we can observe T_{it}, Y_{it}^T when $T_{it} = 1$, Y_{it}^C for $T_{it} = 0$, but that $G_{it} = Y_{it}^T - Y_{it}^C$ is not directly observable for any *i* (or in expectation) since we are missing the data on Y_{it}^T for $T_{it} = 0$ and Y_{it}^C for $T_{it} = 1$. To solve the missing-data problem, the *DD* estimator assumes that the selection bias (the unobserved difference in mean counterfactual outcomes between treated and untreated units) is time invariant, in which case the outcome changes for non-participants reveal the counterfactual outcome changes, i.e.:

$$E(Y_1^C - Y_0^C \mid T_1 = 1) = E(Y_1^C - Y_0^C \mid T_1 = 0) \tag{8}$$

This is clearly a weaker assumption than conditional exogeneity in single-difference estimates; $B_t^{TT} = 0$ for all *t* implies (8) but is not necessary for (8). Since period 0 is a baseline, with $T_{0i} = 0$ for all *i* (by definition), $Y_{0i} = Y_{0i}^C$ for all *i*. Then it is plain that the double-difference estimator gives the mean treatment effect on the treated for period 1:

$$DD = E(Y_1^T - Y_0^C \mid T_1 = 1) - E(Y_1^C - Y_0^C \mid T_1 = 0) = E(G_1 \mid T_1 = 1) \tag{9}$$

Notice that panel data are not necessary for calculating *DD*. All one needs is the set of four means that make up *DD*; the means need not be calculated for the same sample over time.

When the counterfactual means are time-invariant ($E[Y_1^C - Y_0^C | T_1 = 1] = 0$), equations (8) and (9) collapse to a <u>reflexive comparison</u> in which one only monitors outcomes for the treatment units. Unchanging mean outcomes for the counterfactual is 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).

For calculating standard errors and implementing weighted estimators it is convenient to use a regression estimator for *DD*. The data over both time periods and across treatment status are pooled and one runs the regression:

$$Y_{it} = \alpha + \beta T_{i1} t + \gamma T_{i1} + \delta t + \varepsilon_i \quad (t = 0,1;\ i = 1,\ldots,n) \tag{10}$$

The single-difference estimator is $SD_t \equiv E[(Y_{it}|T_{i1}=1)-(Y_{it}|T_{i1}=0)] = \beta t + \gamma$ while the *DD* estimator is $DD_1 \equiv SD_1 - SD_0 = \beta$. Thus the regression coefficient on the interaction effect between the participation dummy variable and time in equation (10) identifies the impact.

Notice that equation (10) does not require a balanced panel; for example, the interviews do not all have to be done at the same time. This property can be useful in survey design, by allowing "rolling survey" approach, whereby the survey teams move from one primary sampling unit to another over time; this has advantages in supervision and likely data quality. Another advantage of the fact that (10) does not require a balanced panel is that the results will be robust to selective attrition. In the case of a balanced panel, we can instead estimate the equivalent regression in the more familiar "fixed-effects" form:

$$Y_{it} = \alpha^* + \beta T_{i1} t + \delta_t + \eta_i + \nu_{it} \tag{11}$$

Here the fixed effect is $\eta_i = \gamma T_{i1} + \overline{\eta}^C + \mu_i$.[40]

Note that the term γT_{i1} in equation (10) picks up differences in the mean of the <u>latent</u> individual effects, such as would arise from initial selection into the program. The single-difference estimate will be biased unless the means of the latent effects are balanced between treated and non-treated units ($\overline{\eta}^T = \overline{\eta}^C$, i.e., $\gamma = 0$). This is implausible in general, as emphasized in section 2. The double-difference estimator removes this source of bias.

This approach can be readily generalized to multiple time periods; *DD* is then estimated by the regression of Y_{it} on the (individual and date-specific) participation dummy variable T_{it} interacted with time, and with individual and time effects. Or one can use a differenced specification in which the changes over time are regressed on T_{it} with time fixed effects.[41]

Examples of DD evaluations: In an early example, Binswanger et al. (1993) used this method to estimate the impacts of rural infrastructure on agricultural productivity in India, using district-level data. Their key identifying assumption was that the endogeneity problem — whereby infrastructure placement reflects omitted determinants of productivity — arose entirely

---

[40] Note that $\eta_i = \eta_i^T T_{i1} + \eta_i^C (1-T_{i1}) = \gamma T_{i1} + \overline{\eta}^C + \mu_i$ ($E(\eta_i|T_{i1}) \neq 0$)) where $\gamma = \overline{\eta}^T - \overline{\eta}^C$, $\mu_i = (\eta_i^T - \overline{\eta}^T)T_{i1} + (\eta_i^C - \overline{\eta}^C)(1-T_{i1})$, $E(\mu_i) = 0$, $\varepsilon_{it} = \nu_{it} + \mu_i$ and $\alpha = \alpha^* + \overline{\eta}^C$.

[41] As is well-known, when the differenced error term is serially correlated one must take account of this fact 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.

through latent agro-climatic factors that could be captured by district-level fixed effects. They found significant productivity gains from rural infrastructure.

In another example, Duflo (2001) estimated the impact on schooling and earnings in Indonesia of building schools. A 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 services through midwives on children's nutritional status (height-for-age), also in Indonesia.

Galiani et al. (2005) used a *DD* design to study the impact of privatizing water services on child mortality in Argentina, exploiting the joint geographic (across municipalities) and inter-temporal variation in both child mortality and ownership of water services. 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 4). An example can be found in Thomas et al. (2003) who randomized assignment of iron-supplementation pills in Indonesia, with a randomized-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 design for a DD estimator tracks the differences <u>over time</u> 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, <u>less</u> 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.[42] 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 — reveals 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.[43] 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 not GB but rather having land and hence becoming ineligible for GB. (I return to this example below.)

The use of different methods and data sets on the same program can be revealing. As compared to the study by Jalan and Ravallion (2002b) on the same program (Argentina's *Trabajar* program), Ravallion et al. (2005) used a lighter survey instrument, with far fewer questions on relevant characteristics of participants and non-participants. These data did not deliver plausible single-difference estimates using PSM when compared to the Jalan-Ravallion estimates 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). However, the lighter survey instrument used by Ravallion et al. (2005) had the advantage that the same households were followed up over time to form a panel data set. It would appear that Ravallion et al. were able to satisfactorily address the problem of bias in the lighter survey instrument by tracking households over time, which allowed them to difference-out the miss-matching errors arising from incomplete data.

---

[42] This is my interpretation; Pitt and Khandker (1998) do not mention the *DD* interpretation of their design. 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 households in eligible villages to obtain the impact parameter for GB.)

[43] 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.

This illustrates the trade-off between collecting cross-sectional data for the purpose of single-difference matching, versus collecting longitudinal data with a lighter survey instrument. An important factor in deciding which method to use is how much we know *ex ante* about the determinants of program placement. If a single survey can convincingly capture these determinants then PSM will work well; if not then one is well advised to do at least two rounds of data collection and use DD, possibly combined with PSM, as discussed below.

While panel data are not essential for estimating *DD*, household-level panel data open up further options for the 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. 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).

Panel data also facilitate the use of dynamic regression estimators for the *DD*. An example of this approach can be found in Jalan and Ravallion (2002), who identified the effects of lagged infrastructure endowments in a dynamic model of consumption growth using a six-year household panel data set. Their econometric specification is an example of the non-stationary fixed-effects model proposed by Holtz-Eakin et al. (1988), which allows for latent individual and geographic effects and can be estimated using the Generalized Method of Moments, treating lagged consumption growth and the time-varying regressors as endogenous (using sufficiently long lags as instrumental variables). The authors found significant longer-term consumption gains from improved infrastructure, such as better rural roads.

Concerns about DD designs: Two key problems have plagued DD estimators. The first is that, in practice, one sometimes does not know at the time the baseline survey is implemented who will participate in the program. One must make an informed guess in designing the

sampling for the baseline survey; knowledge of the program design and setting can provide clues. Types of observation units with characteristics making them more likely to participate will often have to be over-sampled, to help assure adequate coverage of the population treatment group and to provide a sufficiently large pool of similar comparators to draw upon. Problems can arise later if one does not predict well-enough *ex ante* who will participate. For example, Ravallion and Chen (2005) had designed their survey so that the comparison group would be drawn from randomly sampled villages in the same poor counties of rural China in which it was known that the treatment villages were to be found (for a poor-area development program). However, the authors subsequently discovered that there was sufficient heterogeneity within poor counties to mean that many of the selected comparison villages had to be dropped to assure common support. With the benefit of hindsight, greater effort should have been made to over-sample relatively poor villages within poor countries.

The second source of concern is the *DD* assumption of time-invariant selection bias. Infrastructure improvements may well be attracted to places with rising productivity, leading a geographic fixed-effects specification to overestimate the economic returns to new development projects. The opposite bias is also possible. Poor-area development programs are often targeted to places that lack infrastructure and other conditions conducive to economic growth. Again the endogeneity problem cannot be dealt with properly by positing a simple additive fixed effect. The selection bias is not constant over time and the *DD* will then be a biased impact estimator.

Figure 3 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 poorer 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 downward bias in the *DD* estimator, as in panel (c).

Two examples from actual evaluations illustrate the problem. Jalan and Ravallion (1998) show that poor-area development projects in rural China have been targeted to areas with poor infrastructure <u>and</u> that these same characteristics resulted in lower growth rates; presumably, areas with poor infrastructure were less able to participate in the opportunities created by China's growing economy. Jalan and Ravallion 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 is the Pitt and Khandker (1998) study of Grameen Bank. Following my interpretation above of their method of assessing the impacts of GB credit, it is clear that the 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 unusually high or low returns to land. It seems plausible that the returns to land are lower in villages selected for GB (which may be why they are poor in the first place) and low returns to land would also suggest that such villages have a comparative advantage in the non-farm activities facilitated by GB credit. Then the Pitt-Khandker method will overestimate GB's impact.

The upshot of these observations is that controlling for initial heterogeneity is crucial to the credibility of DD estimates. Using PSM for selecting the initial comparison group is an obvious corrective, and this will almost certainly reduce the bias in DD estimates. 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). The DD method can also be combined with a discontinuity design (Jacob and Lefgren, 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.

What if baseline data are unavailable? Anti-poverty programs in developing countries often have to be set up quickly in response to a macroeconomic or agro-climatic crisis; it is not feasible to delay the operation to do a baseline survey. (Needless-to-say, nor is randomization an option.) Even so, under certain conditions, impacts can still be identified by observing participants' outcomes in the absence of the program <u>after</u> the program rather than before it. To see what is involved, recall that the key identifying assumption in all double-difference studies is

that the selection bias into the program is additively separable from outcomes and time invariant. In the standard set-up described earlier in this section, date 0 precedes the intervention and *DD* gives the mean current gain to participants in date 1. However, suppose now that the program is in operation at date 0. The scope for identification arises from the fact that some participants at date 0 subsequently drop out of the program. The triple-difference (*DDD*) estimator proposed by Ravallion et al. (2005) is the difference between the double differences for stayers and leavers. Ravallion et al. show that their *DDD* estimator consistently identifies the mean gain to participants at date 1 (*TT*) if two conditions hold: (i) there is no selection bias in terms of who leaves the program and (ii) there are no current gains to non-participants. They also show that a third survey round allows a joint test of these two conditions. If these conditions hold and there is no selection bias in period 2, then there should be no difference in the estimate of gains to participants in period 1 according to whether or not they drop out in period 2.

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 they find evidence of a post-program "Ashenfelter's dip," namely when earnings drop sharply at retrenchment, but then recover.[44]

Suppose instead that we do not have a comparison group of nonparticipants; we calculate the *DD* 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 their specific setting, Ravallion et al. find that 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 may not hold in other applications.

[44] "Ashenfelter's dip" refers to the bias in using *DD* for inferring long-term impacts of training programs that can arise when there is a pre-program earnings dip (as was found in Ashenfelter, 1978).

8. Instrumental variables

The nonexperimental estimators discussed so far require some form of (conditional) exogeneity assumption for program placement. The single-difference methods assume that placement is uncorrelated with the latent determinants of outcome levels while the double-difference assumes that changes in placement are uncorrelated with the changes in these latent factors. We now turn to a popular method that relaxes these assumptions, but adds new ones.

The instrumental variables estimator (IVE): Returning to the archetypal model in section 2, the standard linear IVE makes two extra assumptions. The first is that there exists an instrumental variable (IV), denoted *Z*, which influences program placement independently of *X*:

$$T_i = \gamma Z_i + X_i \delta + \nu_i \quad (\gamma \neq 0) \tag{11}$$

(*Z* is exogeneous, as is *X*.) The second assumption is that impacts are homogeneous, in that outcomes respond identically across all units at given *X* ($\mu_i^T = \mu_i^C$ for all *i* in the archetypal model of section 2); a common special case in practice is the common-impact model:

$$Y_i = ATE.T_i + X_i \beta^C + \mu_i^C \tag{5}$$

We do not, however, assume conditional exogeneity of placement. Thus ν_i and μ_i^C are potentially correlated, inducing selection bias ($E(\mu^C | X, T) \neq 0$). But now there is a solution. Substituting (11) into (5) we obtain the reduced form equation for outcomes:

$$Y_i = \pi Z_i + X_i (\beta^C + ATE.\delta) + \mu_i \tag{12}$$

where $\pi = ATE\gamma$ and $\mu_i = ATE\nu_i + \mu_i^C$. Since OLS gives consistent estimates of both (11) and (12), the Instrumental Variables Estimator (IVE), $\hat{\pi}_{OLS} / \hat{\gamma}_{OLS}$, consistently estimates *ATE*.[45] The assumption that Z_i is not an element of X_i allows us to identify π in (12) separately from β^C. This is called the "exclusion restriction" (in that Z_i is excluded from (5)).

Strengths and weaknesses of the IVE method. The standard (linear) IVE method described above shares some of the weaknesses of other NX methods. As with OLS, the validity of causal inferences typically rests on *ad hoc* assumptions about the outcome regression, including its functional form. PSM, by contrast, is non-parametric in the outcome space.

[45] A variation is to re-write (11) as a nonlinear binary response model (such as a probit or logit) and use the predicted propensity score as the IV for program placement (Wooldridge, 2002, Chapter 18).

However, when a valid IV is available, the real strength of IVE over most other NX estimators is its robustness to the existence of unobserved variables that jointly influence program placement and outcomes.[46] This also means that (under its assumptions) IVE is less demanding of our ability to model the program's assignment than PSM. IVE gives a consistent estimate of *ATE* in the presence of omitted determinants of program placement. And if one has a valid IV then one can use this to test the exogeneity assumption of PSM or OLS.

This strength of IVE rests on the validity of its assumptions, and large biases in IVE can arise if they do not hold.[47] The best work in the IVE tradition gives close scrutiny to those assumptions. It is easy to test if $\gamma \neq 0$ in (11). The exclusion restriction is more difficult. If one has more than one valid IV then (under the other assumptions of IVE) one can do the standard over-identification test. However, one must still have at least one IV and so the exclusion restriction is fundamentally un-testable within the confines of the data available. Nonetheless, appeals to theoretical arguments or other evidence (external to the data used for the evaluation) can often leave one reasonably confident in accepting, or rejecting, a postulated exclusion restriction in the specific context.

Note that the exclusion restriction is not strictly required when a <u>nonlinear</u> binary response model is used for the first stage, instead of the linear model in (11). Then the impact 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. This is really a matter of judgment; identification off nonlinearity is still identification. Nonetheless, it is worrying when identification rests on an *ad hoc* assumption about functional form and the distribution of an error term.

---

[46] IVE is not the only NX method that relaxes conditional exogeneity. The control-function approach mentioned in section 5 also provides a method of addressing endogeneity; by adding a suitable control function (or "generalized residual") to the outcome regression one can eliminate the selection bias. Todd (2007) provides a useful overview of these approaches. In general, the CF approach should give similar results to IVE. Indeed, the two estimates are formally identical for a linear first-stage regression (as in equation 11), since then the control function approach amounts to running OLS on (5) augmented to include $\hat{v}_i = T_i - \hat{\gamma} Z_i$ as an additional regressor (Hausman, 1978). This CF removes the source of selection bias, arising from the fact that $Cov(\nu_i, \mu_i^C) \neq 0$.

[47] Recall that Glazeman et al. (2003) found that this type of method of correcting for selection bias tended in fact to be bias-increasing, when compared to experimental results on the same programs; they point to invalid exclusion restrictions as the likely culprit.

Heterogeneity in impacts. The common-impact assumption in (5) is not a harmless simplification, but is crucial to identifying mean impact using IVE (Heckman, 1997). To see why, return to the more general model in section 2, in which impact heterogeneity arises from differences between the latent factors in outcomes under treatment versus those under the counterfactual; write this as: $G_i = ATE + \mu_i^T - \mu_i^C$. Then (5) becomes:

$$Y_i = ATE.T_i + X_i\beta^C + [\mu_i^C + (\mu_i^T - \mu_i^C)T_i)] \tag{13}$$

where the term in [.] is the new error term. For IVE to consistently estimate *ATE* we now require that $Cov[Z,(\mu^T - \mu^C)T] = 0$ (on top of $\gamma \neq 0$ and $Cov(Z,\mu^C) = 0$). This will fail to hold if selection into the program is informed by the idiosyncratic differences in impact ($\mu^T - \mu^C$); likely "winners" will no doubt be attracted to a program, or be favored by the implementing agency. This is what Heckman et al. (2006) call "essential heterogeneity." (Note that this is no less of a concern in social experiments with randomized assignment but selective compliance.) To interpret the IVE as an estimate of ATE, we must assume that the relevant agents do not know μ^T or μ^C, or do not act on that information. These are strong assumptions.

With heterogeneous impacts, IVE identifies impact for a specific population sub-group, namely those induced to take up the program by the exogenous variation attributable to the IV.[48] This sub-group is rarely identified explicitly in IVE studies, so it remains worryingly unclear how one should interpret the estimated IVE. It is presumably the impact for someone, but who?

The local instrumental variables (LIV) estimator of Heckman and Vytlacil (2005) directly addresses this issue. LIV entails a nonparametric regression of outcomes *Y* on the propensity score, $P(X,Z)$.[49] Intuitively, the slope of this regression function gives the impact at the specific values of *X* and *Z*; in fact this slope is the marginal treatment effect introduced by Björklund and Moffitt (1987), from which any of the standard impact parameters can be calculated using appropriate weights (Heckman and Vytlacil, 2005). To see whether impact heterogeneity is an issue in practice, Heckman et al. (2006) recommend that one should first test for linearity of *Y* in the propensity score. (A standard functional form test, such as RESET,

---

[48] The outcome gain for this sub-group is sometimes called the "local average treatment effect" (LATE) (Imbens and Angrist, 1994). Also see the discussion of LATE in Heckman (1997).

[49] Linear controls *X* can be included, making this a "partial linear model"(for a continuous propensity score); see Yatchew (1998) on parlial linear models and other topics in nonparametric regression analysis.

would be appropriate.) If nonlinearity is indicated then LIV is the appropriate estimator; but if not then the standard IVE is justified; indeed, the OLS regression coefficient of Y on the propensity score $P(X,Z)$ directly gives *ATE* in this case (Heckman, 1997).

Bounds on impact: In practice, IVE sometimes gives seemingly implausible impact estimates (either too small or too large). One might suspect that a violation of the exclusion restriction is the reason. But how can we form judgments about this issue in a more scientific way? If it is possible to rule out certain values for Y on *a priori* grounds then this can allow us to establish plausible bounds to the impact estimates (following an approach introduced by Manski, 1990). This is easily done if the outcome variable is being "poor" versus "non-poor" (or some other binary outcome). Then $0 \le TT \le E(Y^T | T=1)(\le 1)$ and:[50]

$$\begin{aligned}&(E[Y^T|T=1]-1)\Pr(T=1)-E[Y^C|T=0]\Pr(T=0)\\&\le ATE \le\\&(1-E[Y^C|T=0])\Pr(T=0)+E[Y^T|T=1]\Pr(T=1)\end{aligned}$$

The width of these bounds will (of course) depend on the specifics of the setting. The bounds may not be of much use in the (common) case of continuous outcome variables.

Another approach to setting bounds has been proposed by Altonji, Elder and Taber (AET) (2005a,b). The authors recognize the likely bias in OLS for the relationship of interest (in their case probably overestimating the true impact), but they also question the exclusion restrictions used in past IV estimates. Recall that OLS assumes that the unobservables affecting outcomes are uncorrelated with program placement. AET study the implications of the extreme alternative assumption: that the unobservables in outcomes have the same effect on placement as does the index of the observables (the term $X_i\beta^C$ in (5)); in other words, the selection on unobservables is assumed to be as great as that for the observables.[51] Implementing this assumption requires constraining the correlation coefficient between the error terms of the equations for outcomes and participation (μ^C in (5) and ν in (11)) to a value given by the

---

[50] The lower bound for *ATE* is found by setting $E[Y^T|T=0]=0$ and $E[Y^C|T=1]=1$ while the upper bound is found at $E[Y^T|T=0]=1, E[Y^C|T=1]=0$.

[51] Altonji et al. (2005a) gives conditions under which this will hold. However (as they note) these conditions are not expected to hold in practice; their estimator provides a bound to the true estimate, rather than an alternative point estimate.

regression coefficient of the score function for observables in the participation equation ($X_i\delta$ in equation (11) with $\gamma = 0$) on the corresponding score function for outcomes ($X_i\beta^C$).

AET argue that their estimator is a lower bound to the true impact when the latter is positive; this rests on the (*a priori* reasonable) presumption that the error term in the outcomes equation includes at least some factors that are truly uncorrelated with participation. OLS provides an upper bound. Thus, the AET estimator gives a useful indication of how sensitive OLS is to any selection bias based on unobservables. Altonji et al. (2005b) also show how their method can be used to assess the potential bias in IVE due to an invalid exclusion restriction. One would question an IVE that was outside the interval spanning the AET and OLS estimators.

Examples if IVE in practice. There are two main sources of IVs, namely experimental design features and *a priori* arguments about the determinants of program placement and outcomes. The following discussion considers examples of each.

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. So there is a problem that actual participation is endogenous. However, the randomized assignment provides a natural IV in this case (Dubin and Rivers, 1993; Angrist et al., 1996; the latter paper provides a formal justification for using IVE for this problem). The exclusion restriction is that being randomly assigned to the program only affects outcomes via actual program participation

An example is found in the aforementioned MTO experiment, in which randomly-selected inner-city families in US cities were given vouchers to buy housing in better-off areas. Not everyone offered a voucher took up the opportunity. The difference in outcomes (such as school drop-out rates) only reveal the extent of the external (neighborhood) effect if one corrects for the endogenous take-up using the randomized assignment as the IV (Katz et al., 2001).

An example for a developing country is the *Proempleo* experiment. Recall that this included a randomly-assigned training component. Under the assumption of perfect take-up or random non-compliance, neither the employment nor incomes of those receiving the training were significantly different to those of the control group 18 months after the experiment began.[52] However, some of those assigned the training component did not want it, and this selection

[52] 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.

process was correlated with the outcomes from training. An impact of training was revealed for those with secondary schooling, but only when the authors corrected for compliance bias using assignment as the IV for treatment (Galasso et al., 2004).

Randomized outreach (often called an "encouragement design") can also provide a valid IV.[53] The idea is that, for the purpose of the evaluation, one disseminates extra information/publicity on the (non-randomly placed) program to a random sample. One expects this to be correlated with program take-up and the exclusion restriction is also plausible.

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 9.

The bulk of the applications of IVE have used nonexperimental IVs. In the literature in labor economics, wage regressions often allow for endogenous labor-force participation (and hence selection of those observed to have wages). A common source of IVs is found in modeling the choice problem, whereby it is postulated that there are variables that influence the costs and benefits of labor-force participation but do not affect earnings given that choice; there is a large literature on such applications of IVE and related control function estimators.[54]

The validity of such exclusion restrictions can be questioned. For 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. These characteristics influence take-up of the program but are unlikely to be directly observable to employers; on this basis it is argued that they 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, this exclusion restriction is questionable when there are productivity-relevant spillover effects within households. For example, in developing-country settings it has been argued that the presence of a literate person in the household can exercise a strong effect on an illiterate worker's productivity; this is argued in theory and with supporting evidence (for Bangladesh) in Basu et al. (2002).

---

[53] Useful discussions of encouragement designs can be found in Bradlow (1998) and Hirano et al. (2000). I do not know of any examples for anti-poverty programs in developing countries.

[54] For an excellent overview see Heckman et al. (1999).

In evaluating anti-poverty programs in developing countries, three popular sources of instrumental variables have been the geographic placement of programs, political variables and discontinuities created by program design.[55] I consider examples of each.

The geography of program placement has been used for identification in a number of studies. I discuss three examples. Ravallion and Wodon (2000) test the widely heard claim that child labor displaces schooling and so perpetuates poverty in the longer-term. They used the presence of a targeted school enrollment-subsidy in rural Bangladesh (the *Food-for-Education Program*) as the source of a change in the price of schooling in their model of schooling and child labor. To address the endogeneity of placement at the individual level they used prior placement at the village level as the IV. The worry here is the possibility that village placement is correlated with geographic factors relevant to outcomes. Drawing on external information on the administrative assignment rules, Ravallion and Wodon provide exogeneity tests that offer some support their identification strategy, although this ultimately rests on an un-testable exclusion restriction and/or nonlinearity for identification. Their results indicate that the subsidy increased schooling by far more than it reduced child labor. Substitution effects appear to have helped protect current incomes from the higher school attendance induced by the subsidy.

A second 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. These authors also address the objections that can be raised against the exclusion restriction.[56] 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).

---

[55] Natural events (twins, birth dates, rainfall) have also been used as IVs in a number of studies (for an excellent review see Rosenzweig and Wolpin, 2000). These are less common as IVs for poverty programs, although we consider one example (for infrastructure).

[56] As in the Ravallion-Wodon example, the other main requirement of a valid IV, namely that it is correlated with treatment, is more easily satisfied in this case.

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 be affected, such as child birth weight. However, they do not find such effects, supporting the choice of IV.

The geography of program placement sometimes involves natural, topographic or agro-climatic, features that can aid identification. An example is provided by the Duflo and Pande (2005) study of the district-level poverty impacts of dam construction in India. To address the likely endogeneity of dam placement they exploit the fact that the distribution of land by gradient affects the suitability of a district for dams (with more positive gradients making a dam less costly).[57] Their key identifying assumption is that gradient does not have an independent effect on outcomes. To assess whether that is a plausible assumption one needs to know more about other possible implications of differences in land gradient; for example, for most crops land gradient matters to productivity (positively for some negatively for others), which would invalidate the IV. The authors provide little defense for their exclusion restriction. They find that a dam increases poverty in its vicinity but helps the poor downstream; on balance their results suggest that dams are poverty increasing.

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. The authors assume that female law makers favor workers' compensation but that this does not have an independent effect on the labor market. The latter condition would fail to hold if a higher incidence of women in parliament in a given state reflected latent social factors that lead to higher female labor force participation generally, with implications for aggregate labor market outcomes of both men and women.

To give another example, 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 social fund spending in Peru, when modeling schooling outcomes. Their idea was that the geographic allocation of 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

[57] Note that what the authors refer to as "river gradient" is based on the distribution of land gradients in an area that includes a river.

hypothesis.) It must also be assumed that the fact that an area turned against the government in the last election is not correlated with latent factors influencing schooling. The variation in spending attributed to this IV was found to significantly increase school attendance rates.

The third set of examples exploit discontinuities in program design, as discussed in section 6.[58] Here the impact estimator 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 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 used eligibility as the IV for receipt of a pension in her regressions for anthropometric outcome variables. Duflo found 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.

Again, this assumes we know eligibility, which is not always the case. Furthermore, eligibility for anti-poverty programs is often based on poverty criteria, which are also the relevant outcome variables. Then one must be careful not to make assumptions in estimating who is eligible (for constructing the IV) that pre-judge the impacts of the program.

The use of the discontinuity in the eligibility rule as an IV for actual program placement can also address concerns about selective compliance with those rules; this is discussed further in Battistin and Rettore (2002).

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

---

[58] Using discontinuities in IVE 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.

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. While the exclusion restriction is ultimately untestable, some studies do a better job than others of justifying in their specific case. Almost all applied work has assumed homogeneous impacts, despite the repeated warnings of Heckman and others. Relaxing this assumption in specific applications appears to be a fertile ground for future research.

9. Learning from evaluations

So far we have focused on the "internal validity" question: does the evaluation design allow us to obtain a reliable estimate of the counterfactual outcomes in the specific context? This has been the main focus of the literature to date. However, there are equally important concerns related to what can be learnt from an impact evaluation beyond its specific setting. This section turns to the "external validity" question as to whether the results from specific evaluations can be applied in other settings (places and/or dates) and what lessons can be drawn for development knowledge and future policy from evaluative research.

Do publishing biases inhibit learning from evaluations? Development policy-making draws on accumulated knowledge built up from published evaluations. Thus publication processes and the incentives facing researchers are relevant to our success against poverty and in achieving other development goals. It would not be too surprising to find that it is harder to publish a paper that reports unexpected or ambiguous impacts, when judged against received theories and/or past evidence. Reviewers and editors may well apply different standards in judging data and methods according to whether they believe the results on *a priori* grounds. To the extent that impacts are generally expected from anti-poverty programs (for that is presumably the main reason why the programs exist) this will mean that our knowledge is biased in favor of positive impacts. In exploring a new type of program, the results of the early studies will set the priors against which later work is judged. An initial bad draw from the true distribution of impacts may then distort knowledge for some time after. Such biases would no doubt affect the production of evaluative research as well as publications; researchers may well work harder to obtain positive findings to improve their chances of getting their work published. No doubt, extreme biases (in either direction) will be eventually exposed, but this may take some time.

These are largely conjectures on my part. Rigorous testing requires some way of inferring the counterfactual distribution of impacts, in the absence of publication biases. Clearly this is difficult. However, there is at least one strand of evaluative research where publication bias is unlikely, namely replication studies that have compared NX results with experimental findings for the same programs (as in the meta-study for labor programs in developed countries by Glazerman et al. 2003). Comparing the distribution of published impact estimates from (non-replication) NX studies with a counterfactual drawn from replication studies of the same type of programs could throw useful light on the extent of publication bias.

Can the lessons from an evaluation be scaled up? The context of an intervention often matters to its outcomes, thus confounding inferences for "scaling up" from an impact evaluation. Such "site effects" arise whenever aspects of a program's setting (geographic or institutional) interact with the treatment. The same program works well in one village but fail hopelessly in another. For example, in studying Bangladesh's *Food for Education* Program, Galassso and Ravallion (2005) found that the program worked well in reaching the poor in some villages but not in others, even in relatively close proximity. Site effects clearly make it difficult to draw valid inferences for scaling up and replication based on trials.

The local institutional context of an intervention is likely to be relevant 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 trials are tied to the activities of specific Non-Governmental Organizations (NGOs) as the facilitators, 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 external-validity concern is that, while partial equilibrium assumptions may be fine for a pilot, general equilibrium effects (sometimes called "feedback" or "macro" effects in the evaluation literature) can be important when it is scaled up nationally. 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.[59] 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 (2004) 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 becomes more pro-poor with scaling up. Trial results may well underestimate how pro-poor a program is likely to be after scaling up because the political economy entails that the initial benefits tend to be captured more by the non-poor (Lanjouw and Ravallion, 1999).

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).

[59] Heckman et al., (1998) demonstrate that partial equilibrium analysis can greatly overestimate the impact of a tuition subsidy once relative wages adjust; Lee (2005) finds a much smaller difference between the general and partial equilibrium effects of a tuition subsidy in a slightly different model.

Repeating evaluations across different settings and at different scales can help address these concerns, although it will not always be feasible to do a sufficient number of trials to span the relevant domain of variation found in reality. The scale of a randomized trial needed to test a large national program could be prohibitive. Nonetheless, varying contexts for trials is clearly a good idea, subject to feasibility. The failure to systematically plan locational variation into their design has been identified as a serious weakness in the randomized field trials that have been done of welfare reforms in the US (Moffitt, 2003).

Important clues to understanding impacts can often be found in the geographic differences in impacts and how this relates to the characteristics of locations, such as the extent of local inequality, "social capital" and the effectiveness of local public agencies; again, see the example in Galasso and Ravallion (2005). This calls for adequate samples at (say) village level. However, this leads to a further problem. For a given aggregate sample size, larger samples at local level will increase the design effect on the standard error of the overall impact estimate.[60] Evaluations can thus face a serious trade off between the need for precision in estimating overall impacts and the ability to measure and explain the underlying heterogeneity in impacts. Small-area estimation methods can sometimes improve the trade off by exploiting Census (or larger sample survey) data on covariates of the relevant outcomes and/or explanatory variables at local level; a good example in the present context can be found in Araujo et al (2006).

Another lens for understanding impacts is to study what can be called "intermediate" outcome measures. The typical evaluation design identifies a small number of "final outcome" indicators, and aims to assess the program's impact on those indicators. 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 were partly invested, with second-round effects on future incomes (Sadoulet et al., 2001). Similarly, Ravallion and Chen (2005) found that participants in a poor-area development program in China saved a large share of the income gains from the program (as estimated using the matched

[60] The design effect (*DE*) is the ratio of the actual variance (for a given variable in the specific survey design) to the variance in a simple random sample; this is given by $DE = 1 + \rho(B-1)$ where ρ is the intra-cluster correlation coefficient and *B* is the cluster sample size (Kish, 1965, Chapter 5).

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.

This also illustrates a common concern in evaluation studies, given behavioral responses, namely that 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 information external to the evaluation. Qualitative research (intensive interviews with participants and administrators) can be a useful source of information on the underlying processes determining outcomes.[61] One approach is to use such methods to test the assumptions made by an intervention; this has been called "theory-based evaluation," although that is hardly an ideal term given that NX 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 (2005) tested the assumption that a SF works by

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

local communities collectively proposing 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. For these reasons, the so-called "theory-based evaluation" approach cannot be considered a serious substitute for assessing impacts on final outcomes by credible (experimental or NX) methods, although it can still 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 locally). 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 chapter 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, 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.[62] These simulation methods make it far easier to attribute impacts to the policy change, although 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.

---

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

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, though this is rare in practice. 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.[63]

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*. However, as noted in section 3, there is nothing to stop us re-interpreting the outcome measure such that equation (2) gives the program's impact on the headcount index of poverty (% below the poverty line). By repeating the impact calculation for multiple "poverty lines" one can then trace out the impact on the cumulative distribution of income. This is feasible with the same tools, though evaluation practice has been rather narrow in its focus.

There is often interest in better understanding the <u>horizontal impacts</u> of program, meaning the differences in impacts at a given level of counterfactual outcomes, as revealed by the joint distribution of Y^T and Y^C. We cannot know this from a social experiment, which only

[63] For another example of the same result see Murgai and Ravallion (2005).

reveals net counterfactual mean outcomes for those treated; *TT* gives the mean gain net of losses amongst participants. 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. We already discussed an example in section 7, namely the use of panel data in studying impacts of an anti-poverty program on poverty dynamics. 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).

Heterogeneity in the impacts of anti-poverty programs can be expected. Eligibility criteria impose differential costs on participants. For example, the foregone labor earnings incurred by participants in workfare or conditional cash transfer schemes (via the loss of earnings from child labor) will vary according to skills and local labor-market conditions. Knowing more about this heterogeneity is relevant to the political economy of anti-poverty policies, and may also point to the need for supplementary policies for better protecting the losers.

Heterogeneity of impacts in terms of observables is readily allowed for by adding interaction effects with the treatment dummy variable, as in equation (5.1), though this is still surprisingly far from universal practice. One can also allow for latent heterogeneity, using a random coefficients estimator in which the impact estimate (the coefficient on the treatment dummy variable) contains a stochastic component (i.e., $\mu_i^T \neq \mu_i^C$ in the error term of equation 4). Applying this type of estimator to the evaluation data for *PROGRESA*, Djebbari and Smith (2005) find that they can convincingly reject the common effects assumption in past evaluations. When there is such heterogeneity, one will often want to distinguish marginal impacts from average impacts. Following Björklund and Moffitt (1987), the marginal treatment effect can be defined as the mean gain to units that are indifferent between participating or not. This requires that we model explicitly the choice problem facing participants (Björklund and Moffitt, 1987; Heckman and Navarro-Lozano, 2004). We may also want to estimate the joint distribution of Y^T and Y^C, and a method for doing so is outlined in Heckman et al. (1997a).

However, it is questionable how relevant the choice models found in this literature are to the present setting. The models have stemmed mainly from the literature on evaluating training and other programs in developed countries, in which selection is seen as a largely a matter of

individual choice, amongst those eligible. This approach does not sit easily with what we know about many anti-poverty programs in developing countries, in which the choices made by politicians and administrators appear to be at least as important to the selection process as the choices made by those eligible to participate.

This speaks to the need for a richer theoretical characterization of the selection problem in future work. An example of one effort in this direction can be found in the Galasso and Ravallion (2005) model of the assignment of a decentralized anti-poverty program; their model focuses on the public-choice problem facing the central government and the local collective action problem facing communities, with individual participation choices treated as a trivial problem. Such models can also point to instrumental variables for identifying impacts and studying their heterogeneity.

When the policy issue is whether to expand a given program at the margin, the classic estimator of mean-impact on the treated is actually of rather little interest. The problem of estimating the marginal impact of a greater duration of exposure to the program on those treated was considered in section 7, 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 6 (in its non-parametric form) and section 8 (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.[64] 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

[64] For a useful overview of *ex ante* methods see Bourguignon and Ferreira (2003). Todd and Wolpin (2006) provide a number of examples, including for a schooling subsidy program, using the PROGRESA data.

with a structural model of education choices and exploiting the randomized design for identification, one can greatly expand the set of policy-relevant questions about the design of *PROGRESA* that a conventional evaluation can answer (Todd and Wolpin, 2002; Attanasio et al., 2004, and de Janvry and Sadoulet, 2006). This strand of the literature has revealed that a budget-neutral switch of the enrolment subsidy from primary to secondary school would have delivered a net gain in school attainments, by increasing the proportion of children who continue onto secondary school. While *PROGRESA* had an impact on schooling, it could have had a larger impact. However, it should be recalled that this type of program has two objectives: increasing schooling (reducing future poverty) and reducing current poverty, through the targeted transfers. To the extent that refocusing the subsidies on secondary schooling would reduce the impact on current income poverty (by increasing the forgone income from children's employment), the case for this change in the program's design would need further analysis.

10. 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 mean impact, it is neither necessary nor sufficient for a good evaluation, and nor is it always feasible, notably for large public programs. While economists have sometimes been too uncritical of their nonexperimental 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. This may involve randomizing some aspects and using econometric methods to deal with the non-random elements, or by using randomized elements of a program as a source of instrumental variables. Likely biases in specific non-experimental methods can also be reduced by combining with other methods. For example, depending on the application at hand (including the data available and its quality), single-difference matching methods can be vulnerable to biases stemming from latent non-ignorable factors, while standard double-difference estimators are vulnerable to biases arising from the way differing initial conditions can influence the subsequent outcome changes over time (creating a time-varying selection bias). With adequate data, combining matching (or weighting) using propensity scores with double-difference methods can reduce the biases in each method.

(In other words, conditional exogeneity of placement with respect to changes in outcomes will often be a more plausible assumption than conditional exogeneity in levels.) Data quality is key to all these methods, as is knowledge of the program and context. Good evaluations typically require that the evaluator is involved from the program's 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, the standard tools of counter-factual analysis for mean impacts can be seen to have some severe limitations for informing development policy making. We have learnt that the context in which a program is placed and the characteristics of the participants can exercise a powerful influence on outcomes. And not all of this heterogeneity in impacts can readily be attributed to observables, which greatly clouds the policy interpretation of standard methods. We need a deeper understanding of this heterogeneity in impacts; this can be helped by systematic replications across differing contexts and the new econometric tools that are available for identifying local impacts. The assumptions made in a program's design also need close scrutiny, such as by tracking intermediate variables of relevance or by drawing on supplementary theories or evidence external to the evaluation. In drawing lessons for anti-poverty policy, we also need a richer set of impact parameters than has been traditional in evaluation practice, including distinguishing the gainers from the 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

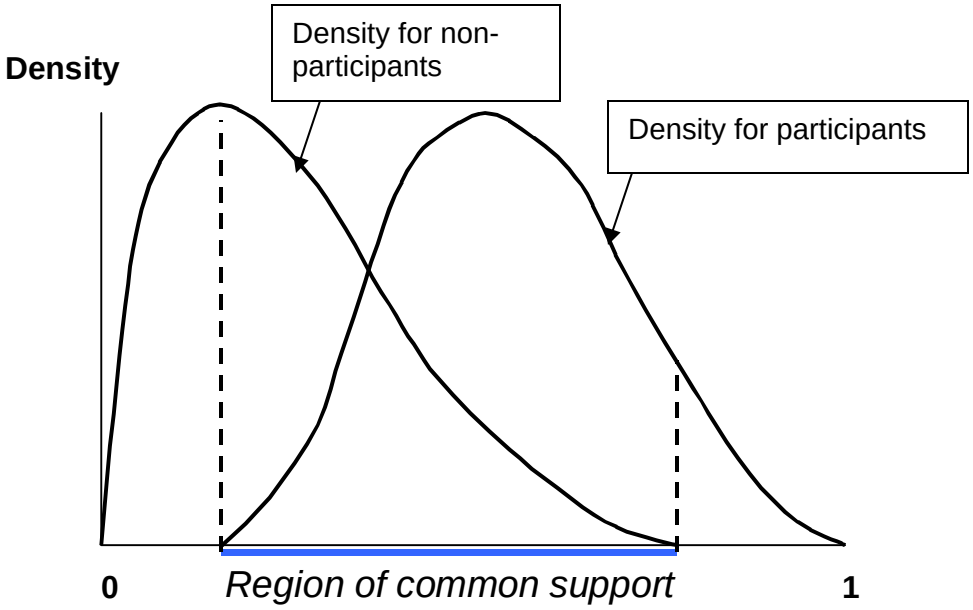
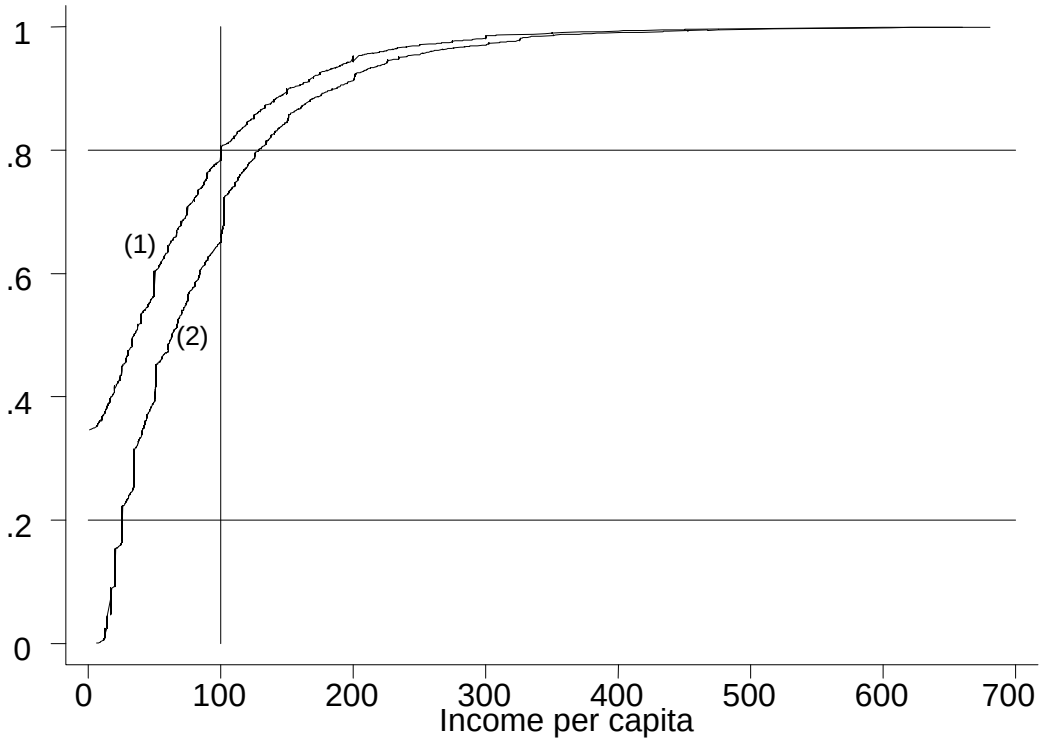


Figure 2: Poverty impacts of disbursements under Argentina's Trabajar program



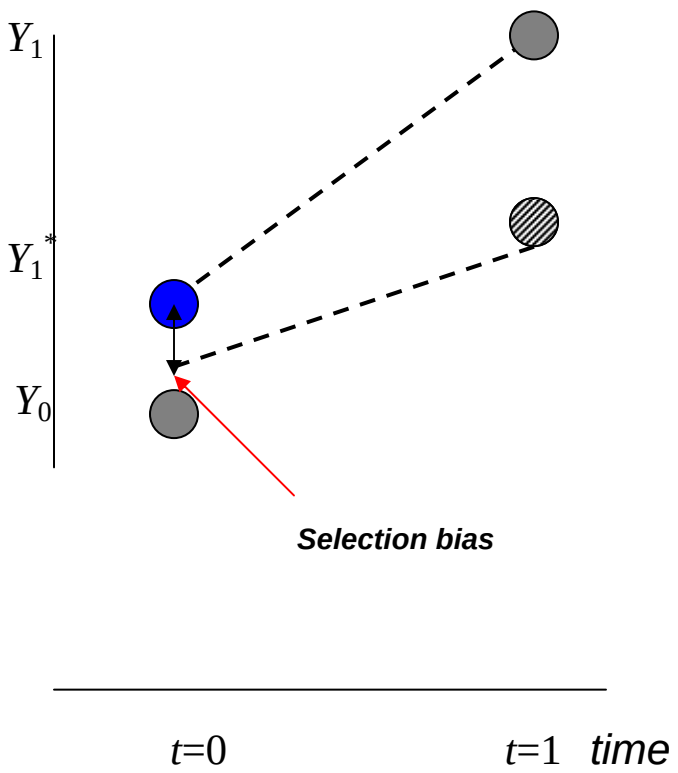
(1) Participant sample pre-intervention (estimated)

(2) Participant sample post-intervention (observed)

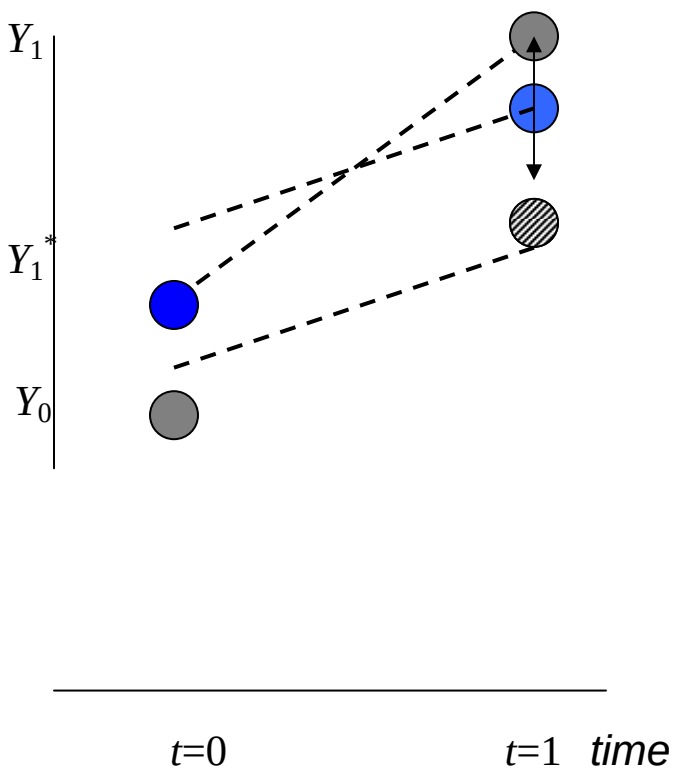
Source: Jalan and Ravallion (2003b).

Figure 3: Bias in double-difference estimates for a targeted anti-poverty program

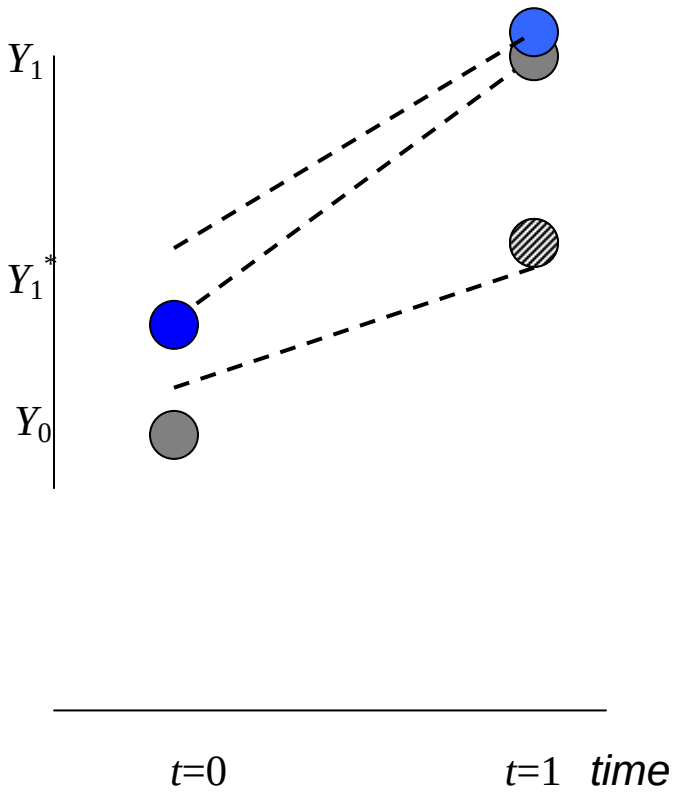
(a)



(b)



(c)



References

Aakvik, Arils, 2001, "Bounding a Matching Estimator: The Case of a Norwegian Training Program," *Oxford Bulletin of Economics and Statistics* 6391): 115-143.

Abadie, Alberto and Guido Imbens, 2006, "Large Sample Properties of Matching Estimators for Average Treatment Effects," *Econometrica* 74(1): 235-267.

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.

Altonji, Joseph, Todd E. Elder and Christopher R. Taber, 2005a, "Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools," *Journal of Political Economy* 113(1): 151-183.

_____, _____ and _____, 2005b, "An Evaluation of Instrumental Variable Strategies for Estimating the Effects of Catholic Schools," *Journal of Human Resources* 40(4): 791-821.

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.

Araujo, Caridad, Francisco Ferreira, Peter Lanjouw and Berk Ozler, 2006, "Local Inequality and Project Choice: Theory and Evidence from Ecuador," Policy Research Working Paper 3997, World Bank, Washington DC. (*Journal of Public Economics*, forthcoming.)

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, 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.

Binswanger, Hans, Shahidur R. Khandker, and Mark Rosenzweig, 1993, "How Infrastructure and Financial Institutions affect Agricultural Output and Investment in India," *Journal of Development Economics* 41: 337-366.

Björklund, Anders and Robert Moffitt, 1987, The Estimation of Wage Gains and Welfare Gains in Self-Selection, *Review of Economics and Statistics* 69(1): 42-49.

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.

Bradlow, Eric, 1998, "Encouragement Designs: An Approach to Self-Selected Samples in an Experimental Design". *Marketing Letters* 9(4): 383-391.

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*, 39:105-115.

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

Caliendo, Marco and Sabine Kopeinig, 2005, "Some Practical Guidance for the Implementation of Propensity Score Matching," Paper 1588, Institute for the Study of Labor (IZA).

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.

Chen, Shaohua, Ren Mu and Martin Ravallion, 2006, "Are there Lasting Impacts of Aid to Poor Areas? Evidence from Rural China," Policy Research Working Paper 4084, World Bank, Washington DC.

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, 1997, *The Analysis of Household Surveys: A Microeconometric Approach to Development Policy*, Baltimore: Johns Hopkins University Press for the World Bank.

_____, 2005, "Measuring Poverty in a Growing World (or Measuring Growth in a Poor World," *Review of Economics and Statistics*, 87(1): 1-19.

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.

De Janvry, Alain and Elisabeth Sadoulet, 2006, "Making Conditional Cash Transfer Programs More Efficient: Designing for Maximum Effect of the Conditonality," *World Bank EconomicReview* 20(1): 1-29.

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

Djebbari, Habiba and Jeffrey Smith, 2005, "Heterogeneous Program Impacts of PROGRESA," mimeo, Laval University and University of Michigan.

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

Duflo, Esther and Rohini Pande, 2005, "Dams," NBER Working Paper 11711. (*Quarterly Journal of Economics*, forthcoming.)

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.

Fraker, Thomas and Rebecca Maynard, 1987, "The Adequacy of Comparison Group Designs for Evaluations of Employment-Related Programs," *Journal of Human Resources* 22(2): 194-227.

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.

Glazerman, Steven, Dan Levy and David Myers, 2003, "Non-Experimental versus Experimental Estimates of Earnings Impacts," *Annals of the American Academy of Political and Social Sciences* 589: 63-93.

Glewwe, Paul, Michael Kremer, Sylvie Moulin and Eric Zitzewitz, 2004, "Retrospective vs. Prospective Analysis of School Inputs: The Case 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, 1998, "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects," *Econometrica* 66: 315-331.

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.

Hausman, Jerry, 1978, "Specification Tests in Econometrics," *Econometrica* 46: 1251-1271.

Heckman, James, 1979, "Sample Selection Bias as a Specification Error," *Econometrica* 47(1): 153-161.

_____, 1997, "Instrumental Variables: A Study of Implicit Behavioral Assumptions Used in Making Program Evaluations," *Journal of Human Resources* 32(3): 441-462.

Heckman, James and Joseph Hotz, 1989, "Choosing Among Alternative NX Methods for Estimating the Impact of Social Programs: The Case of Manpower Training," *Journal of the American Statistical Association* 84: 862-874.

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 Salvador, Navarro-Lozano, 2004, "Using Matching, Instrumental Variables

and Control Functions to Estimate Economic Choice Models," *Review of Economics and Statistics* 86(1): 30-57.

Heckman, James and Richard Robb, 1985, "Alternative Methods of Evaluating the Impact of Interventions", in J. Heckman and B. Singer (eds), *Longitudinal Analysis of Labor Market Data*, Cambridge: Cambridge University Press.

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

Heckman, James and Petra Todd, 1995, "Adapting Propensity Score Matching and Selection Model to Choice-Based Samples," Working Paper, Department of Economics, University of Chicago.

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, Serio Urzua and Edward Vytlacil, 2006, "Understanding Instrumental Variables in Models with Essential Heterogeneity," *Review of Economics and Statistics* 88(3), 389-432.

Heckman, James and Edward Vytlacil, 2005, "Structural Equations, Treatment Effects and Econometric Policy Evaluation," *Econometrica* 73(3), 669-738.

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

Hirano, Keisuke, Guido Imbens and G. Ridder, 2003, "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," *Econometrica* 71: 1161-1189.

Hirano, Keisuke, Guido W. Imbens, Donald B. Ruben and Xiao-Hua Zhou, 2000, "Assessing the Effect of an Influenza Vaccine in an Encouragement Design," *Biostatistics*, 1(1): 69-88.

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

Holland, Paul, 1986, "Statistics and Causal Inference," *Journal of the American Statistical Association* 81: 945-960.

Holtz-Eakin, D., W. Newey and H. Rosen, 1988, "Estimating Vector Autoregressions with Panel Data", *Econometrica*, 56: 1371-1395.

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.

Katz, Lawrence F., Jeffrey R. Kling and Jeffrey B. Liebman, 2001, "Moving to Opportunity in Boston: Early Results of a Randomized Mobility Experiment," *Quarterly Journal of Economics*, 116(2): 607-654.

Keane, Michael, 2006, "Structural vs. Atheoretical Approaches to Econometrics," mimeo, Yale University.

Kish, Leslie, 1965, *Survey Sampling*, New York: John Wiley.

Korinek, Anton, Johan Mistiaen and Martin Ravallion, 2006, "Survey Nonresponse and the Distribution of Income," *Journal of Economic Inequality* 4(2): 33-55.

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

Lanjouw, Peter and Martin Ravallion, 1999, "Benefit Incidence and the Timing of Program Capture," *World Bank Economic Review*, 13(2): 257-274.

Lechner, M., 2001, "Identification and Estimation of Causal Effects of Multiple Treatments Under the Conditional Independence Assumption," in: M. Lechner and F. Pfeiffer (eds) *Econometric Evaluations of Labour Market Policies*, Heidelberg: Physica-Verlag.

Lee, Donghoon, 2005, "An Estimable Dynamic General Equilibrium Model of Work, Schooling, and Occupational Choice," *International Economic Review*, 46(1): 1-34.

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.

Manski, Charles, 1990, "Nonparametric Bounds on Treatment Effects," *American Economic Review Papers and Proceedings* 80: 319-323.

_____, 1993, "Identification of Endogenous Social Effects: The Reflection Problem," *Review of Economic Studies* 60: 531-542.

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 NX Data," *Evaluation Review*, 15(3): 291-314.

_____, 2001, "Policy Interventions, Low-Level Equilibria and Social Interactions," in Steven Durlauf and H. Peyton Young (eds) *Social Dynamics*, Cambridge Mass.: MIT Press.

_____, 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, 2005, "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* 41(5): 788-838.

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.

_____, 2003a, "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.

_____, 2003b, "Measuring Aggregate Economic Welfare in Developing Countries: How Well do National Accounts and Surveys Agree?," *Review of Economics and Statistics*, 85: 645-652.

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

_____, 2005, "Poverty Lines," in *New Palgrave Dictionary of Economics*, 2nd edition, Larry Blume and Steven Durlauf (eds) London: Palgrave Macmillan.

Ravallion, Martin and Shaohua Chen, 2005, "Hidden Impact: Household Saving in Response to a Poor-Area Development Project," *Journal of Public Economics*, 89: 2183-2204.

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.

Ravallion, Martin and Quentin Wodon, 2000, "Does Child Labor Displace Schooling? Evidence on Behavioral Responses to an Enrolment Subsidy," *Economic Journal* 110: C158-C176.

Rosenbaum, Paul, 1995, *Observational Studies*, New York: Springer-Verlag.

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.

_____ and _____, 2000, "Natural "Natural Experiments" in Economics," *Journal of Economic Literature* 38(4): 827-874.

Roy, A., 1951, "Some Thoughts on the Distribution of Earnings," *Oxford Economic Papers* 3: 135-145.

Rubin, Donald B., 1974, "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Education Psychology* 66: 688-701.

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

_____, 1980, "Discussion of the Paper by D. Basu," *Journal of the American Statistical Association* 75: 591-593.

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, Alain 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.

Skoufias, Emmanuel, 2005, *PROGRESA and Its Impact on the Welfare of Rural Households in Mexico*, Research Report 139, International Food Research Institute, Washington DC.

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 NX 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.

Todd, Petra, 2007, "Evaluating Social programs with Endogeneous Program Placement and Selection of the Treated," *Handbook of Development Economics Volume 4*, edited by T. Paul Schultz and John Strauss, Amsterdam, North-Holland.

Todd, Petra and Kenneth Wolpin, 2002, "Using a Social Experiment to Validate a Dynamic Behavioral Model of Child Schooling and fertility: Assessing the Impact of a School Subsidy Program in Mexico," Penn Institute for Economic Research Working Paper 03-022, Department of Economics, University of Pennsylvania.

_____ and _____, 2006, "Ex-Ante Evaluation of Social Programs," mimeo, Department of Economics, University of Pennsylvania.

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 Ren Mu, 2006. "Fungibility and the Flypaper Effect of Project Aid: Micro-Evidence for Vietnam," *Journal of Development Economics*, forthcoming.

Vella, Francis and Marno Verbeek, 1999, "Estimating and Interpreting Models with Endogenous Treatment Effects," *Journal of Business and Economic Statistics* 17(4):

473-478.

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.

Yatchew A., 1998, "Nonparametric Regression Techniques in Economics," *Journal of Economic Literature* 36(June): 669-721.