Evaluating Anti-Poverty Programs

Martin Ravallion

Development Research Group, World Bank

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

Contents

1. Introduction 2
2. The archetypal evaluation problem 3
3. Generic issues in evaluating anti-poverty programs 8
4. Single difference comparisons with randomized assignment 17
5. Single difference matched comparisons 23
6. Exploiting program design in single difference comparisons 31
7. Double difference methods 35
8. Higher-order differencing: following up ex-participants 43
9. Instrumental variables 46
10. Learning more from evaluations 55
11. Conclusions 66
Figures 68
References 70

1 Prepared for the Handbook of Agricultural Economics Volume 4, North-Holland, edited by Robert E. Evenson and T. Paul Schultz. 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, Ajay Chhibber, Aline Coudouel, Jishnu Das, Jed Friedman, Emanuela Galasso, Markus Goldstein, Jose Garcia-Montalvo, David McKenzie, Alice Mesnard, Norbert Schady, Deborah Schlichting, Emmanuel Skoufias, Petra Todd, Dominique van de Walle and participants at a number of presentations at the World Bank and at a workshop held at the Rockefeller Foundation Center at Bellagio, Italy, May 2005.
1. Introduction

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

This paper critically reviews the main methods available for the counterfactual analysis of programs that are assigned exclusively to certain observational units. These may be people, households, villages or larger geographic areas. The key characteristic is that some units get the program and others do not. For example, a social fund 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 eligible households by certain criteria.

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

² 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.
2. The archetypal evaluation problem

Impact evaluation (or “counterfactual analysis”) assesses outcomes for a specific program relative to one or more explicit counterfactuals. It will be assumed that the program is already in place — making the task *ex-post* impact evaluation. That includes the evaluation of a pilot project, as an input to the *ex-ante* assessment of whether the project should be scaled up. However, doing *ex-post* evaluations does not mean that the evaluation should start after the program finishes, or even after it begins. Indeed, the best *ex-post* evaluations are designed *ex-ante* — often side-by-side with the program itself.

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

We will also need some way of inferring the counterfactual. This is inherently unobserved, since it is physically impossible to observe someone in two states of nature at the same time (participating in a program and not participating). Thus evaluation is essentially a problem of missing data. To see the problem more clearly, suppose that data are collected on an outcome indicator \(Y_i\) for unit \(i\) in a sample of size \(n\). For example, \(Y_i\) might be the income of household \(i\) normalized by a household-specific poverty line (reflecting differences in the prices faced in different locations and differences in household size and composition). Some of the sampled units receive the program and some do not; a dummy variable takes the value \(D_i = 1\) for

---

\(^3\) *PROGRESA* stands for *Program for Education, Health and Nutrition*. 
units that receive the program and $D_i = 0$ for those that do not. The value of $Y_i$ if unit $i$ receives the program is $Y_i^T$ (T for “treated”) and it is $Y_i^C$ (C for “counterfactual”) if the program is not received.\footnote{In the literature, $Y_1$ or $Y(1)$ and $Y_0$ or $Y(0)$ are more commonly used for $Y^T$ and $Y^C$ (respectively). My notation will make it easier to follow which group is which.} The individual’s gain from the program is $G_i \equiv Y_i^T - Y_i^C$.

We also collect data on a vector of “control variables”, $X_i$, which includes unity as one element. The most common method of controlling for covariates assumes that outcomes are linear in the control parameters ($\beta^T$ and $\beta^C$) and error terms ($\mu_i^T$ and $\mu_i^C$), giving:

\begin{align*}
Y_i^T &= X_i \beta^T + \mu_i^T \quad (i=1, \ldots, n) \\
Y_i^C &= X_i \beta^C + \mu_i^C \quad (i=1, \ldots, n)
\end{align*}

(1.1) \quad (1.2)

The control variables are taken to be exogenous, i.e., $E(\mu_i^T | X_i) = E(\mu_i^C | X_i) = 0$.

Two widely used impact parameters are the “average treatment effect” ($ATE$) and the “average treatment effect on the treated” ($ATET$), given by:

\begin{align*}
ATE &\equiv E(G_i | X_i) = X_i (\beta^T - \beta^C) \\
ATET &\equiv E(G_i | X_i, D_i = 1) = X_i (\beta^T - \beta^C) + E(\mu_i^T - \mu_i^C | X_i, D_i = 1)
\end{align*}

(2.1) \quad (2.2)

(Sometimes one is interested in the unconditional values, $E(G_i)$ and $E(G_i | D_i = 1)$ respectively.)

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

\[ Y_i^T = X_i \beta^T + \mu_i^T \text{ if } D_i = 1 \]  (3.1)

\[ Y_i^C = X_i \beta^C + \mu_i^C \text{ if } D_i = 0 \]  (3.2)

Equivalently, one can follow the more common practice in applied work of estimating a single
(“switching”) regression for the observed outcome measure on the pooled sample: \(^5\)

\[ Y_i = X_i \beta^C + X_i (\beta^T - \beta^C)D_i + \epsilon_i \text{ (i=1,...,n)} \]  (4)

where the error term has the form:

\[ \epsilon_i = D_i(\mu_i^T - \mu_i^C) + \mu_i^C \]  (5)

The impact parameters are then given by the coefficients on \( D_i \) in (4). A special case that is
popular in practice is the “common effect” specification in which all except the intercepts in the
parameter vectors \( \beta^T \) and \( \beta^C \) are assumed to be invariant to treatment and hence the same in
(1.1) and (1.2). (This assumption is rarely made with any obvious justification beyond the fact
that one can immediately read off the mean impact from the standard regression output.) Then
(4) collapses to a regression of outcomes on participation and the control variables:

\[ Y_i = (\beta_0^T - \beta_0^C)D_i + X_i \beta^C + \epsilon_i \]  (6)

where \( \beta_0^T \) and \( \beta_0^C \) are the intercepts in (1.1) and (1.2).

The simplest (nonparametric) impact estimator is the difference in conditional mean
outcomes between participants and nonparticipants, \( E[Y_i^T | X_i, D_i = 1] - E[Y_i^C | X_i, D_i = 0] \). (For
the parametric model above, this estimator is simply the Ordinary Least Squares (OLS)

\[ Y_i = D_i Y_i^T + (1 - D_i) Y_i^C . \]

\(^5\) The following equation is derived from (3.1) and (3.2) using the identity that:

\[ Y_i = D_i Y_i^T + (1 - D_i) Y_i^C . \]
regression coefficient on the participation dummy variable in (4)). This estimator will not in general give an unbiased estimate of $ATE$ or $ATET$. The bias is obvious if we note the identity:

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

(7)

where $BIAS \equiv E[Y_i^C | X_i, D_i = 1] - E[Y_i^C | X_i, D_i = 0]$. We see that the necessary and sufficient condition for the difference in means between participants and nonparticipants to deliver the impact parameter $ATET$ is that mean counterfactual outcomes do not vary with treatment, i.e.,

$$E[Y_i^C | X_i, D_i = 1] = E[Y_i^C | X_i, D_i = 0].$$

In the context of the parametric regression set-up for the evaluation problem outlined above, the assumption that $BIAS = 0$ is equivalent to requiring that there is no selection bias in placement conditional on $X$, i.e., that the assignment of units between (3.1) and (3.2) is exogenous. Then the error term defined by (5) vanishes in expectation given the regressors — assuring that OLS on (4) gives consistent estimates of the impact parameters under standard conditions. Note also that $ATE$ and $ATET$ become identical under these conditions (since in the absence of selection bias, the error terms vanish in expectation conditional on program participation, so we can set $E(\mu_i^T - \mu_i^C | X_i, D_i = 1) = 0$ in (4)).

However, exogeneity will be a strong assumption in many applications. Concerns will naturally arise when there is purposive placement of the program on the basis of variables that are almost certainly not fully observed to the evaluator. Then the (conditional) counterfactual outcomes depend on treatment status, inducing selection bias.

We will return to this issue when we review the main generic issues confronting all evaluation methods in the next section. The rest of this chapter is then organized around the main methods found in practice for estimating program impacts in the archetypal formulation of
the problem above. One way to assure exogeneity is to randomize placement, in which case we
are dealing with an experimental evaluation, to be considered in detail in section 4. By contrast,
in a non-experimental evaluation (also called an “observational study” or “quasi-experimental
evaluation”) the program is taken to be purposively (non-randomly) placed. (As we will see
later, experimental and non-experimental methods are sometimes combined in practice, but the
distinction is still useful for expository purposes.)

The rest of the chapter then turns to non-experimental methods. These differ in the
assumptions they make in identifying impacts and their related data requirements. The methods
fall into two main groups, depending on which of two (non-nested) conditional independence
assumptions is made. The first group essentially assumes exogeneity of program placement (or
changes in placement) given the observable covariates. Methods that draw on some form of this
exogeneity assumption are discussed in sections 5-8. Sections 5 and 6 look at “single difference”
methods that compare outcome indicators between a sample of participants and one of non-
participants, where those samples are chosen purposively with the aim of reducing selection bias.
Sections 7 and 8 turn to “double or triple difference” methods. These exploit data on outcomes
in the absence of the program, such as using a “baseline survey” done prior to the intervention.

The second, alternative, conditional independence assumption is that there exists an
“instrumental variable” that does not alter outcomes conditional on participation (and other
covariates of outcomes) but nonetheless does influence participation. The instrumental variable
thus isolates a part of the variation in program placement that can effectively be treated as
exogeneous. This is the method is discussed in section 9.

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

3. **Generic issues in evaluating anti-poverty programs**

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

   That objection might be persuasive if eligible people have been denied the program for the purpose of the evaluation and the knowledge from that evaluation does not benefit them. However, in practice, the main reason 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 what resources are made available. Poor people benefit from good evaluations, which weed out defective anti-poverty programs and identify good programs.

   Having secured an agreement to doing the evaluation, one must confront essentially three classes of problems that confound all efforts to identify impact. The first has already been mentioned, namely selection bias, which can arise from either observables or un-observables. The second is the existence of spillover effects, confounding efforts to locate a program’s impacts amongst only its direct participants (when compared to non-participants). After examining these issues in more detail, the section reviews a third set of problems related to data and measurement.
**Have we dealt adequately with selection on observables?** A common concern in non-experimental evaluations is whether the selection process for the program being evaluated is captured adequately by the control variables \( X \). This concern cannot be separated from the problem of non-random placement conditional on observables. One cannot (of course) judge whether exogeneity of placement is a plausible assumption without first establishing whether one has dealt adequately with the observable heterogeneity.

Equations (3) and (4) deal with selection on observables in a rather special way, in that the controls enter in a linear-in-parameters form. This *ad hoc* assumption is rarely justified by anything more than computational convenience (which is rather lame these days). Section 5 will consider formulations of the impact estimation problem under exogeneity that attempt to deal with this source of bias in a more general way.

In non-experimental evaluations it can sometimes be difficult to assure that observables are balanced between treatment and comparison observations. To see why, suppose that placement is determined by a “proxy-means test,” as often used for targeting anti-poverty programs in developing countries. This assigns a score to all potential participants as a function of observed characteristics. When strictly applied, the program is assigned if and only if a unit’s score is below some critical level, as determined by the budget allocation to the scheme. (The pass-score is non-decreasing in the budget under plausible conditions.) 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 called “failure of common support” in the evaluation literature. The problem is plain enough: how can we infer the counterfactual for participants on the basis of non-participants who do not share the same characteristics, as summarized by their score on the proxy means test? If we want to infer mean impact for those
receiving the program then we must have a serious concern about the validity of any comparison group design in this case.\textsuperscript{6} While this example has pedagogic value, it is an extreme case. Thankfully, in practice, there is typically some degree of fuzziness in the application of the proxy-means test and there is typically incomplete coverage of those who pass the test.

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

\textit{Is there a latent selection process?} Like most public programs, participation in direct interventions against poverty is almost never random. This is a problem if some of the variables that jointly influence outcomes and program placement are unobserved to the evaluator. If this is the case then we cannot attribute to the program the observed differences in measured outcomes between units that receive the program and those who do not (conditional on the control variables). The differences in conditional means that we see in the data could just be due to the fact that the program participants were purposely selected by a process that we do not fully observe. When program take-up is a matter of individual choice, there must be a reasonable presumption that selection into the program depends on the gains from participation, which are not fully observed by the evaluator.

\textsuperscript{6} If we don’t need to know impact for the treatment group as whole then the concern is diminished. For example, consider the policy choice of whether to increase the program’s budget allocation by raising the pass mark in the proxy-means test. In this case, we only need focus on impacts in a neighborhood of the pass-mark. Section 6 further discusses “discontinuity designs” for such cases.
In terms of the classic formulation of the evaluation problem above, suppose that participants have latent attributes that yield higher outcomes than non-participants (at given $X$). Then the error terms in the equation for participants (3.1) will be centered to the right relative to those for non-participants (3.2). The error term in (4) cannot vanish in expectation and OLS will give biased and inconsistent estimates. The corresponding nonparametric impact estimate using the conditional means is then biased in the amount: $BIAS = E[Y^*_i | X_i, D_i = 1] - E[Y^*_i | X_i, D_i = 0]$ (section 2). Again it should be emphasized that the extent of concern about this form of selection bias in practice cannot be separated from the prior question above as to how well we have controlled for observable heterogeneity.

There are examples suggesting that $BIAS$ can be large in non-experimental impact estimates in specific cases. A widely-cited study by Lalonde (1986) found large biases when he compared the results of various non-experimental methods with a randomized evaluation of a U.S. training program. Different non-experimental methods also gave quite different results. Similarly, Glewwe et al. (2004) find that non-experimental methods give a larger estimated impact of “flip charts” on the test scores of Kenyan school children than implied by an experiment; they argue that biases in their non-experimental methods account for the difference. Using a different approach to testing non-experimental methods, van de Walle (2002) gives an example for rural road evaluation in which a naïve comparison of the incomes of villages that have a rural road with those that do not indicates large income gains when in fact there are none. Van de Walle used simulation methods in which the data were constructed from a model in which the true benefits were known with certainty and the roads were placed in part as a function of the average incomes of different villages. Only a seemingly small weight on village income in determining road placement was enough to severely bias the mean impact estimate.
Of course, one cannot reject non-experimental methods in other applications on the basis of such studies; arguably the lesson is that better data and methods are needed, informed by past knowledge of how such programs work. In a critique of the Lalonde study, Heckman and Smith (1995) point out that (amongst other things) the data used contained too little information relevant to eligibility for the program studied, that the methods used had limited power for addressing selection bias and did not include adequate specification tests. Similarly, controls for observable correlates of income would almost certainly reduce the bias suggested by van de Walle’s simulations. In the presence of severe data problems it cannot be surprising that observational studies perform poorly in correcting for bias.

**Are there hidden impacts for “non-participants”?** Eliminating selection bias does not assure that impacts can be identified. The classic formulation of the evaluation problem outlined in section 2 assumes that the treatment of unit $i$ can only affect outcomes for that unit. Then we can observe a comparison group that is in no way affected by the program. We may do this at the time the program is in place (giving a “single-difference” design, as discussed further in sections 4-6) or we may do it for participants before the program is in place, giving a “reflexive comparison” (discussed in section 7); when we do both we have a “double difference” (also discussed in section 7). Under certain conditions, we can also infer impacts by comparing those who leave a program with those who stay (section 8). However, all these cases assume that one can observe the non-participation state in a way that is uncontaminated by the program.

That can be a problematic assumption for anti-poverty programs. For example, suppose that we are evaluating a workfare program whereby the government commits to give work to anyone who wants it at a stipulated wage rate; this was the aim of the famous Employment

---

7 Also see the discussion in Heckman et al., (1999).
8 This is sometimes called the “stable unit treatment assumption” in the evaluation literature; see for example, Angrist et al. (1996).
Guarantee Scheme (EGS) in the state of Maharashtra in India and at the time of writing the Government of India is planning to expand the idea to the country as a whole. The attractions of an EGS as a safety net stem from the fact that access to the 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 survey data on both groups and compare outcome indicators between the two, as a means of identifying impact (possibly after cleaning out any observable heterogeneity).

However, this classic evaluation design could give a severely biased result. The gains from such a program must spill over 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 — for no able-bodied worker would accept non-EGS work at any wage rate below the EGS wage. So even if one picks the observationally perfect comparison group, one will conclude that the scheme has no impact, since wages will be the same for participants and non-participants. But that would entirely miss the impact, which could be large for both groups.

Spillover effects can also arise from the behavior of governments. Whether the resources transferred to participants actually financed the identified project is often unclear. To some degree, all external aid is fungible. Yes, it could be verified in supervision that the proposed sub-project was actually completed. But one cannot rule out the possibility that it would have been done anyhow. Participants and local leaders would naturally have put forward the best
development option they saw, even if it was something they planned to do anyway with the resources already available. Then there is some other (infra-marginal) expenditure that was really being financed by the aid. Similarly, there is no way of ruling out the possibility that non-project villages benefited 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 Cratty (2005) in the context of a rural-roads project in Vietnam. The authors find no impact on comparing kilometers of roads rehabilitated by the (aid-financed) project with a comparison group of non-participating communes. This is interpreted as reflecting in part the fungibility of aid, though it turns out that selection bias is also at work, such that the degree of fungibility is overstated unless one controls adequately for the purposive geographic targeting of the development project.

**How are outcomes for the poor to be measured?** The archetypal formulation of the evaluation problem focuses on mean impacts. This can be readily adapted to allow for a measure of poverty by re-interpreting the outcome measure as a variable taking the value $Y_i = 1$ if unit $i$ is poor and $Y_i = 0$ otherwise. That assessment will typically be based on a set of poverty lines, which ideally give the minimum income that is deemed to be necessary for unit $i$ to achieve a given reference utility, interpretable as the minimum “standard of living” needed to be judged non-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.$^9$

---

$^9$ 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 (2005).
With this re-interpretation of the outcome variable, equation (2) now gives the program’s impact on the headcount index of poverty (% below the poverty line).\(^{10}\) By repeating the impact calculation for multiple “poverty lines” one can then trace out the impact on the cumulative distribution of income. Higher order poverty measures (that penalize inequality amongst the poor) can also be accommodated as long as they are members of the (broad) class of additive measures, by which the aggregate poverty measure can be written as the population-weighed mean of all individual poverty measures in that population.\(^{11}\) We may have multiple indicators, such as for different poverty lines or other non-income dimensions of poverty, such as non-income indicators of future poverty (such as schooling).

**What data are required?** As is clear from the above discussion, concerns about inadequate or imperfect data lie at the heart of the evaluation problem. When embarking on any impact evaluation, it is important to first know a lot about the administrative/institutional details of the program; that information typically comes from the program administration. For non-experimental evaluations, such information is key to designing a survey that collects the right data to control for the selection process. Knowledge of the program’s context and design features can also help in dealing with selection on unobservables, since it can sometimes generate plausible identifying restrictions, as discussed further in sections 6 and 9.

Non-experimental evaluations of anti-poverty programs can be very demanding in their data requirements. The precise sources of data used in an evaluation can embrace both informal, unstructured, interviews with participants in the program as well as quantitative data from representative samples. However, it is extremely difficult to ask counter-factual questions in interviews or focus groups; try asking someone who is currently participating in a public

---

\(^{10}\) I leave aside the econometric issues that arise with linear binary response models (see, for example, Wooldridge, 2002).

\(^{11}\) See Atkinson (1987) on the general form of these measures and examples in the literature.
program: “what would you be doing now if this program did not exist?” Talking to program participants can be valuable, but it is unlikely to provide a credible impact evaluation on its own. One also needs data on the outcome indicators and relevant explanatory variables.

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

A potentially serious concern is the comparability of different data sources, particularly those used for the observations on participants and non-participants. Differences in the design of the survey instruments can entail non-negligible differences in the outcome measures. 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.\textsuperscript{12}

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).\textsuperscript{13} Yet here 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 do encounter

\textsuperscript{12} Also see the example in Diaz and Handa (2004).

\textsuperscript{13} 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).
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).\textsuperscript{14}

Problems of measurement errors in surveys can to some extent be dealt with by the same methods used for addressing selection bias. For example, if the measurement problem affects the measured outcomes for treatment and comparison units identically (and additively) and is uncorrelated with the control variables ($X$ in 1.1 and 1.2) then it will not be a problem when estimate the average treatment effect. This again points to the importance of the controls. But even if there are obvious omitted variables correlated with the measurement error, there is still hope for obtaining reliable estimates using the class of 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.

4. Single difference comparisons with randomized assignment

A social experiment aims to randomize placement, such that all units (within some well-defined set) have the same chance \textit{ex-ante} of receiving the program. Unconditional randomization is virtually inconceivable for anti-poverty programs, which policy makers are often keen to target on the basis of observed characteristics, such as households with many dependents living in poor areas. More commonly, program assignment is randomized conditional on certain 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

\textsuperscript{14} In measuring poverty some researchers have replaced the survey mean by the mean from the national accounts (GDP or consumption per capita); see, for example, Bhalla (2002) and Sala-i-Martin (2002). This assumes that the discrepancy is distribution neutral, which is unlikely to be the case; for example, selective non-response to surveys can generate highly non-neutral errors (Korinek et al., 2005).
whether or not a unit actually receives the program. By implication, $BIAS=0$ and so the observed 
ex-post difference in mean outcomes between the two groups is attributable to the program. In 
terms of the formulation of the evaluation problem in section 2, randomization guarantees that 
there is no sample selection bias in estimating (4.1) and (4.2) or (equivalently) that the error term 
in equation (5) is orthogonal to the regressors. The non-participants are then a valid control 
group for identifying the counterfactual,\textsuperscript{15} and ATE is consistently estimated (nonparametrically) 
by the difference between the sample means of $Y_i^T$ and $Y_i^C$ (including at given values of $X_i$).

**Examples:** Social experiments have been done on a number of evaluations in the US, 
often applied to a pilot scheme; much has been learnt about welfare policy reform 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 done by Burtless 

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 non-experimental tests for neighborhood effects (Manski, 1993).\textsuperscript{16}

\textsuperscript{15} The term “control group” is often confined to social experiments, with the term “comparison 
group” used in non-experimental evaluations.

\textsuperscript{16} Note that the design of the MO experiment does not identify neighborhood effects at the origin, 
given that attributes of the destination also matter to outcomes (Moffitt, 2001).
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 (considerable) influence that this program has had in the development community clearly stems in no small measure from the substantial, and public, effort that went into its evaluation. One third of the sampled communities deemed eligible for the program were chose 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 in a developing country, Newman et al. (2002) were able to randomize eligibility to a World Bank supported social fund within one region of Bolivia. The fund-supported investments in education were found to have had significant impacts on school infrastructure but not education outcomes within the evaluation period. Randomization was also used by Angrist et al. (2002) to evaluate a program in Colombia that allocated vouchers for schooling by a lottery. Three years later, the lottery winners had significantly better school attainments, with lower 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 paper to examine some lessons from this evaluation more closely.)

It has been argued that development agencies such as the World Bank should make much greater use of social experiments. While the World Bank has supported a number of social 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;\textsuperscript{17} indeed, only 21 of the evaluations used any form of counterfactual analysis. Cook (2001) and Duflo and Kremer (2005) have advocated that OED should do many more social experiments.\textsuperscript{18} However, before accepting that advice one should be aware of some of the concerns raised by social experiments, to which we now turn.

\textbf{Issues with social experiments:} There has been much debate about whether randomized designs are in fact the ideal for evaluating anti-poverty programs.\textsuperscript{19} Social experiments have often raised ethical objections and generated political sensitivities, which have sometimes stalled attempts to implement them. There is a perception that social experiments treat people like “guinea pigs,” deliberately denying access to the program for some of those who need it (to form the control group) in favor of some who don’t (in that 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

\textsuperscript{17} From Kapoor’s description it is not clear that even this one evaluation was a genuine social experiment.

\textsuperscript{18} OED only assesses Bank projects (including the evaluations done by the Bank’s project staff) 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, which invariably uses counterfactual analysis and sometimes randomization.

\textsuperscript{19} On the arguments for and against social experiments see (\textit{inter alia}) Heckman and Smith (1995), Burtless (1995) and Moffitt (2003).
denying the program to poor people who need it — in both cases running counter to the aim of fighting poverty.

As noted in the last section, incomplete coverage for anti-poverty programs is rarely because of the evaluation, but rather it is because there are too few resources available. In fact, when there are poor people who can’t get the program with the currently available resources, it can be argued that the ethical concerns are less persuasive for social experiments. The fairest solution in such a situation is surely to assign the program randomly, so that everyone has an equal opportunity of getting the limited resources available. Against that argument, it is hard to appreciate the “fairness” of an anti-poverty program that deliberately ignores available information on differences in the extent of deprivation. The “equal opportunity” argument is more persuasive when the randomization is conditional on observables, which is the case in most applications.

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

\[\text{From the description of the Newman et al. (2003) study it appears that this is how randomization was defended to the relevant authorities in their case.}\]
methods can help address the problem, and how partially randomized designs can help identify impacts using non-experimental methods.

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

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

This is an instance of a more general and fundamental problem with randomized designs for anti-poverty programs, namely that the very process of randomization can alter the way a program works in practice. There may well be systematic differences between the characteristics of people normally attracted to a program and those randomly assigned the program from the
same population. (This is sometimes called “randomization bias.”) Heckman and Smith (1995) discuss an example from the evaluation of the JTPA, whereby substantial changes in the program’s recruiting procedures were required to form the control group. The evaluated pilot program is not then the same as the program that gets implemented — casting doubt on the validity of the inferences drawn from the evaluation.

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

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

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

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

5. Single difference matched comparisons

As section 3 emphasized, selection bias is to be expected in comparing a random sample from the population of participants with a random sample of non-participants. There must be a general presumption that such comparisons misinform policy. How much so is an empirical
question. On *a priori* grounds it is worrying that many non-experimental evaluations in practice provide too little information to assess properly whether the “comparison group” of non-participants is similar to the participants in the absence of the intervention.\textsuperscript{21}

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

**Propensity-Score Matching:** Rosenbaum and Rubin (1983) offer a solution to this problem.\textsuperscript{22} The method selects comparators according to their predicted probabilities of participation (called their “propensity scores”). The key to PSM is understanding and modeling how the program is assigned. Participants are matched to non-participants on the basis of the propensity score, $P(Z_i) = E(D_i|Z_i) \ (0 < P(Z_i) < 1)$, where $Z_i$ is a vector of pre-exposure control variables (which can include pre-treatment values of the outcome indicator).\textsuperscript{23} PSM uses $P(Z_i)$ (or a monotone function of $P(Z_i)$) to select comparison units. It is known from Rosenbaum and Rubin (1983) that if (i) the $D_i$’s are independent over all $i$, and (ii) outcomes are independent of participation given $Z_i$, then outcomes are also independent of participation given $P(Z_i)$.\textsuperscript{24} Assumption (ii) is essentially a more general version of the exogeneity-of-placement assumption discussed in sections 2 and 3. Under these conditions, exact matching on $P(Z_i)$ eliminates

\begin{footnotesize}
\begin{enumerate}
\item See, for example, Kapoor’s (2002) comments on OED’s evaluations.
\item The Rosenbaum-Rubin paper built on a series of papers in the statistical literature by Rubin and others. For a thorough recent review of the theory of propensity score matching see Imbens (2004).
\item The present discussion is confined to the standard case of binary treatment. In generalizing to the case of multi-valued or continuous treatments one defines the generalized propensity score given by the conditional probability of a specific level of treatment (Imbens, 2000; also see Hirano and Imbens, 2004).
\item For a clear recent statement and proof of the Rosenbaum-Rubin theorem see Imbens (2004).
\end{enumerate}
\end{footnotesize}
selection bias.\textsuperscript{25} As in a social experiment, ATE is non-parametrically identified by the difference between the sample means of $Y_i^T$ and $Y_i^C$ for the matched comparison group.

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

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

\textsuperscript{25} On the efficiency of PSM relative to covariate matching see Angist and Hahn (2004).
the available data do not include important determinants of participation then the presence of these unobserved characteristics will mean that PSM will not be able to reproduce (to a reasonable approximation) the results of a social experiment.

Common practice is to use the predicted values from a standard logit or probit regression to estimate the propensity score for each observation in the participant and the non-participant samples (though non-parametric binary response models can also be used; see Heckman et al., 1997). The participation regression is of interest in its own right as it can provide useful insights into the targeting performance of an anti-poverty program (see, for example, the discussion in Jalan and Ravallion, 2003b). The comparison group is then formed by picking the “nearest neighbor” for each participant, defined as the non-participant that minimizes $|\hat{P}(Z_i) - \hat{P}(Z_j)|$ as long as this does not exceed some caliper bound. Given measurement errors, more robust estimates are likely by taking the mean of the nearest (say) five neighbors, though this does not necessarily reduce bias.\textsuperscript{26} 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.

More generally, the estimator for mean impact is $\sum_{j=1}^{NT}(Y_j - \sum_{i=1}^{NC} W_{ij} Y_{ij}^c)/NT$ where $NT$ is the number receiving the program, $NC$ is the number of non-participant households and the $W_{ij}$’s are the weights applied in calculating the average outcome of the matched non-participants. (Sampling weights may also be needed, depending on the survey design.) There are several weighting schemes that have been used, ranging from nearest-neighbor weights to non-parametric weights based on kernel functions of the differences in scores whereby all the

\textsuperscript{26} Rubin and Thomas (2000) use simulations to compare the bias in using the nearest five neighbors to just the nearest neighbor; no clear pattern emerges.
comparison units are used in forming the counterfactual for each participating unit, but with a
weight that reaches its maximum for the nearest neighbor but declines as the absolute difference
in propensity scores increases; Heckman et al. (1997b) discuss this weighting scheme.\footnote{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.}

Mean impacts can be calculated conditional on observed characteristics. For anti-poverty programs one is interested in comparing the conditional mean impact across different pre-intervention incomes. For each sampled participant, one estimates the income gain from the program by comparing that participant’s income with the income for matched non-participants. Subtracting the estimated gain from observed post-intervention income, it is then possible to estimate where each participant would have been in the distribution of income without the program. On averaging this across different strata defined by pre-intervention incomes one can 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. Further discussion of how the results of an impact assessment by PSM can be used to assess impacts on poverty measures robustly to the choice of those measures and the poverty line can be found in Ravallion (2003b), with a detailed worked example for an actual anti-poverty program (Argentina’s Trabajar program).

\textbf{How does PSM differ from other methods?} In a social experiment (at least in its pure form), randomized assignment assures that the distributions of both observables and unobservables are balanced between treatment and comparison units. PSM only attempts to
balance the distributions of observables. Thus PSM will face concerns about selection bias whenever one can postulate the existence of a latent variable that jointly influences placement and outcomes (thus invalidating the key conditional independence assumption made by PSM). This must be judged for the application in hand.

Nor can it be assumed that eliminating selection bias based on observables will reduce the aggregate bias; that will only be the case if the two sources of bias — that associated with observables and that due to unobserved factors — go in the same direction, which cannot be assured on a priori grounds. If the selection bias based on unobservables counteracts that based on observables then eliminating only the latter bias will increase aggregate bias. While this is possible in theory, I do not know of any plausible example from practice.

A natural comparison is between PSM and an OLS regression of the outcome indicators on dummy variables for program placement, allowing for the observable covariates entering as linear controls (as in equations 4 and 6). OLS requires essentially the same conditional independence (exogeneity) assumption as PSM, but also imposes arbitrary functional form assumptions concerning the treatment effects and the control variables. By contrast, PSM (in common with experimental methods) does not require a parametric model linking outcomes to program participation. Thus PSM allows estimation of mean impacts without arbitrary assumptions about functional forms and error distributions. This can also facilitate testing for the presence of potentially complex interaction effects. For example, Jalan and Ravallion (2002a) use PSM to study how the interaction effects between income and education influence the child-health gains from access to piped water in rural India. The authors find a complex pattern of interaction effects; for example, poverty attenuates the child-health gains from piped water, but less so the higher the level of maternal education.
A variation on the standard OLS estimator is to add the estimated propensity score $\hat{P}(Z)$ to a regression of the outcome variable on the treatment dummy variable, $D$. Under assumptions of PSM this will clearly eliminate any omitted variable bias in having excluded $Z$ from that regression, given that $Z$ is independent of treatment given $P(Z)$. Yet another variation is to include an interaction effect between $\hat{P}(Z_i)$ and $D_i$. These variations on the standard OLS estimate of the common effects model (7) may or may not be an improvement or more convenient (certainly the OLS standard errors will need correction given that $\hat{P}(Z_i)$ requires previously estimated parameters), and they do not have the non-parametric flexibility of PSM.

PSM also differs from standard regression methods with respect to the sample. In PSM one confines attention to matched sub-samples; unmatched comparison units are dropped. Some of the non-participants may be excluded because they have a score that is outside the range found for the participant sample. The range of scores estimated for the participants should correspond closely to that for the retained sub-sample of non-participants. In other words, matching is confined to the region of common support, as illustrated in Figure 1. One may also want to restrict potential matches in other ways, depending on the setting. For example, one may want to 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.

---

28 This provides a further intuition as to how PSM works; see the discussion in Imbens (2004).
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, possibly including variables that are poor predictors of outcomes. Indeed, analytic results and simulations indicate that variables with weak predictive ability for outcomes can still help reduce bias in estimating causal effects using PSM (Rubin and Thomas, 2000).

It is an empirical question as to how much difference it would make to mean-impact estimates by using PSM rather than OLS. Comparative methodological studies have been rare. In one exception, Godtland et al. (2004) use both an outcome regression and PSM for assessing the impacts of field schools on farmers’ knowledge of good practices for pest management in potato cultivation. They report that their results were robust to changing the method used.

**How well does PSM perform?** Returning to the same data set used by the Lalonde (1986) study (described in section 3), an influential paper by Dehejia and Wahba (1999) found that propensity-score matching achieved a good approximation — much better than the non-experimental methods studied by Lalonde. However, the robustness of the Dehejia-Wahba findings to sample selection and the specification chosen for calculating the propensity scores has been questioned by Smith and Todd (2005a), who argue that PSM does not solve the selection problem in the program studied by Lalonde.29

Similar attempts to test PSM against randomized evaluations have shown mixed results. Agodini and Dynarski (2004) find no consistent evidence that PSM can replicate experimental results from evaluations of school dropout programs in the US. Using the PROGRESA data base, Diaz and Handa (2004) find that PSM performs well when the same survey instrument is

used for measuring outcomes for the treatment and comparison groups, but not when the survey instruments differ. The importance of using the same survey instrument in PSM is also emphasized by Heckman et al. (1997a, 1998) in the context of their evaluation of a US training program. The latter study also points to the importance of both participants and non-participants coming from the same local labor markets, and of being able to control for employment history.

**In summary:** PSM is an important addition to the menu of tools available for counterfactual analysis. The emphasis on understanding the assignment mechanism — the observable determinants of program placement — is welcome, and is certainly a more convincing starting point than the presumption that one has found a “natural experiment,” which seems unlikely in general. In evaluating anti-poverty programs in developing countries, single-difference comparisons using PSM also have the advantage that they do not require either randomization or baseline (pre-intervention) data. While this is can be a huge advantage, it comes at a cost. To accept the exogeneity assumption one must be confident that one has controlled for the factors that jointly influence program placement and outcomes. So PSM requires good data; Section 8 will give an example of how far wrong the method can go if data are inadequate.

6. **Exploiting program design in single difference comparisons**

Single difference non-experimental estimators can sometimes usefully exploit features of program design for identification. Discontinuities generated by program eligibility criteria can help identify impacts in a neighborhood of the cut-off points for eligibility. Delays in the implementation of a program can also facilitate forming comparison groups, which can also help pick up some sources of latent heterogeneity.
**Discontinuity designs:** 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. Examples include a proxy-means test that sets a maximum score for eligibility (as discussed in section 3) and programs that confine eligibility within geographic boundaries. In a test of how well discontinuity designs perform, Buddelmeyer and Skoufias (2004) use the cut-offs in PROGRESA’s eligibility rules to measure impacts and compare the results to those obtained by exploiting the program’s randomized design. The authors find that the discontinuity design gives good approximations for almost all outcome indicators.

To see more clearly what this method involves, let $M_i$ denote the score received by unit $i$ in a proxy-means test (say) and let $m$ denote the cut-off point for eligibility, such that $D_i = 1$ for $M_i \leq m$ and $D_i = 0$ otherwise. Impact is then $E(Y_i^T | M_i = m - \varepsilon) - E(Y_i^C | M_i = m + \varepsilon)$ 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, $P(M_i) = E(D_i | M_i)$, which is an increasing function of $M_i$ with a discontinuity at $m$. The essential idea remains the same, in that impacts are measured by the difference in mean outcomes in a neighborhood of $m$.

The key identifying assumption is that there is no discontinuity in counterfactual outcomes at $m$. However, the fact that a program has more-or-less strict eligibility rules does not (of itself) mean that continuity is a plausible assumption. For example, the geographic boundaries for program eligibility will often coincide with local political jurisdictions, entailing current or past geographic differences in (say) local fiscal policies and institutions that cloud

---

30 Hahn et al. (2001) provide a formal analysis of identification and estimation of impacts for discontinuity designs under this assumption.
identification. The plausibility of the continuity assumption for counterfactual outcomes must be judged in each application.

This method assumes that the evaluator knows $M_i$ and (hence) eligibility for the program. That will not always be the case. Consider again a means-tested income 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 implicitly assuming 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 below.

How does this method compare to PSM? A discontinuity design gives mean impact for a selected sample of the participants, while PSM aims to give mean impact for the treatment group as a whole. However, the aforementioned common support problem that is sometimes generated by eligibility criteria can mean that PSM is also confined to a selected sub-sample; the question is then whether that is an interesting sub-sample. The truncation of treatment group samples under PSM will most likely tend to exclude those with the highest probability of participating (for which non-participating comparators are hardest to find), while discontinuity designs will tend to include only those with the lowest probability. The latter sub-sample can, nonetheless, be relevant for deciding about program expansion; section 10 returns to this point.
Although mean impacts are non-parametrically identified for discontinuity designs, the literature in economics has more often used an alternative parametric method in which the discontinuity in the eligibility criterion is used as an instrumental variable for program placement; we will return to give examples in section 9.

**Pipeline comparisons:** The idea here is to use people who have applied for a program but have not yet received it as the comparison group.\(^{31}\) 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. Non-experimental pipeline comparisons have also been used in developing countries. An example can be found in Chase (2002) who used communities who had applied for a social fund (in Armenia) as the source of the comparison group in estimating the fund’s impacts on communities that received its support. In another example, Galasso and Ravallion (2004) evaluated a large social protection program in Argentina, namely the Government’s *Plan Jefes y Jefas*, which was the main social policy response to the severe economic crisis of 2002. To form a comparison group for participants they used those individuals who had successfully applied for the program, but had not yet received it, as the comparison group. Notice that this method does to some extent address the problem of latent heterogeneity in other single-difference estimators, such as PSM; the successful applicants will tend to have similar unobserved characteristics, whether they have yet received the program or not.

The key assumption here is that the timing of treatment is random given application. In practice, one must anticipate a potential bias arising from selective treatment amongst the applicants or behavioral responses by applicants awaiting treatment. This is a greater concern in

\[^{31}\text{This is sometimes called “pipeline matching” though this term is less than ideal given that no matching is actually done.}\]
some settings than others. For example, Galasso and Ravallion argued that it was not a serious concern in their case given that they assessed the program during a period of rapid scaling up, during the 2002 crisis in Argentina. The authors also tested for observable differences between the two sub-sets of applicants, and found that observables (including idiosyncratic income shocks during the crisis) were well balanced between the two groups, alleviating concerns about bias. Using longitudinal observations also helped; we return to this 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 results should, however, be tested for selection bias based on observables and (if need be) PSM can be used to clean out the observable heterogeneity prior to making the pipeline comparison (Galasso and Ravallion, 2004).

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 (such as defined by pre-existing infrastructure) 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. One would probably also want to allow for observable heterogeneity and time effects.

7. Double difference methods

A popular approach for addressing concerns about the exogeneity assumption in single-difference cross-sectional comparisons is the double difference (or “difference-in-difference”) (DD) method. This method compares samples of participants and non-participants before and after the intervention. Typically one has a baseline survey before the intervention, covering both non-participants and participants. Often one does not know who will participate, and must make
an informed guess in designing the sampling for the baseline survey; knowledge of the program
design and setting can help. One then does one or more follow-up surveys. Finally one
calculates the mean difference between the “after” and “before” values of the outcome indicator
for each of the treatment and comparison groups. The difference between these two mean
differences (hence “double difference”) is the estimate of the impact of the program.

To see what is involved in more formal terms, suppose that we have collected data on an
outcome measure $Y$ for a set $\Psi$ of participants and a comparison group. We can write the
outcome measure, $Y_{it}$, for the $i$’th treatment unit ($D_i = 1$) observed at two dates $t=1,2$ as

$$(Y_{it} \mid D_i = 1) = Y_{it}^C + G_{it} + \varepsilon_{it}$$

where $Y_{it}^C$ is the counter-factual outcome measure for treatment unit $i$ if the program had not existed, $G_{it}$ is the gain attributable to the program and $\varepsilon_{it}$ is a zero-mean innovation error term uncorrelated with program participation, to allow for measurement error in $Y_{it}$. An indicator of the counter-factual is available from a comparison group and is given by $\hat{Y}_{it}^C$. This may be a noisy indicator due to selection bias or measurement errors (such as arising from a change in survey design between the baseline and follow-up surveys).

The two key assumptions of a DD estimator are: (i) on average, the changes over time in $\hat{Y}_{it}^C$ reveal the changes in $Y_{it}^C$, i.e., $E(\Delta \hat{Y}_{it}^C) = E(\Delta Y_{it}^C)$; this will hold if the selection bias or other sources of error are separable and time invariant, and so they are swept away by taking differences over time, and (ii) period 1 outcomes are not contaminated by the expectation of the program’s future placement, i.e., $G_{i1} = 0$. Under these assumptions, on taking the expectation over all participants, the DD estimator of mean impact is the single-difference impact estimate for the second period less the single difference in the baseline:

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

(8)
Equivalently, DD is the outcome gain observed over time for the treatment group less that for the comparison group. This can be readily generalized to multiple time periods and DD can then be estimated by the common effect regression of \( Y_{it} \) on the (individual and date-specific) participation dummy variable \( D_{it} \), with individual and time fixed effects.\(^{32}\)

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

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

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

---

\(^{32}\) As is well-known in econometrics, when the error term is serially correlated one must take account of this in calculating the standard errors of the DD estimator; Bertrand et al. (2004) demonstrate the possibility for large biases in the uncorrected (OLS) standard errors for DD estimators.
ownership of water services to identify impacts. Their results suggest that privatization of water services reduced child mortality.

A DD design can also be used to address possible biases in a social experiment, whereby there is some form of selective compliance or other distortion to the randomized assignment (as discussed in section 3). An example can be found in Thomas et al. (2003) who randomized assignment of iron-supplementation pills in Indonesia, with a 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 DD design tracks the differences over time between participants and non-participants, that is not the only possibility. Jacoby (2002) used a DD design to test whether intra-household resource allocation shifted in response to a school-feeding program, to neutralize the latter’s effect on child nutrition. Some schools had the feeding program and some did not, and some children attended school and some did not. The author’s DD estimate of impact was then the difference between the mean food-energy intake of children who attended a school (on the previous day) that had a feeding program and the mean for those who did not attend such schools, less the corresponding difference between attending and non-attending children found in schools that did not have the program.

Another example can be found in Pitt and Khandker (1998) who assessed the impact of participation in Bangladesh’s Grameen Bank (GB) on various indicators relevant to current and future living standards. GB credit is targeted to landless households in poor villages. Some of their sampled villages were not eligible for the program and within the eligible villages, some households were not eligible, namely those with land (though it is not clear how well this was
enforced). The authors implicitly use an unusual DD design to estimate impact.\textsuperscript{33} Naturally, the returns to having land are higher in villages that do not have access to GB credit (given that access to GB raises the returns to being landless). Comparing the returns to having land between two otherwise identical sets of villages — one eligible for GB and one not — will thus reveal the impact of GB credit. So the Pitt-Khandker estimate of the impact of GB is actually the impact on the returns to land of taking away village-level access to the GB.\textsuperscript{34} 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 \textit{Trabajar} program), Ravallion et al. (2005) used a lighter survey instrument, with far fewer questions on relevant characteristics of participants and non-participants. This did not deliver plausible single-difference estimates using PSM when compared to the Jalan and Ravallion estimates using single-difference PSM for the same program on richer data. The likely explanation is that using the lighter survey instrument meant that there were many unobservable differences; in other words the conditional independence assumption of PSM was not valid. Given the sequence of the two evaluations, the key omitted variables in the later study were known — they mainly related to local level connections (as evident in memberships of various neighborhood associations and length of time living in the same barrio). However, the lighter

\textsuperscript{33} This is my interpretation; Pitt and Khandker (1998) do not mention the DD interpretation of their design given here. However, it is readily verified that the impact estimator implied by solving equations (4a-d) in their paper is the DD estimator described here. (Note that the resulting DD must be normalized by the proportion of landless household in eligible villages to obtain the impact parameter for GB.)

\textsuperscript{34} 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.
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.

This illustrates an important point about evaluation design. A trade-off exists between the resources devoted to 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 \textit{ex ante} about the determinants of program placement (both on the side of program administrators and participants). If a single survey can be implemented that convincingly captures 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.

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

Even with good data, the DD assumption of time-invariant and additive selection bias is implausible for some anti-poverty programs in developing countries. DD will give a biased impact estimate if the subsequent outcome changes are a function of initial conditions that also
influenced the assignment of the sample between the two groups. Anti-poverty programs are often targeted to poor sub-groups, such as poor areas. And these same targeting criteria could well influence subsequent growth rates.\footnote{There is also the well-known bias in using DD for inferring long-term impacts of training programs that can arise when there is a pre-program earnings “Ashenfelter’s dip” (Ashenfelter, 1978).}

Figure 2 illustrates the point. Mean outcomes are plotted over time, before and after the intervention. The lightly-shaded circles represent the observed means for the treatment units, while the hatched circle is the counterfactual at date $t=1$. Panel (a) shows the initial selection bias, arising from the fact that the program targeted 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 potentially large bias in the DD estimate, as in panel (c).

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

The second example draws on the Pitt and Khandker (1998) study of Grameen Bank. Following my interpretation of the Pitt-Khandker method of assessing the impacts of GB credit, it is clear that the authors’ key assumption is that the returns to having land are independent of village-level GB eligibility. A bias will arise if GB tends to select villages that have either
unusually high or low returns to land. It seems plausible that the returns to land are lower in villages selected for GB, which may well be why they are poor in the first place, and low returns to land would also suggest to GB that such villages have a comparative advantage in the non-farm activities facilitated by GB credit. Then the Pitt-Khandker method will overestimate the impact of the Grameen Bank.

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

Notice that panel data are not necessary for calculating the DD impact estimator. All one needs is the set of four means that make up DD, and the means for each of the participant and non-participants groups (possibly after matching) need not be calculated for the same sample.
For example, recall the above interpretation of the Pitt-Khandker evaluation of Grameen Bank; one does not need longitudinal observations of the villages with and without the GB.

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

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

Pre-intervention baseline data are sometimes unavailable. Safety-net interventions such as workfare programs and social funds often have to be set up quickly in response to a macroeconomic or agro-climatic crisis, and it is not feasible to delay the operation in order to do a baseline survey. (Nor is randomization feasible in such settings.) Nonetheless, under certain conditions, impacts can still be identified by observing participants’ outcomes in the absence of the program after the program rather than before it.
To see what is involved, write the gain for unit i at date t as $G_{it} = Y_{it}^{T*} - Y_{it}^{C*}$ where $Y_{it}^{T*}$ is the true value of the outcome variable and $Y_{it}^{C*}$ is the true value of the counter-factual outcome for the participant. The observed values are (dropping the i subscripts):

$$Y_i^T = Y_{it}^{T*} + \eta_i + \varepsilon_i^T \tag{9.1}$$
$$Y_i^C = Y_{it}^{C*} + \eta_i + \varepsilon_i^C \tag{9.2}$$

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

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

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

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

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

In the standard DD set-up, date 1 precedes the intervention, $G_1 = 0$ and $DD$ gives the mean current gain to participants at $t=2$, $DD = E(G_2 | D_2 = 1)$. However, in this case, the program is in operation at date 1. The scope for identification arises from the fact that some participants at date 1 subsequently drop out of the program.

The triple-difference estimator proposed by Ravallion et al. (2005) is the difference between the double differences for stayers and leavers:

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

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

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

In applying the above approach, Ravallion et al. (2005) examine what happens to participants’ incomes when they leave Argentina’s Trabajar program as compared to the incomes of continuing participants, after netting out economy-wide changes, as revealed by a
matched comparison group of non-participants. The authors find partial income replacement, amounting to one-quarter of the *Trabajar* wage within six months of leaving the program, though rising to one half in 12 months. Thus there is evidence of a post-program “Ashenfelter’s dip,” namely when earnings drop sharply at retrenchment, but then recover.

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

9. **Instrumental variables**

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

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

$$D_i = \gamma Z_i + \nu_i$$

(14)

To simplify the exposition, I focus on the common effects specification (equation 6), for which the reduced form equation for outcomes is simply:
\[ Y_i = \pi Z_i + X_i \beta^C + \mu_i \]

where \( \pi = (\beta^T - \beta_0^T) \gamma \) and \( \mu_i = (\beta^T - \beta_0^T) \nu_i + \varepsilon_i \). When it exists, the Instrumental Variables Estimator (IVE) for the impact parameter is \((\hat{\beta}_0^T - \hat{\beta}_0^T)_{IVE} = \hat{\pi}_{OLS} / \hat{\gamma}_{OLS} \) (in obvious notation). A variation on this is to estimate (14) as a nonlinear binary response model (a probit or logit) and use the predicted values as the IV for program placement in equation (7). (The first stage regression can include \( X \) as well.)\(^{36}\)

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

Like all the preceding non-experimental methods, the IVE requires an untestable conditional independence assumption, though it is a different assumption to PSM or OLS; in the case of IVE this is the exclusion restriction.\(^{37}\) Also note that this is not strictly required when a nonlinear binary response regression is used for the first stage. Then the model is identified off

\(^{36}\) A good discussion of this estimator can be found in Wooldridge (2002, Chapter 18).
\(^{37}\) If \( Z \) is a vector (with more than one variable) then the model is over-identified and one can test whether all but one of the IVs is significant when added to the main equation of interest (7). However, one must still leave one IV and so the exclusion restriction is un-testable.
the nonlinearity of the first stage regression. (If instead a linear-probability model had been used then identification would be lost.) 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. However, it is worrying when identification rests on a somewhat ad hoc assumption about the distribution of an error term. Avoiding this requires a justification for excluding $Z_i$ from (7). We shall return to this issue.

There are similarities too. As with OLS, the validity of causal inferences for (parametric) IVE rests on mostly *ad hoc* functional form assumptions for the outcome regression. Note also that the first stage equation (14) echoes the first stage of PSM method. However, IVE is arguably less demanding of our ability to model the program’s assignment than PSM; while the instrumental variable $Z$ needs to be a significant predictor for participation, one is not typically as concerned about a low $R^2$ in the first stage equation for IVE than for the model used to estimate propensity scores for matching.

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

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

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 in developing country settings in which 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 rural Bangladesh) in Basu et al. (2002).

The validity of widely used exclusion restrictions can be particularly questionable with only a single cross-sectional data set; while one can imagine many variables that are correlated with placement, such as geographic characteristics of an area, it is questionable that those variables are uncorrelated with outcomes given placement. There can be more potential for identification with longitudinal (panel) data, on the assumption that lagged effects die out sufficiently rapidly to justify the exclusion restrictions. Examples of this approach include Rosenzweig and Wolpin (1986), Pitt et al., (1995) and Jalan and Ravallion (2002).

Where do we find an IV? There are essentially two sources, experimental design features and theoretical arguments about the determinants of program placement and outcomes. The following discussion considers these in turn.
Partially randomized designs as a source of instrumental variables: As noted in section 4, it is often the case in social experiments that some of those randomly selected for the program do not want to participate. While finding a valid IV is often difficult, the randomized assignment is a natural choice in this case. Here the exclusion restriction is plausible, namely that being randomly assigned to the program only affects outcomes via actual participation. Looking more closely at how this works is instructive for understanding IVE in this context.38

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

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

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

Under what conditions does $\pi / \gamma$ give the mean impact, as would be obtained by simply comparing the outcome means for treatment and control observations in a social experiment?39

38 For a more general characterization of the theoretical conditions under which an IVE delivers the mean impact of a program see Angrist et al. (1996). Also see the discussion in Dubin and Rivers (1997).

39 This is essentially the same question addressed by Angrist et al. (1996) who provide a more complete discussion of the conditions under which causal effects can be identified using instrumental variables. There is a similarity to the analysis in Dubin and Rivers (1993).
By construction, $\pi$ is the weighted mean of $E(Y|Z = 1, D = 1) - E(Y|Z = 0, D = 1)$ and

$E(Y|Z = 1, D = 0) - E(Y|Z = 0, D = 0)$ with weights $\gamma$ and $1 - \gamma$ respectively. This can be re-written as:

$$\pi = \gamma \text{ATE} - \gamma [E(Y|Z = 0, D = 1) - E(Y|Z = 0, D = 0)]$$

$$+ E(Y|Z = 1, D = 0) - E(Y|Z = 0, D = 0)$$  \hspace{1cm} (17)

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

An example of the above approach to correcting for bias in randomized designs can be found in the aforementioned MTO experiment, in which randomly-selected inner-city families in US cities were given vouchers to buy housing in better-off areas. Naturally, not everyone offered such a voucher takes up the opportunity. The difference in outcomes (such as school drop-out rates) only reveals 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).

Another example can be found in the aforementioned Proempleo experiment. Recall that this included a training component that was assigned randomly. Under the assumption of perfect take-up or random non-compliance, neither the employment nor incomes of those receiving the training were significantly different to those of the control group 18 months after the experiment.
began. However, some of those assigned the training component did not want it, and this selection process was correlated with the outcomes from training. An impact of training was revealed for those with secondary schooling, but only when the authors corrected for compliance bias using assignment as the IV for treatment (Galasso et al., 2004).

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

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

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

40 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.
either households or the community centers. Amongst the justifications they give for their choice of IV, the authors note that survey respondents who have moved recently never identified the desire to move closer to a community center as one of the reasons for choosing their location (even though this was one of the options). Tellingly, they also note that if their results were in fact driven by endogeneity of their IV then they would find (spurious) effects on variables that should not be affected, such as child birth weight. However, they do not find such effects, supporting the choice of IV.

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

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

The third set of examples exploit **discontinuities in program design**, as discussed in section 6. Here the LATE is in the neighborhood of a cut-off for program eligibility. An example of this approach can be found in Angrist and Lavy (1999) who assessed the impact on
school attainments in Israel of class size. For identification they exploited the fact that an extra
teacher (in Israel) was assigned when the class size went above 40. Yet there is no plausible
reason why this cut-off point in class size would have an independent effect on attainments, thus
justifying the exclusion restriction. The authors find sizeable gains from smaller class sizes,
which were not evident using OLS.

Another example is found in Duflo’s (2003) study of the impacts of old-age pensions in
South Africa on child anthropometric indicators. Women only become eligible for a pension at
age 60, while for men it is 65. It is implausible that there would be a discontinuity in outcomes
(conditional on treatment) at these critical ages. Following Case and Deaton (1998), Duflo 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.

Notice again that this assumes we know eligibility, which is not always the case in the
data sets available (Section 6). 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.

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

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

10. Learning more from evaluations

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

Can the lessons from an evaluation be scaled up? The context of an intervention often matters to its outcomes, thus confounding inferences for “scaling up” from an impact evaluation. These “external validity” concerns relate to both experimental and non-experimental evaluations. If one allows for contextual factors then it can be hard to make meaningful generalizations for scaling up and replication from trials. The same program works well in one village but fails hopelessly in another. This is illustrated by the results of Galasso and Ravallion (2005)
studying Bangladesh’s Food for Education Program. The program worked well in reaching the poor in some villages but not others, even in relatively close proximity.

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

A further concern about external validity is that while partial equilibrium assumptions may be fine for a pilot, that can cease to be so when the program is scaled up nationally, and general equilibrium effects become important (sometimes called “feedback” or “macro” effects in the evaluation literature). For example, an estimate of the impact on schooling of a tuition subsidy based on a randomized trial may be deceptive when scaled up, given that the structure of returns to schooling will alter. 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

---

41 Heckman et al., (1998) demonstrate that the partial equilibrium analysis can greatly overestimate the impact of a tuition subsidy once relative wages adjust, though Lee (2005) finds a much smaller difference between the general and partial equilibrium effects of a tuition subsidy in a slightly different model.
validity concern stems from the context-specificity of trials; outcomes in the context of the trial may differ appreciably (in either direction) once the intervention is scaled up and prices and wages respond.

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

The realized impacts on scaling up can also differ from the trial results (whether randomized or not) because the socio-economic composition of program participation varies with scale. Ravallion (2004a) discusses how this can happen, and presents results from a series of country case studies, all of which suggest that the incidence of program benefits 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).

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

An alternative approach is to probe more deeply into why a program has (or does not have) impact in a specific context, as a basis for inferring whether it would work in a different context. Here better use can often be made of intermediate indicators. The most common evaluation design identifies a relatively small number of “final outcome” indicators, and aim to assess the program’s impact on those indicators. However, instead of using only final outcome indicators, one may choose to also study impacts on certain intermediate indicators of behavior. For example, the inter-temporal behavioral responses of participants in anti-poverty programs are of obvious relevance to understanding their impacts. An impact evaluation of a program of compensatory cash transfers to Mexican farmers found that the transfers partly invested, with second-round effects on future incomes (Sadoulet et al., 2001). Similarly, Ravallion and Chen (2005) found that participants in a World Bank poor-area development program in China saved a large share of the income gains from the program (as estimated using the matched double-difference method described in section 7). Identifying responses through savings and investment provides a clue to understanding current impacts on living standards and the possible future welfare gains beyond the project’s current life span. Instead of focusing solely on the agreed
welfare indicator, one collects and analyzes data on a potentially wide range of intermediate indicators relevant to understanding the processes determining impacts.

As an aside, this also illustrates a common concern in evaluation studies, given behavioral responses, 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 other sources of information external to the evaluation. There are many possible sources, including qualitative research (intensive interviews with participants and administrators). The essential idea is to “test” the assumptions made by an intervention. This is sometimes called “theory-based evaluation,” though that is hardly an ideal term given that non-experimental identification strategies for mean impacts are often theory-based (as discussed in the last section). Weiss (2001) illustrates this approach in the abstract in the context of evaluating the impacts of community-based anti-poverty programs. An example is found in an evaluation of social funds (SFs) by the World Bank’s Operations Evaluation Department, as summarized in Carvalho and White (2004). While the overall aim of a SF is typically to reduce poverty, the OED study was interested in seeing whether SFs worked the way that was intended by their designers. For example, did local communities participate? Who participated? Was there “capture” of the SF by local elites (as

See the discussion on “mixed methods” in Rao and Woolcock (2003).
some critics have argued)? Building on Weiss (2001), the OED evaluation identified a series of key hypothesized links connecting the intervention to outcomes and tested whether each one worked. For example, in one of the country studies for the OED evaluation of SFs, Rao and Ibanez (2003) tested the assumption that a SF works by local communities collectively proposing the sub-projects that they want; for a SF in Jamaica, the authors found that the process was often dominated by local elites.

In practice, it is very unlikely that all the relevant assumptions are testable (including alternative assumptions made by different theories that might yield similar impacts). Nor is it clear that the process determining the impact of a program can always be decomposed into a neat series of testable links within a unique causal chain; there may be more complex forms of interaction and simultaneity that do not lend themselves to this type of analysis. So the so-called “theory-based evaluation” approach cannot be considered a serious substitute for assessing impacts on final outcomes by credible (experimental or non-experimental) methods, though 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 paper are for assigned programs, but have little obvious role for economy-wide programs in which no explicit assignment process is evident, or if it is, the spillover effects are likely to be pervasive. When some countries get the economy-wide program but some do not, cross-country comparative work (such as growth regressions) can reveal impacts. That identification task is often difficult, notably because there are typically latent factors at country level that simultaneously influence outcomes and whether a country adopts the policy in question. And even when the identification strategy is accepted, carrying the generalized lessons from cross-country regressions to inform policy-making in any one country can be highly problematic. There are also a number of promising examples of how simulation tools for economy wide policies such as Computable General Equilibrium models can be combined with household-level survey data to assess impacts on poverty and inequality. These simulation methods make it far easier to attribute impacts to the policy change, though this advantage comes at the cost of the need to make many more assumptions about how the economy works.

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

One issue is the choice of counterfactual. The classic formulation of the evaluation problem assesses mean impacts on those who receive the program, relative to counterfactual outcomes in the absence of the program. However, this may fall well short of addressing the concerns of policy makers. While common practice is to use outcomes in the absence of the

---

43 See, for example, Bourguignon et al. (2003) and Chen and Ravallion (2004).
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.44

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.

44 For another example of the same result see Murgai and Ravallion (2005).
Poverty measures are still mean impacts, albeit for a summary statistic about the marginal distribution of outcomes. They are also conditional means, in that the impacts vary across the sample of participants, but only as functions of the observables used as control variables. More generally, we may want to know about the joint distribution of $Y^T$ and $Y^C$. We cannot know this from a social experiment, which only reveals net counterfactual mean outcomes for those treated; $ATET$ gives the mean gain net of losses amongst participants. However, there are anti-poverty programs in practice that can generate losses amongst participants; for example, an agricultural development project invariably imposes costs on participants (such as their own time) yet with uncertain future gains (such as due to the risk that prices will fall).

This points to the need for estimates of a wider range of impact parameters than the simple mean impact ($ATET$) or the marginal distribution of impacts. (Recall that 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.) Instead of focusing solely on the net gains to the poor (say) we may ask how many losers there are amongst the poor, and how many gainers. Some interventions may yield losers even though mean impact is positive and policy makers will understandably want to know about those losers, as well as the gainers. (This can be true at any given poverty line.) Thus one can relax the “anonymity” or “veil of ignorance” assumption of traditional welfare analysis, whereby outcomes are judged solely by changes in the marginal distribution (Carneiro et al., 2001). This is clearly of greater relevance to some applications than others. For example, trade reforms are very likely to generate both losers and gainers at any given level of living, given general equilibrium effects and the heterogeneity in net trading positions in relevant markets (Ravallion, 2004b). But even for policies with only non-negative impacts, horizontal impacts at given levels of living can be expected. Methods for estimating the
joint distribution of $Y^T$ and $Y^C$ are developed in Heckman et al. (1997a). The analogous estimator to OLS applied to the classic “common effects” specification (equation 7) is a random coefficients estimator in which the coefficient on treatment dummy variable contains a stochastic components.

When the policy issue is whether to expand or contract a given program at the margin, the classic estimator of mean-impact on the treated (by experimental or non-experimental methods) is actually of little or no interest. The problem of estimating the marginal impact of a greater duration of exposure to the program on those treated was considered in section 8, using the example of comparing “leavers” and “stayers” in a workfare program (Ravallion et al., 2005). Another example can be found in the study by Behrman et al. (2004) of the impacts on children’s cognitive skills and health status of longer exposure to a preschool program in Bolivia. The authors provide an estimate of the marginal impact of higher program duration by comparing the cumulative effects of different durations using a matching estimator. In such cases, selection into the program is not an issue, and we do not even need data on units who never participated. The discontinuity design method discussed in section 6 (in its non-parametric form) and section 9 (in its parametric IV form) is also delivering an estimate of the marginal gain from a program, namely the gain when the program is expanded (or contracted) by a small change in the eligibility cut-off point.

A deeper understanding of the factors determining outcomes in ex post evaluations can also help in simulating the likely impacts of changes in program or policy design ex ante. Naturally, ex ante simulations require many more assumptions about how an economy works.45

As far as possible one would like to see those assumptions anchored to past knowledge built up

---

45 For a useful overview of ex ante methods see Bourguignon and Ferreira (2003).
from rigorous *ex post* evaluations. For example, by combining a randomized evaluation design with a structural model of education choices, Todd and Wolpin (2002) and Attanasio et al. (2004) are able to greatly expand the set of policy-relevant questions about the design of *PROGRESA* that a conventional evaluation can answer. By modeling the determinants of school choice — exploiting the randomized design for identification — they show that a budget-neutral switch of the enrolment subsidy from primary to secondary school would have delivered a net gain. *PROGRESA* had impact, but it could have had more impact.

11. **Conclusions**

Two main lessons for future evaluations of anti-poverty programs emerge from this survey. Firstly, no single evaluation tool can claim to be ideal in all circumstances. While randomization can be a powerful tool for assessing impact, it is neither necessary nor sufficient for a good evaluation. While economists have sometimes been too uncritical of their non-experimental identification strategies, credible means of isolating at least a share of the exogenous variation in an endogenously placed program can still be found in practice. Good evaluations draw pragmatically from the full range of tools available, often combining methods: randomizing some aspects and using econometric methods to deal with the non-random elements, using randomized elements of a program as a source of instrumental variables, or by combining score matching methods with longitudinal observations to try to eliminate matching errors with imperfect data. Good evaluations typically also require that the evaluator is involved from the programs’ inception and is very well informed about how the program works on the ground; the features of program design and implementation can sometimes provide important clues for assessing impact by non-experimental means.
Secondly, even putting internal validity concerns to one side, it is unlikely that the tools of counter-factual analysis for mean impacts on well-defined outcome variables are ever going to be sufficient for drawing reliable lessons for future development projects and policies. The context in which a program is placed can exercise a powerful influence on outcomes. This points to the need for a deeper understanding of why a program does or does not have impact. It also calls for an eclectic approach drawing on various sources, including replications across differing contexts when feasible, and testing the assumptions made in a program’s design, such as by tracking intermediate variables of relevance or by drawing on supplementary theories or evidence external to the evaluation. In drawing useful lessons for anti-poverty policy, we need a richer set of impact parameters than has been traditional in evaluation practice, including distinguishing the impacts on gainers from losers at any given level of living. The choice of parameters to be estimated in an evaluation must ultimately depend on the policy question to be answered; for policy makers this is a mundane point, but for evaluators it seems to be ignored too often.
Figure 1: Region of common support

Density for non-participants

Density for participants

Region of common support
Figure 2: Bias in double-difference estimates for a targeted anti-poverty program

(a)

(b)

(c)

Selection bias
References


Carneiro, Pedro, Karsten Hansen and James Heckman, 2001, “Removing the Veil of Ignorance


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.


Ravallion, Martin, Emanuela Galasso, Teodoro Lazo and Ernesto Philipp, 2005, “What Can Ex-Participants Reveal About a Program’s Impact?” *Journal of Human Resources*,


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.


