

Chapter 5

Assessing the Poverty Impact of an Assigned Program

Martin Ravallion

5.1 Introduction

Some public programs are assigned more-or-less exclusively to certain observational units. These may be people, households, villages or larger geographic areas. The key thing is that some units get the program and some do not. This chapter reviews tools that can help assess the impacts of such a program, judged against its agreed objectives. The following are examples of the types of programs that can be assessed with the tools discussed in this chapter:

- A social fund asks for proposals from community organizations, with preference for proposals from poor areas. Some areas do not apply, and some do, but are rejected.
- A workfare program entails extra earnings for participating workers, and gains to the residents of the areas in which the work is done. Others receive nothing.
- Infrastructure projects (road or water connections, for example) are targeted to areas that are both poor and poorly endowed in that infrastructure. Other areas do not participate.

Notice that in each of these examples, there may be some indirect (or “second-round”) impacts on non-participants. 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 following discussion will assume that the program is already in place, which makes this a case of *ex-post* impact assessment.¹ 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. This can

¹ For an example of an *ex-ante* impact assessment of an anti-poverty program see Ravallion (1999).

greatly facilitate the evaluation, such as by allowing pre-intervention data to be collected on probable participants and non-participants.

The indicators by which a program is to be assessed are taken to be given, as appropriate to the type of program. For example, for direct anti-poverty programs we are usually concerned about the impacts on incomes of the participants and possibly also effects on other indicators such as school attendance. Knowing impact is of obvious interest in its own right as a means of measuring the aggregate benefits from the program. However, when reducing poverty is the overall objective of the program we also want to know the incidence of the welfare gains. We can only know that by knowing the welfare impact at given values of the pre-intervention welfare indicator. To know incidence we must know impact.

Figure 5.1 is an example of the type of “impact-incidence” assessment we might make for an assignable anti-poverty program; in this example, it is Argentina’s Trabajar program (a combination of workfare program and social fund). The figure gives the poverty incidence curves (PICs) showing how the headcount index of poverty (% below the poverty line) varies across a wide range of possible poverty lines (when that range covers all incomes we have the standard cumulative distribution function). The vertical line is an indicative poverty line for Argentina. The figure also gives the estimated counter-factual PIC, after deducting the imputed income gains from the observed (post-intervention) incomes of all the sampled participants. Thus we can see the gain at each percentile of the distribution (looking horizontally) or the impact on the incidence of poverty at any given poverty line (looking vertically).

In this chapter we will learn how Figure 5.1 is estimated.² Along the way we will also learn about other tools used for impact assessments of assignable programs. The methods share some common features related to their data requirements, as summarized in Box 5.1.

² This article does not attempt to review all of the tools that have been used in the past for impact evaluation. The focus will be on more recent developments that appear likely to have relevance to the assessment of anti-poverty programs, active labor market and other “social protection” programs in developing country-settings. More comprehensive discussions of the methods found in practice can be found in Mofitt (1991), Meyer (1995), Blundell and Costa Dias (2000) and Ravallion (2001).

5.2 Randomization

Clearly we are going to need data on an appropriate outcome indicator for the participants. That will not be sufficient, however, since to assess impact we will also need some way of inferring the counter-factual of what we expect would have been the value of the outcome indicator in the absence of the program. This calls for data on non-participants.

But even with good data on outcome measures for both participants and non-participants, retrieving a reliable estimate of the program's impact is far from easy. The main reason is that public programs are generally not assigned randomly across the population of units. So we cannot attribute to the program the observed differences in measured outcome indicators between units who receive the program and those who do not. The measured differences we see in the data could just be due to the fact that the program participants were purposely selected. (This is often called "selection bias.")

This is not a problem with randomized assignment (a genuine experiment), since everyone then has the same chance *ex-ante* of receiving the program. The distributions of both observed and unobserved attributes prior to the program intervention are the same, whether or not a unit receives the program. Then the observed *ex-post* differences in the outcome indicators are attributable to the program.

Randomization is the theoretical ideal, and a natural benchmark for assessing non-experimental (sometimes called "quasi-experimental") methods. There are sometimes opportunities for randomizing the assignment of an anti-poverty program, possibly on a pilot basis. A number of evaluations of active labor market programs have used randomized assignment. In the case of training programs, two examples are the US Job Training Partnership Act (see, for example, Heckman et al., 1997), and the US National Supported Work Demonstration (studied by Lalonde, 1986, and Dehejia and Wahba, 1999, amongst others). For wage subsidy programs, randomized evaluations have been done by Burtless (1985), Woodbury and Spiegelman (1987) and Dubin and Rivers (1993) — all for targeted wage subsidy schemes in the US. A recent example for a World Bank supported program can be found in Galasso et al., (2001) who randomized a wage subsidy and training program for assisting workfare participants in Argentina to find regular, private-sector jobs. Besides labor market programs, randomization has also been used in assessing (*inter alia*) residential relocation programs (Katz et al., 2001) and school voucher programs (Angrist et al., 2001).

In practice, it is sometimes the case that the chosen participants do not want to comply with the randomized assignment. This is to be expected in almost any social experiment. We typically want to know the impact of receiving the treatment, which clearly cannot be assumed to be exogenous when there is selective compliance. Angrist et al., (1996) have addressed this issue and shown how one can correct for selective compliance by using the randomized assignment as the instrumental variable for treatment in a regression for the outcome measure. Applications can be found in Galasso et al., (2001) and Katz et al., (2001).

However, it is frequently the case in practice that randomization is not a feasible option. The government does not want to randomly assign the program, but rather to purposively target it to certain groups, such as the income poor or those with low current access to the facilities provided by the program. What can be done to assess impact when it is known that a program was not randomly placed? The rest of this chapter aims to provide an overview of the best methods currently available for addressing this question.

5.3 Propensity-score matching methods

Along with randomization, matching is one of the oldest tools of evaluation. The idea is to find a comparison group that looks like the treatment group in all respects except one: the comparison group did not get the program. However, the problem in practice was always how to define “looks like;” there are potentially many characteristics one might look for to match on, and it was not clear whether a match has to be “identical” in all these characteristics, and (if not) how each characteristic should be weighted.

The method of Propensity-Score Matching (PSM) due to Rosenbaum and Rubin (1983) can justifiably claim to be the solution to this problem, and thus to be the observational analog of a randomized experiment. The method balances the observed covariates between the treatment group and a control group (sometimes called “comparison group” for non-random evaluations) based on similarity of their predicted probabilities of receiving the treatment (called their “propensity scores”). The difference between PSM and a pure experiment is that the latter also assures that the treatment and comparison groups are identical in terms of the distribution of unobserved characteristics. Box 5.2 summarizes the steps in PSM.

The key to PSM is understanding and modeling the assignment mechanism for the program. Two groups are identified: those households that have the treatment (denoted $D_i = 1$ for household i) and those that do not ($D_i = 0$). Treated units are matched to non-treated units on the basis of the propensity score:

$$P(X_i) = \text{Prob}(D_i = 1 | X_i) \quad (0 < P(X_i) < 1) \quad (1)$$

where X_i is a vector of pre-exposure control variables. The choice of variables must be based on knowledge of the program, and will often also be informed by theories of the economic, social or political factors influencing the assignment of a program. Clearly if your 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 the results of a pure experiment.

PSM uses $P(X)$ (or a monotone function of $P(X)$) to select controls for each of those treated. 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 X_i , then outcomes are also independent of participation given

$P(X_i)$, just as they would be if participation were assigned randomly.³ Exact matching on $P(X)$ implies that the resulting matched control and treated subjects have the same distribution of the covariates. This is the sense in which PSM is the observational analog to an experiment; just like an experiment, PSM equalizes the probability of participation across the population — the difference is that with PSM it is the conditional probability, conditional on the X variables.

Common practice is to use the predicted values from standard logit or probit models to estimate the propensity score for each observation in the participant and the comparison-group samples.⁴ Using the estimated propensity scores, $\hat{P}(X)$, matched-pairs are constructed on the basis of how close the scores are across the two samples. The “nearest neighbor” to the i ’th participant is defined as the non-participant that minimizes $[p(X) - p(X_j)]^2$ over all j in the set of non-participants, where $p(X_k)$ is the predicted odds ratio for observation k i.e., $p(X_k) = \hat{P}(X_k)/(1 - \hat{P}(X_k))$. One can apply caliper bounds; for example, matches might only be accepted if $[p(X) - p(X_j)]^2$ is less than (say) 0.001.

Letting ΔY_j denote the gain in a welfare indicator for the j ’th unit attributable to access to the program, the PSM estimator of mean impact is:

$$\Delta \bar{Y} = \sum_{j=1}^T \omega_j (Y_{j1} - \sum_{i=1}^C W_{ij} Y_{ij0}) \quad (2)$$

where Y_{j1} is the post-intervention welfare indicator, Y_{ij0} is the outcome indicator of the i ’th non-treated matched to the j ’th treated, T is the total number of treatments, C is the total number of non-treated households, ω_j ’s are the sampling weights used to construct the mean impact estimator, and the W_{ij} ’s are the weights applied in calculating the average income of the matched non-participants.

There are several weights that one can use, ranging from nearest-neighbor weights to non-parametric weights based on kernel functions of the differences in scores (Heckman et al., 1997, 1998).⁵ It is a good idea to use more than just the nearest neighbor; for example, one can use the mean for the nearest five neighbors, i.e., take the average outcome measure of the closest five matched non-participants as the counter-factual for each participant.⁶

One can also use a regression-adjusted estimator. This assumes a conventional linear model for outcomes in the matched comparison group, $Y_0 = X\beta_0 + \mu_0$ in obvious notation. (The regression is only

³ Assumption (ii) is sometimes referred to in the literature as the “conditional independence” assumption, and sometimes as “strong ignorability.”

⁴ Dehejia and Wahba (1999) report that their PSM results are robust to alternative estimators and alternative specifications for the logit regression.

⁵ Jalan and Ravallion (2001b) discuss the choice further. They used a range of weighting schemes, including nearest neighbor, nearest five neighbors and a kernel-based weighting scheme in which the weight is a function of the absolute difference in propensity scores. They found that their results for estimating income gains from an anti-poverty program are reasonably robust to the choice. However, that may not be so in other applications.

⁶ 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.

run for the matched comparison group, so it is not contaminated by the endogeneity of access to the program.) The impact estimator in this case is then defined as:

$$\Delta \bar{Y} = \sum_{j=1}^T \omega_j [(Y_{j1} - X_j \hat{\beta}_0) - \sum_{i=1}^C W_{ij} (Y_{ij0} - X_i \hat{\beta}_0)] \quad (3)$$

where $\hat{\beta}_0$ is the Ordinary Least Squares (OLS) estimate for the comparison group sample.

Conditional mean impact estimators can be obtained by calculating equation (2) conditional on certain observed characteristics. For anti-poverty programs one is interested in comparing the conditional mean impact across different pre-intervention income levels. 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 know where each participant would have been in the distribution of income without the program. Thus one can construct the empirical and counter-factual PICs, as in Figure 5.1. (These can be smoothed, using locally weighted means for example.) Box 5.3 summarizes the steps in how this is done, and how the results should be interpreted to form a qualitative assessment of poverty impact. This can all be repeated for multiple programs, which can then be compared with each other.

One can also construct the concentration curve, showing the cumulative share of benefits going to the poorest $x\%$ of the population, ranked by household income per person, with x ranging from 1 to 100. Figure 5.2 give the concentration curve for the earnings gains from the Trabajar program. Of course, the concentration curve does not give the impact on poverty; for that purpose one needs the PIC, as in Figure 5.1.

5.4 How does PSM compare to other methods?

Probably the most common method in practice for assessing the impact of an assigned program is to compare average outcome indicators between units that have the program and those that do not. For example, past methods of assessing health gains from water and sanitation have often compared villages with piped water and those without. Similarly, assessments of the impacts of providing new rural roads often compare the incomes or other outcome indicators of villages with roads and those without. Clearly failure to control for differences in the pre-intervention characteristics of the participants and non-participants could severely bias such comparisons. Van de Walle (2002) gives an example for rural road evaluation in which a naïve comparison of the incomes of villages who get the program with those that do not indicates large income gains when in fact there are none.⁷

Another method found in the literature is to run a regression of the outcome indicator on a dummy variable for treatment or facility placement, allowing for the observable covariates entering as linear

⁷ 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.

controls. The widely used OLS regression method requires the same conditional independence assumption as PSM, but also imposes arbitrary functional form assumptions concerning the treatment effects and the control variables. By contrast, PSM does not require a parametric model linking program participation to outcomes. 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; see, for example, the analysis of the interaction effects between income and education in influencing the child-health gains from access to piped water in Jalan and Ravallion (2002a).

A variation on this regression method is to use an instrumental variables estimator (IVE), treating placement as endogenous. This method also makes an untestable conditional independence assumption; in the case of IVE this is the exclusion restriction that the instrumental variable is independent of outcomes given participation. And again the validity of causal inferences rests on the *ad hoc* functional form assumptions required by standard (parametric) IVE. Under these assumptions, IVE identifies the causal effect robustly to unobserved heterogeneity. The validity of the exclusion restriction required by IVE is particularly questionable with only a single cross-sectional data set; while one can imagine many variables that are correlated with placement, such as geographic characteristics of an area, it is questionable on *a priori* grounds that those variables are uncorrelated with outcomes given placement.

PSM also differs from commonly-used regression methods with respect to the sample used. In PSM one confines attention to the matched sub-samples; unmatched comparison units are dropped. (In the terminology of the literature on PSM, matching is confined to the region of “common support”, where the “support” refers to the estimated propensity scores.) 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 misspecification of the regression function, than those based on matched samples.

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

The possibility that some treatment units may have to be dropped for lack of sufficiently similar comparators points to the possibility of a trade off between two possible sources of bias in the resulting estimates of the mean impact. On the one hand, there is the need to assure comparability in terms of initial characteristics, which speaks to the importance of assuring common support. 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 common support; this is a well known problem in the evaluation

literature.⁸ Recognizing this trade-off, it is wise to check robustness of your estimates to only eliminating non-participating units that are outside the propensity-score range found for treatment units, while retaining the original sample of treatment villages.⁹

There has been some recent work comparing PSM with other methods. A classic study by Lalonde (1986) found large biases in non-experimental methods when compared to a randomized evaluation of a US training program. On the same data set, Dehejia and Wahba (1999) found that propensity-score matching achieved a good approximation — much better than the non-experimental methods studied by Lalonde.¹⁰

5.5 Double difference

A popular approach to non-experimental evaluations in the literature is the double difference (or “difference-in-difference”) (DD) method. This compares treatment and comparison groups in terms of outcome changes over time relative to the outcomes observed for a pre-intervention baseline. DD allows for conditional dependence in the levels arising from additive time-invariant latent heterogeneity. Box 5.4 summarizes the steps in constructing a DD estimate of program impacts.

Since PSM optimally balances observed covariates between the treatment and comparison groups, it is the obvious method for selecting the comparison group in DD studies. The changes over time in the outcome indicator will no doubt contain heterogeneity in observables which would bias an unmatched DD.¹¹ PSM is the obvious method to clean this out prior to doing the differencing. If there is no observable heterogeneity in the differences (i.e., it has all been washed out by differencing) then there is no gain from matching on top of DD. 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.

Nonetheless, DD estimators have their limitations. In some circumstances it is implausible that the selection-bias (due to unobserved heterogeneity) is time invariant. There is a potential bias in DD estimators when the changes over time are a function of initial conditions that also influence program placement. There is also the well-known bias for inferring long-term impacts that can arise when there is a pre-program earnings dip (known as “Ashenfelter’s dip” after Ashenfelter, 1978).

For safety-net interventions, such as workfare programs, that have to be set up quickly in response to a macroeconomic or agro-climatic crisis, it is often unfeasible to delay the operation in order to do a baseline survey. Nor is randomization usually feasible in such settings. Suppose instead that we

⁸ Also see the discussion of the problem of “nonoverlapping support bias” in Heckman et al. (1997, 1998).

⁹ For further discussion and an example see Chen and Ravallion (2003).

¹⁰ Also see Heckman et al, (1998) and Smith and Todd (2001), who question the robustness of the Dehejia and Wahba PSM estimates to the choices made in sample selection and model specification.

¹¹ For example, Jalan and Ravallion (1998) show that this can seriously bias evaluations of poor-area development programs that are targeted on the basis of initial geographic characteristics that also influence the growth process.

follow up samples of participants and non-participants over time, post-intervention, and that some participants become non-participants. What can we then learn about the program's impacts?

The approach proposed by Ravallion et al., (2001) is to examine what happens to participants' incomes (or other welfare indicator) when they leave the program, and to compare this with the incomes of continuing participants, after netting out economy-wide changes, as revealed by a matched comparison group of non-participants. The authors wanted to estimate the net income gain to participants, net of their foregone income from the work displaced by the program. The standard DD estimate of program impact is the difference in the income gains over time between a treatment group of program participants and the matched comparison group of non-participants. The double-matched triple difference estimator of Ravallion et al., is the difference between the value of the double difference (between matched participants and non-participants) for the matched stayers and leavers. The difference between the program's benefit level and the triple-difference estimate of impact gives an estimate of the mean gain to participants.

While this approach is feasible without a baseline survey, it brings its own problems. Firstly, while differencing over time can eliminate bias due to latent (time-invariant) matching errors, there remains a potential bias due to any selective retrenchment from the program based on unobservables. Ravallion et al., argue that the direction of bias can be determined under plausible assumptions. Secondly, there may well be a post-program "Ashenfelter's dip," namely when earnings drop sharply at retrenchment, but then recover. As in the pre-program dip, this is a potential source of bias in assessing the longer-term impact, although (as with the pre-program version) to the extent that the dip entails a welfare change it can still be relevant to assessing the short-term impact of a safety-net intervention. And the post-program dip is of interest in assessing the dynamics of recovery from retrenchment. To help address this issue one can follow up initial participants over multiple survey rounds (Ravallion et al., 2001).

Under certain conditions, this type of follow-up study of participants can identify the gains to current participants from a program. There are concerns about selection bias, and there is the problem that past participation may bring current gains to those who leave the program. Assuming these lagged gains are positive, the net loss from leaving the program will be less than the gain from participation relative to the counter-factual of never participating. Ravallion et al., derive a test for the joint conditions needed to identify the mean gains to participants from this type of study, also exploiting further follow-up surveys of past participants.

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

valid. However, it would appear that Ravallion et al., were able to satisfactorily address this problem by tracking households over time, even using their lighter survey instrument. It appears that the follow-up evaluation design was able to difference out the miss-matching errors. From the point of view of evaluation design, this suggests that a trade off exists between the resources devoted to cross-sectional data collection for the purpose of single-difference matching, versus collecting longitudinal data with a lighter survey instrument.

5.6 On behavioral responses

Behavioral responses to a program can often be identified using the same methods discussed above, but using instead some intermediate indicators of behavior as the “outcome” variable, rather than the actual outcome variable(s) relevant to the program’s objective(s).

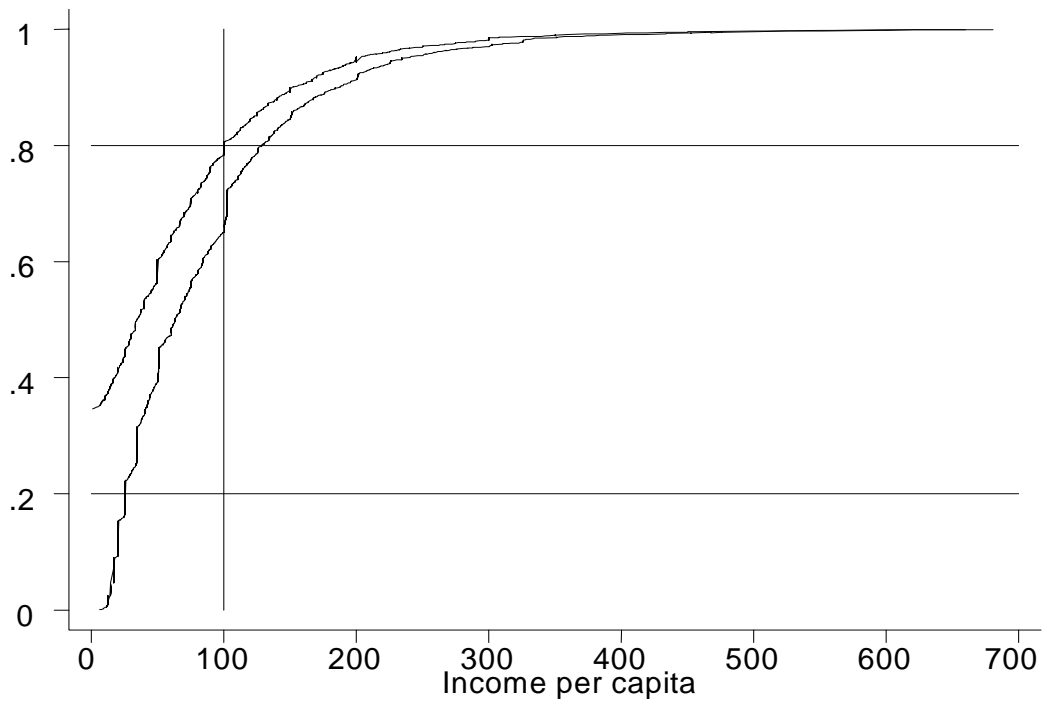
For example, Chen and Ravallion (2003) were interested in how much the participants in a World Bank supported poor-area development program in China saved the income gains from the program. It was agreed that the program’s aim was to raise living standards of the poor, but there was also a concern about how well this would be captured within the evaluation period. Identifying the savings response of participants provided a clue as to the possible future welfare gains beyond the project’s current life span. Indeed, Chen and Ravallion found that the participants saved about half of the income gain from the program, as estimated using the matched double-difference method described above.

This also illustrates a common concern in evaluation studies, given behavioral responses. The study period is rarely much longer than the period of the program’s disbursements. However, a share of the impact on peoples’ living standards may occur beyond the life of the project. This does not necessarily mean that credible evaluations will need to track welfare impacts over much longer periods than is typically the case — raising concerns about feasibility. But it does suggest that evaluations need to look carefully at impacts on partial intermediate indicators of longer-term impacts — such as incomes in the Chen-Ravallion example — 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.

5.7 Conclusions

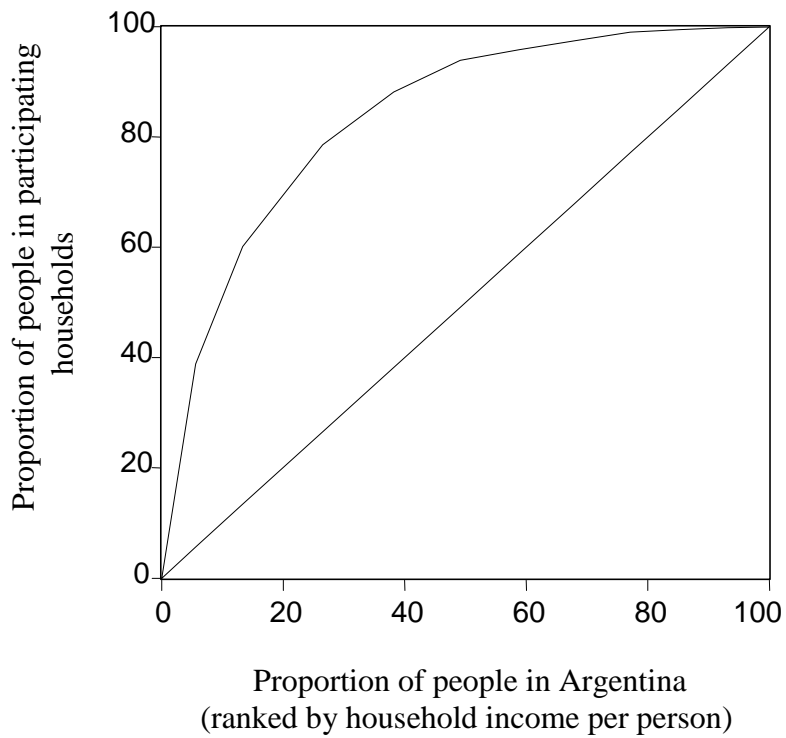
No single evaluation tool can claim to be ideal in all circumstances. The art of good evaluation is to draw carefully from the full range of tools available to deal pragmatically with the problem at hand in its specific context. The best evaluations often combine multiple methods: randomizing some aspects and using econometric methods to deal with the non-random elements, for example, or by combining score matching methods with longitudinal observations to try to eliminate matching errors with imperfect data. Good evaluations also need to be designed early in the program cycle, both to assure quality and to allow more rapid feedback into decision making about the program.

Figure 5.1: Poverty impacts of disbursements under Argentina's Trabajar program



Source: Jalan and Ravallion (2002b).

Figure 5.2: Concentration curve of participation in Argentina's Trabajar program



Box 5.1: Data for impact evaluation

- Know the program well. It is risky to embark on an evaluation without knowing a lot about the administrative/institutional details of the program; that information typically comes from the program administration.
- It helps a lot to have a firm grip on the relevant “stylized facts” about the setting. The relevant facts might include the poverty map, the way the labor market works, the major ethnic divisions, other relevant public programs, etc.
- Be eclectic about data. Sources can embrace both informal, unstructured, interviews with participants in the program as well as quantitative data from representative samples.
- However, it is extremely difficult to ask counter-factual questions in interviews or focus groups; try asking someone who is currently participating in a public program: “what would you be doing now if this program did not exist?” Talking to program participants can be valuable, but it is unlikely to provide a credible evaluation on its own.
- One also needs data on the outcome indicators and relevant explanatory variables. You need the latter to deal with heterogeneity in outcomes conditional on program participation. Outcomes can differ depending on whether one is educated, say. It may not be possible to see the impact of the program unless one controls for that heterogeneity.
- You might also need data on variables that influence participation but do not influence outcomes given participation. Such instrumental variables can be valuable in sorting out the likely causal effects of non-random programs.
- The data on outcomes and other relevant explanatory variables can be either quantitative or qualitative. But it has to be possible to organize it in some sort of systematic data structure. A simple and common example is that one has values of various variables including one or more outcome indicators for various observation units (individuals, households, firms, communities).
- The variables one has data on and the observation units one uses are often chosen as part of the evaluation method. These choices should be anchored to the prior knowledge about the program (its objectives of course, but also how it is run) and the setting in which it is introduced.
- The specific source of the data on outcomes and their determinants, including program participation, typically comes from survey data of some sort. The observation unit could be the household, firm, geographic area, depending on the type of program one is studying.
- Survey data can often be supplemented with useful other data on the program (such as from the project monitoring data base) or setting (such as from geographic data bases).

Box 5.2: Propensity score matching

The aim of matching is to find the closest comparison group from a sample of non-participants to the sample of program participants. “Closest” is measured in terms of observable characteristics. If there are only one or two such characteristics then matching should be easy. But typically there are many potential characteristics. This is where propensity score matching comes in. The main steps in matching based on propensity scores are as follows:

Step 1: You need a representative sample survey of eligible non-participants as well as one for the participants. The larger the sample of eligible non-participants the better, to facilitate good matching. If the two samples come from different surveys, then they should be highly comparable surveys (same questionnaire, same interviewers or interviewer training, same survey period and so on).

Step 2: Pool the two samples and estimate a logit model of program participation as a function of all the variables in the data that are likely to determine participation.

Step 3: Create the predicted values of the probability of participation from the logit regression; these are called the “propensity scores”. You will have a propensity score for every sampled participant and non-participant.

Step 4: Some of the non-participant sample may have to be excluded at the outset because they have a propensity score which is outside the range (typically too low) found for the treatment sample. The range of propensity scores estimated for the treatment group should correspond closely to that for the retained sub-sample of non-participants. You may also want to restrict potential matches in other ways, depending on the setting. For example, you may want to only allow matches within the same geographic area to help assure that the matches come from the same economic environment.

Step 5: For each individual in the treatment sample, you now want to find the observation in the non-participant sample that has the closest propensity score, as measured by the absolute difference in scores. This is called the “nearest neighbor”. You will get more precise estimates if you use the nearest five neighbors (say).

Step 6: Calculate the mean value of the outcome indicator (or each of the indicators if there is more than one) for the five nearest neighbors. The difference between that mean and the actual value for the treated observation is the estimate of the gain due to the program for that observation.

Step 7: Calculate the mean of these individual gains to obtain the average overall gain. This can be stratified by some variable of interest such as incomes in the non-participant sample.

Box 5.3: Graphical representation of poverty impact

The empirical and counter-factual poverty incidence curves (as in Figure 5.1) are constructed as follows:

Step 1: You should already have the post-intervention income (or other welfare indicator) for each household in the whole sample (comprising both participants and non-participants); this is data. You also know how many people are in each household. And, of course, you know the total number of people in the sample (N; or this might be the estimated population size, if inverse sampling rates have been used to “expend up” each sample observation).

Step 2: You can plot this information in the form of a PIC. This gives (on the vertical axis) the percentage of the population living in households with an income less than or equal to that value on the horizontal axis. To make this graph, you can start with the poorest household, mark its income on the horizontal axis, and then count up on the vertical axis by 100 times the number of people in that household divided by N. The next point is the proportion living in the two poorest households, and so on. This gives the post-intervention PIC.

Step 3: Now calculate the distribution of pre-intervention income. To get this you subtract the estimated gain for each household from its post-intervention income. You then have a list of post-intervention incomes, one for each sampled household. Then repeat Step 2 to get the pre-intervention PIC.

How should these curves be interpreted? If we think of any given income level on the horizontal axis as a “poverty line” then the difference between the two PICs at that point gives the impact on the headcount index for that poverty line. Alternatively, looking horizontally gives you the income gain at that percentile. If none of the gains are negative then the post-intervention PIC must lie below the pre-intervention one. Poverty will have fallen no matter what poverty line is used. Indeed, this also holds for a very broad class of poverty measures; see Atkinson (1987). If some gains are negative, then the PICs will intersect. The poverty comparison is then ambiguous; the answer will depend on which poverty lines and which poverty measures one uses. You might then use *a priori* restrictions on the range of admissible poverty lines. For example, you may be confident that the poverty line does not exceed some maximum value, and if the intersection occurs above that value then the poverty comparison is unambiguous. If the intersection point (and there may be more than one) is below the maximum admissible poverty line then a robust poverty comparison is only possible for a restricted set of poverty measures. To check how restricted the set needs to be, you can calculate the poverty depth curves (PDCs). These are obtained by simply forming the cumulative sum up to each point on the PIC. (So the second point on the PDC is the first point on the PIC plus the second point, and so on.)

If the PDCs do not intersect then the direction of the program's impact on poverty is unambiguous as long as one restricts attention to the poverty gap index or any of the distribution sensitive poverty measures, such as the Watts (1968) measure or the squared poverty gap index of Foster, Greer and Thorbecke (1984). If the PDCs intersect then you can calculate the poverty severity curves with and without the program, by forming the cumulative sums under the PDCs. If these do not intersect over the range of admissible poverty lines then the direction of impact on any of the distribution-sensitive poverty measures.

For further discussion see Ravallion (1994).

Box 5.4: Double difference

The “double difference” method entails comparing a treatment group with a comparison group (as might ideally be determined by the score matching method described above) both before and after the intervention. The main steps are as follows:

Step 1: You need a “baseline” survey before the intervention is in place, and the survey must cover both non-participants and participants. If you do not know who will participate, you have to make an informed guess. Talk to the program administrators.

Step 2: You then need one or more follow-up surveys, after the program is put in place. These should be highly comparable to the baseline surveys (in terms of the questionnaire, the interviewing, etc). Ideally the follow-up surveys should be of the same sampled observations as the baseline survey. If this is not possible then they should be the same geographic clusters, or strata in terms of some other variable.

Step 3: Calculate the mean difference between the “after” and “before” values of the outcome indicator for each of the treatment and comparison groups.

Step 4: Calculate the difference between these two mean differences. That is your estimate of the impact of the program.

5.8 References

- Angrist, Joshua, Eric Bettinger, Erik Bloom, Elizabeth King and Michael Kremer, 2001, Vouchers for Private Schooling in Colombia: Evidence from a Randomized Natural Experiment, NBER Working Paper 8343.
- Angrist, Joshua, Guido Imbens and Donald Rubin, 1996, Identification of Causal Effects Using Instrumental Variables, *Journal of the American Statistical Association*, XCI, 444-455.
- Ashenfelter, Orley, 1978, Estimating the Effect of Training Programs on Earnings, *Review of Economic Studies* 60: 47-57.
- Atkinson, Anthony, 1987, On the Measurement of Poverty, *Econometrica*, 55: 749-64.
- Blundell, Richard and Monica Costa Dias, 2000, Evaluation Methods for Non-Experimental Data, *Fiscal Studies* 21(4): 427-468.
- Burtless, Gary, 1985, Are Targeted Wage Subsidies Harmful? Evidence from a Wage Voucher Experiment, *Industrial & Labor Relations Review*, Vol. 39, pp. 105-115.
- Chen, Shaohua and Martin Ravallion, 2003, Are the Income Gains from a Development Project Consumed or Saved?, Development Research Group, World Bank.
- Dehejia, R.H. and S. Wahba, 1999, Causal effects in non-experimental studies: Re-evaluating the evaluation of training programs, *Journal of the American Statistical Association* 94, 1053-1062.
- Dubin, Jeffrey A., and Douglas Rivers, 1993, Experimental Estimates of the Impact of Wage Subsidies, *Journal of Econometrics*, 56(1/2), 219-242.
- Foster, James, J. Greer, and Erik Thorbecke, 1984, A Class of Decomposable Poverty Measures, *Econometrica*, 52: 761-765.
- Galasso, Emanuela, Martin Ravallion and Agustin Salvia, 2001, Assisting the Transition from Workfare to Work: A Randomized Experiment, Policy Research Working Paper 2738, World Bank.
- Heckman, J., H. Ichimura, and P. Todd, 1997, Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme, *Review of Economic Studies* 64, 605-654.
- Heckman, J., H. Ichimura, J. Smith, and P. Todd, 1998, Characterizing selection bias using experimental data, *Econometrica* 66, 1017-1099.
- Jalan, Jyotsna and Martin Ravallion, 1998, Are There Dynamic Gains from a Poor-Area Development Program?, *Journal of Public Economics*, 67(1), 65-86.
- Jalan, Jyotsna and Martin Ravallion, 2002a, Does Piped Water Reduce Diarrhea for Children in Rural India? *Journal of Econometrics*, forthcoming.
- Jalan, Jyotsna and Martin Ravallion, 2002b, Estimating Benefit Incidence for an Anti-poverty Program using Propensity Score Matching, *Journal of Business and Economic Statistics*, forthcoming.
- Katz, Lawrence F., Jeffrey R. Kling and Jeffrey B. Liebman, 2001, Moving to Opportunity in Boston: Early Results of a Randomized Mobility Experiment, *Quarterly Journal of Economics*, 116(2), 607-654.
- Lalonde, Robert, 1986, Evaluating the Econometric Evaluations of Training Programs, *American Economic Review* 76: 604-620.
- Meyer, Bruce D., 1995, Natural and Quasi-Experiments in Economics, *Journal of Business and Economic Statistics*, April.
- Moffitt, Robert, 1991, Program Evaluation with Nonexperimental Data, *Evaluation Review*, 15(3): 291-314.
- Ravallion, Martin, 1994, *Poverty Comparisons*, Fundamentals in Pure and Applied Economics Volume 56, Harwood Academic Publishers.
- _____, 1999, Appraising Workfare, *World Bank Research Observer*, 14(1), 31-48.
- _____, 2001, The Mystery of the Vanishing Benefits: An Introduction to Impact Evaluation, *World Bank Economic Review* 15(1), 115-140.
- Ravallion, Martin, Emanuela Galasso, Teodoro Lazo and Ernesto Philipp, 2001, Do Workfare Participants Recover Quickly from Retrenchment? Policy Research Working Paper 2672, World Bank.
- Rosenbaum, Paul and Donald Rubin, 1983, The central role of the propensity score in observational studies for causal effects, *Biometrika*, 70, 41-55.
- _____ and _____, 1985, Constructing a Control Group using Multivariate Matched Sampling Methods that Incorporate the Propensity Score, *American Statistician* 39: 35-39.
- Rubin, Donald B., and N. Thomas, 2000, Combining propensity score matching with additional

- adjustments for prognostic covariates, *Journal of the American Statistical Association* 95, 573-585.
- Smith, Jeffrey and Petra Todd, 2001, Reconciling Conflicting Evidence on the Performance of Propensity-Score Matching Methods, *American Economic Review*, 91(2), 112-118.
- Van de Walle, Dominique, 2002, Choosing rural road investments to help reduce poverty, *World Development* 30(4).
- Watts, H.W., 1968, An Economic Definition of Poverty, in D.P. Moynihan (ed.), *On Understanding Poverty*. New York, Basic Books.
- Woodbury, Stephen and Robert Spiegelman, 1987, Bonuses to Workers and Employers to Reduce Unemployment," *American Economic Review*, 77, 513-530.