Chapter 3
Applying Analytical Methods for Impact Evaluation: A Case Study*

This case study is based on a hypothetical antipoverty program, PROSCOL, which provides cash transfers targeted to poor families with school-age children in one region of a given developing country. The case is intended to illustrate the analytical steps involved in carrying out an impact evaluation and the options an analyst may face, with the process applicable to any type of antipoverty program. In exploring how to go about evaluating the impact of the program, the policy analyst makes several common errors along the way, seeking input on specific topics from the specialized skills of colleagues—a statistician, an economist, an econometrics professor, and a sociologist.

Among the analytical steps that the analyst goes through in the case are identifying the questions to be addressed in the impact evaluation, assessing data resources, taking a first look at the data, understanding biases, learning about forgone income, adding control variables, understanding the importance of exogeneity, exploring better ways to form a comparison group (propensity score matching), learning about biases due to unobservables, reviewing what could have been done with a baseline survey (double differences), using instrumental variables, testing the various methodologies, incorporating input from the field, and planning for future work.

Description of the Hypothetical Program, PROSCOL

The PROSCOL program identifies families eligible for participation using various poverty proxies, which include the number of people in the household, the education of the head, and various attributes of the dwelling. PROSCOL pays a fixed amount per school-age child to all selected households on the condition that the children attend 85 percent of their school classes, which has to be verified by a note from the school. Households must keep their children in school until 18 years of age.

This program was introduced 12 months ago, is financed by the World Bank, and operates out of the Ministry of Social Development. In an effort

* This chapter draws heavily on a background paper by Martin Ravallion, *The Mystery of the Vanishing Benefits: Ms. Speedy Analyst's Introduction to Evaluation*, Policy Research Working Paper No. 2153, 1999.

to assess PROSCOL's impact on poverty in order to help determine whether the program should be expanded to include the rest of the country or be dropped, the World Bank has requested an impact evaluation by the Ministry of Finance. The request was to the Ministry of Finance so as to help assure an independent evaluation and to help develop capacity for this type of evaluation in a central unit of the government—close to where the budgetary allocations are being made.

Identifying the Questions to Be Addressed in the Impact Evaluation

The first step for the analyst in the Ministry of Finance assigned to the task of carrying out the PROSCOL evaluation is to clarify which project objectives will be looked at in evaluating impact. The project has two policy goals: the cash transfers aim to reduce current poverty, and by insisting that transfer recipients keep their kids in school the program aims to reduce future poverty by raising education levels among the current population of poor children. Two pieces of information would therefore be needed about the program to assess impact. First, are the cash transfers mainly going to low-income families? And second, how much is the program increasing school enrollment rates?

Assessing Data Resources

To carry out the evaluation the analyst has two main resources. The first is a report based on qualitative interviews with program administrators and focus groups of participants. It is not clear, however, whether those interviewed were representative of PROSCOL participants, or how poor they were relative to those who were not picked for the program and were not interviewed. The report says that the children went to school, but it is possible that they might have also gone to school if the program had not existed. Although this report is an important start, it does not tell the analyst how poor PROSCOL participants are and what impact the program has on schooling. The second resource is a recent independent national household survey carried out by the country's Bureau of Statistics, called the Living Standards Survey (LSS). The LSS included a random sample of 10,000 households and asked about household incomes by source, employment, expenditures, health status, education attainments, and demographic and other attributes of the family. The survey had incorporated a question on whether or not the sampled household had participated in PROSCOL and a line item for money received from PROSCOL in the listing of income sources.

Taking a First Look at the Data

The analyst then proceeds with obtaining the raw LSS data set to focus on assessing who is benefiting from the program. She uses a statistical software package such as SPSS or SAS to generate a cross-tabulation of the average amount received from PROSCOL by household deciles, where the deciles are formed by ranking all households in the sample according to their income per person. In calculating the latter, the analyst decides to subtract any monies received from PROSCOL as a good measure of income in the absence of the program with the intent of identifying who gained according to his or her preintervention income.

The cross-tabulation suggests that the cash transfers under the program are quite well-targeted to the poor. By the country's official poverty line, about 30 percent of the population in the Northwest is poor. From the table, calculations show that the poorest 30 percent of the survey sample receive 70 percent of the PROSCOL transfers. At first glance, this appears to be a positive result.

The next question is about the impact on schooling. This is looked at through a cross-tabulation of average school enrollment rates of various age groups for PROSCOL families versus non-PROSCOL families. This suggests almost no difference between the two; the average enrollment rate for kids aged 6 to 18 is about 80 percent in both cases. The analyst then calculates average years of schooling at each age, and the results are plotted separately for PROSCOL families and non-PROSCOL families. This shows that the two figures are not identical, but they are very close. At this stage, the analyst wonders whether there was really no impact on schooling, or whether the approach is wrong.

Understanding Biases

With this uncertainty the analyst next seeks input from a senior statistician to explore why the results suggest that PROSCOL children are no more likely to be in school than non-PROSCOLchildren. The statistician hypothesizes that the results may have a serious bias. In order to assess program impact, we need to know what would have happened without the program. Yet the analyst has not accounted for this; instead the non-PROSCOL families are used as the comparison group for inferring what the schooling of the PROSCOL participants would have been if the program had not existed.

In other words, P_i denotes PROSCOL participation of the *i*th child. This can take two possible values, namely $P_i = 1$ if the child participates in PROSCOL and $P_i = 0$ if he or she does not. If the *i*th child does not participate, then its level of schooling is S_{0i}, which stands for child *i*'s schooling S when $P = 0$. If the child does participate then its schooling is S_{1i}. Its gain in schooling due to PROSCOL is S_{1I}–S_{0i}. The gain for the *i*th child who participates (P =1) is then

$$G_i = S_{1i} - S_{0i} \mid P_i = 1.$$

The | stands for "given that" or "conditional on" and is needed to make it clear that the calculation is the gain for a child who actually participated. If one wants to know the average gain, this is simply the mean of all the *G*'s, which gives the sample mean gain in schooling among all those who participated in PROSCOL. As long as this mean is calculated correctly (using the appropriate sample weights from the survey), it will provide an unbiased estimate of the true mean gain. The latter is the "expected value" of *G*, and it can be written as

$$G = \mathrm{E}(S_{1i} - S_{0i} \mid P_i = 1).$$

This is another way of saying "mean." However, it need not be exactly equal to the mean calculated from the sample data, given that there will be some sampling error. In the evaluation literature, $\mathrm{E}(S_{1I} - S_{0i} \mid P_I = 1)$ is sometimes called the "treatment effect" or the "average treatment effect on the treated." In this case PROSCOL is considered the treatment.

The statistician points out to the analyst that she has not calculated *G*, but rather the difference in mean schooling between children in PROSCOL families and those in non-PROSCOL families. This is the sample estimate of

$$D = \mathrm{E}(S_{1i} \mid P = 1) - \mathrm{E}(S_{0i} \mid P = 0).$$

There is a simple identity linking the *D* and *G*, namely:

$$D = G + B.$$

This term "*B*" is the bias in the estimate, and it is given by

$$B = \mathrm{E}(S_{0i} \mid P_i = 1) - \mathrm{E}(S_{0i} \mid P_i = 0).$$

In other words, the bias is the expected difference in schooling without PROSCOL between children who did in fact participate in the program and those who did not. This bias could be corrected if $\mathrm{E}(S_{0i} \mid P_i = 1)$ were known, but it is not possible to even get a sample estimate of that. One cannot observe what the schooling would have been of children who actually participated in PROSCOL had they not participated; that is missing data—also called a "counterfactual" mean.

This bias presents a major concern. In the absence of the program, PROSCOL parents may well send their children to school less than do other parents. If so, then there will be a bias in the calculation. Going back to the original evaluation questions, we are interested in the extra schooling due to PROSCOL. Presumably this only affects those families who actually participate. In other words, we need to know how much less schooling could be expected without the program. If there is no bias, then the extra schooling under the program is the difference in mean schooling between those who participated and those who did not. Thus the bias arises if there is a difference in mean schooling between PROSCOL parents and non-PROSCOL parents in the absence of the program.

To eliminate this bias, the best approach would be to assign the program randomly. Then participants and nonparticipants will have the same expected schooling in the absence of the program, that is, $E(S_{0i} \mid P_i = 1) = E(S_{0i} \mid P_i = 0)$. The schooling of nonparticipating families will then correctly reveal the counterfactual, that is, the schooling that we would have observed for participants had they not had access to the program. Indeed, random assignment will equate the whole distribution, not just the means. There will still be a bias owing to sampling error, but for large enough samples one can safely assume that any statistically significant difference in the distribution of schooling between participants and nonparticipants is attributable to the program.

Within the existing design of the program, it is clear that participation is not random. Indeed, it would be a serious criticism of PROSCOL to find that it was. The very fact of its purposive targeting to poor families, which are presumably less likely to send their kids to school, would create bias.

This raises the question, if PROSCOL is working well then we should expect participants to have worse schooling in the absence of the program. Then $E(S_{0i} \mid P_i = 1) < E(S_{0i} \mid P_i = 0)$ and the analysts' original calculation will underestimate the gain from the program. We may find little or no benefit even though the program is actually working well.

The analyst now realizes that the magnitude of this bias could be huge. Suppose that poor families send their kids to work rather than school; because they are poor and cannot borrow easily, they need the extra cash now. Nonpoor families send their kids to school. The program selects poor families, who then send their kids to school. One observes negligible difference in mean schooling between PROSCOL families and non-PROSCOL families; indeed, $E(S_{1i} \mid P_i = 1) = E(S_{0i} \mid P_i = 0)$ in expectation. But the impact of the program is positive, and is given by $E(S_{0i} \mid P_i = 0) - E(S_{0i} \mid P_i = 1)$. The failure to take account of the program's purposive, pro-poor targeting could well have led to a substantial underestimation of PROSCOL's benefits from the analyst's comparison of mean schooling between PROSCOL families and non-PROSCOL families.

Learning about Forgone Income

The analyst next shows the results of her cross-tabulation of amounts received from PROSCOL against income to another colleague, an economist in the Ministry of Finance. The economist raises a main concern—that the gains to the poor from PROSCOL have been clearly overestimated because foregone income has been ignored. Children have to go to school if the family is to get the PROSCOL transfer; thus they will not be able to work, either in the family business or in the labor market. For example, children aged 15 to 18 can earn two-thirds or more of the adult wage in agriculture and construction. PROSCOL families will lose this income from their children's work. This foregone income should be taken into account when the net income gains from the program are calculated. And this net income gain should be subtracted, not the gross transfer, to work out preintervention income. This will also matter in determining how poor the family would have been in the absence of the PROSCOL transfer. The current table, therefore, might greatly overstate the program's gains to the poor.

The analyst wonders why she should factor out the forgone income from child labor, assuming that less child labor is a good thing. The economist highlights that she should look at the gains from reducing child labor, of which the main gain is the extra schooling, and hence higher future incomes, for currently poor families. The analyst has produced tables that reflect the two main ways PROSCOL reduces poverty: by increasing the current incomes of the poor and by increasing their future incomes. The impact on child labor matters to both, but in opposite directions; thus PROSCOL faces a tradeoff.

This highlights why it is important to get a good estimate of the impact on schooling; only then will it be possible to determine the forgone income. It is, for example, possible that the extra time at school comes out of nonwork time.

With regard to the second cross-tabulation, the main concern raised by the economist is that there is no allowance for all the other determinants of schooling, besides participation in PROSCOL. The economist suggests running a regression of years of schooling on a set of control variables as well as whether or not the child's family was covered by PROSCOL. For the *i*th child in the sample let

$$S_i = a + bP_i + cX_i + \varepsilon_i \ .$$

Here *a*, *b*, and *c* are parameters; *X* stands for the control variables, such as age of the child, mother's and father's education, the size and demographic composition of the household, and school characteristics; and ε is a residual that includes other determinants of schooling and measurement errors. The estimated value of *b* gives you the impact of PROSCOL on schooling.

Note that if the family of the *i*th child participates in PROSCOL, then $P = 1$ and so its schooling will be $a + b + cX_i + \varepsilon_i$. If it does not participate, then $P = 0$ and so its schooling will be $a + cX_i + \varepsilon_i$. The difference between the two is the gain in schooling due to the program, which is just *b*.

Adding Control Variables

As suggested, the analyst next runs a regression with and without the control variables. When it is run without them, the results show that the estimated value of *b* is not significantly different from zero (using the standard t-test given by the statistical package). These results look very similar to the first results, taking the difference in means between participants and nonparticipants—suggesting that PROSCOL is not having any impact on schooling. However, when several control variables are included in the regression, there is a positive and significant coefficient on PROSCOL participation. The calculation shows that by 18 years of age the program has added two years to schooling.

The analyst wonders why these control variables make such a difference? And are the right controls being used? She next visits her former econometrics professor and shows him her regressions. His first concern related to the regression of schooling on *P* and *X* is that it does not allow the impact of the program to vary with *X*; the impact is the same for everyone, which does not seem very likely. Parents with more schooling would be more likely to send their children to school, so the gains to them from PROSCOL will be lower. To allow the gains to vary with *X*, let mean schooling of nonparticipants be $a_0 + c_0X_i$ while that of participants is $a_1 + c_1X_i$, so the observed level of schooling is

$$S_i = (a_1 + c_1X_i + \varepsilon_{1i})P_i + (a_0 + c_0X_i + \varepsilon_{0i})(1 - P_i)$$

where ε_0 and ε_1 are random errors, each with means of zero and uncorrelated with *X*. To estimate this model, it is necessary to add an extra term for the interaction effects between program participation and observed characteristics to the regression already run. Thus the augmented regression is

$$S_i = a_0 + (a_1 - a_0)P_i + c_0X_i + (c_1 - c_0)P_iX_i + \varepsilon_i$$

where $\varepsilon_i = \varepsilon_{1i}P_i + \varepsilon_{0i}(1 - P_i)$. Then $(a_1 - a_0) + (c_1 - c_0)X$ is the mean program impact at any given value of X. If the mean X in the sample of participants is used, then it will give the mean gain from the program.

Understanding the Importance of Exogeneity

A second concern raised by the econometrics professor is in how the regression has been estimated. In using the regress command in the statistical package, ordinary least squares (OLS), there is concern because the OLS estimates of the parameters will be biased even in large samples unless the right-hand-side variables are exogenous. Exogeneity means that the right-hand-side variables are determined independently of schooling choices and so they are uncorrelated with the error term in the schooling regression. Because participation in the program was purposively targeted, PROSCOL's participation is not exogenous. This can affect the calculation of the program's impact as follows: The equation for years of schooling is

$$S_i = a + bP_i + cX_i + \varepsilon_i.$$

The value of $a + b + cX_i + \varepsilon_i$ was used as the estimate of the ith household's schooling when it participates in PROSCOL, while $a + cX_i + \varepsilon_i$ was used to estimate schooling if it does not participate. Thus the difference, b, is the gain from the program. However, in making this calculation the implicit assumption is that ε_i was the same either way. In other words, the assumption was that ε was independent of P, which would affect the calculation of the program's impact.

This highlights the bias due to nonrandom program placement, which may also be affecting the estimate based on the regression model suggested earlier by the economist ($S_i = a + bP_i + cX_i + \varepsilon_i$). This may not, however, mean that the results will be completely wrong.

The econometrics professor clarifies this with an explicit equation for P, namely,

$$P_i = d + eZ_i + \nu_i$$

where Z is several variables that include all the observed "poverty proxies" used for PROSCOL targeting. There will also be some purely random error term that influences participation; these are poverty proxies that are not in the data, and there will also have been mistakes in selecting participants that end up in this ν term. This equation is linear, yet P can only take two possible values, 0 and 1. Predicted values between zero and one are acceptable, but a linear model cannot rule out the possibility of negative predicted values, or values over one. There are nonlinear models that can deal with this problem, but to simplify the discussion it will be easiest to confine attention to linear models.

There is a special case in which the above OLS regression of S on P and X will give an unbiased estimate of b. That is when X includes all the variables in Z that also influence schooling, and the error term ν is uncorrelated with the error term ε in the regression for schooling. This is sometimes called "selection on observables" in the evaluation literature.

Suppose that the control variables X in the earlier regression for schooling include all the observed variables Z that influence participation P and ν is uncorrelated with ε (so that the unobserved variables affecting program placement do not influence schooling conditional on X). This has then eliminated any possibility of P being correlated with ε. It will now be exogenous in the regression for schooling. In other words, the key idea of selection on observables is that there is some observable X such that the bias vanishes conditional on X.

Adding the control variables to the regression of schooling on PROSCOL participation made a big difference because the X must include variables that were among the poverty proxies used for targeting, or were correlated with them, and they are variables that also influenced schooling. This, however, only works if the assumptions are valid. There are two problems to be aware of. First, the above method breaks down if there are no unobserved determinants of participation; in other words if the error term ν has zero variance, and all of the determinants of participation also affect schooling. Then there is no independent variation in program participation to allow one to identify its impact on schooling; it is possible to predict P perfectly from X, and so the regression will not estimate. This problem is unlikely to arise often, given that there are almost always unobserved determinants of program placement.

The second problem is more common, and more worrying in this case. The error term ε in the schooling regression probably contains variables that are not found in the survey but might well influence participation in the program, that is, they might be correlated with the error term í in the participation equation. If that is the case then $E(\varepsilon \mid X, P) \neq 0$, and ordinary regression methods will still be biased when regressions for schooling are estimated. Thus the key issue is the extent of the correlation between the error term in the equation for participation and that in the equation for schooling.

Exploring Better Ways to Form a Comparison Group—Propensity Score Matching

With further input from the professor, the analyst learns there are better ways to form a comparison group. The objective is to compare schooling levels conditional on observed characteristics. If the sample groups are divided into groups of families with the same or similar values of X, one compares the conditional means for PROSCOL and non-PROSCOL families. If schooling in the absence of the program is independent of participation, given X, then the comparison will give an unbiased estimate of PROSCOL's impact. This is sometimes called "conditional independence," and it is the key assumption made by all comparison-group methods.

Thus, a better way to select a comparison group, given the existing data, is to use as a control for each participant a nonparticipant with the same observed characteristics. This could, however, be very hard because the data set could have a lot of those variables. There may be nobody among the nonparticipants with exactly the same values of all the observed characteristics for any one of the PROSCOL participants.

A statistical approach, propensity score matching, provides techniques for simplifying the problem greatly. Instead of aiming to ensure that the matched control for each participant has exactly the same value of X, the same result can be achieved by matching on the predicted value of P, given X, which is called the propensity score of X. Rosenbaum and Rubin (1983) show that if (in this case) schooling without PROSCOL is independent of participation given X, then participants are also independent of participation given the propensity score of X. Since the propensity score is just one number, it is far easier to control for it than X, which could be many variables. And yet propensity score matching is sufficient to eliminate the bias provided there is conditional independence given X.

In other words, one first regresses P on X to get the predicted value of P for each possible value of X, which is then estimated for the whole sample. For each participant, one should find the nonparticipant with the closest value of this predicted probability. The difference in schooling is then the estimated gain from the program for that participant.

One can then take the mean of all those differences to estimate the impact. Or take the mean for different income groups. This, however, requires caution in how the model of participation is estimated. A linear model could give irregular predicted probabilities, above one, or negative. It is better to use the LOGIT command in the statistical package. This assumes that the error term ν in the participation equation has a logistic distribution, and estimates the parameters consistent with that assumption by maximum likelihood methods. This is based on the principles of the maximum likelihood estimation of binary response models.

Another issue to be aware of is that some of the nonparticipants may have to be excluded as potential matches right from the start. In fact there are some recent results in the literature in econometrics indicating that failure to compare participants and controls at common values of matching variables is a major source of bias in evaluations (see Heckman and others 1998).

The intuition is that one wants the comparison group to be as similar as possible to the treatment group in terms of the observables, as summarized by the propensity score. We might find that some of the nonparticipant sample has a lower propensity score than any of those in the treatment sample. This is sometimes called called "lack of common support." In forming the comparison group, one should eliminate those observations from the set of nonparticipants to ensure that only gains over the same range of propensity scores are being compared. One should also exclude those nonparticipants for whom the probability of participating is zero. It is advisable to trim a small proportion of the sample, say 2 percent, from the top and bottom of the nonparticipant distribution in terms of the propensity scores. Once the participants have been identified and nonparticipants have been identified over a common matching region, it is recommended to take an average of (say) the five or so nearest neighbors in terms of the absolute difference in propensity scores (box 3.1).

Box 3.1 Steps in Propensity Score Matching

The aim of matching is to find the closest comparison group from a sample of nonparticipants 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. The main steps in matching based on propensity scores are as follows:

Step 1: You need a representative sample survey of eligible nonparticipants as well as one for the participants. The larger the sample of eligible nonparticipants 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 nonparticipant.

Step 4: Some in the nonparticipant sample may have to be excluded at the outset because they have a propensity score that 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 subsample of nonparticipants. You may also want to restrict potential matches in other ways, depending on the setting. For example, you may want to allow only matches within the same geographic area to help ensure 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 nonparticipant sample that has the closest propensity score, as measured by the absolute difference in scores. This is called the "nearest neighbor." You can find the five (say) nearest neighbors.

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 income, in the nonparticipant sample.

This is the simplest form of propensity score matching. Complications can arise in practice. For example, if there is oversampling of participants, you can use choice-based sampling methods to correct for this (Manski and Lerman 1977); alternatively you can use the odds ratio ($p/(1-p)$, where p is the propensity score) for matching. Instead of relying on the nearest neighbor you can instead use all the nonparticipants as potential matches but weight them differently, according to how close they are (Heckman and others 1998).

Next, all the variables in the data set that are, or could proxy for, the poverty indicators that were used in selecting PROSCOL participants should be included. Again, X should include the variables in Z. This, however, brings out a weakness of propensity score matching. With matching, a different X will yield a different estimate of impact. With randomization, the ideal experiment, the results do not depend on what X you choose. Nor does randomization require that one specify a model for participation, whether a logit or something else. Box 3.1 summarizes the steps for doing propensity score matching.

Learning about Biases Due to Unobservables

Even after forming the comparison group, the analyst cannot be sure that this will give a much better estimate of the programs' impact. The methods described above will only eliminate the bias if there is conditional independence, such that the unobservable determinants of schooling—not included in the set of control variables *X*—are uncorrelated with program placement. There are two distinct sources of bias, that due to differences in observables and that due to differences in unobservables; the latter is often called "selection bias." Box 3.2 elaborates on this difference.

Going back to the professor's last equation shows that conditional independence will hold if *P* is exogenous, for then $E(\varepsilon_i \mid X_i, P_i) = 0$. However, endogenous program placement due to purposive targeting based on unobservables will still leave a bias. This is sometimes called selection on observables. Thus the conditions required for justifying the method raised earlier by the economist are no less restrictive than those needed to justify a version of the first method based on comparing PROSCOL families with non-PROSCOL families for households with similar values of *X*. Both rest on believing that these unobservables are not jointly influencing schooling and program participation, conditional on *X*.

Intuitively, one might think that careful matching reduces the bias, but that is not necessarily so. Matching eliminates part of the bias in the first naïve estimate of PROSCOL's impact. That leaves the bias due to any troublesome unobservables. However, these two sources of bias could be offsetting—one positive, the other negative. Heckman and others (1998) make this point. So the matching estimate could well have more bias than the naïve estimate. One cannot know on a priori grounds how much better off one is with even a well-chosen comparison group, which is an empirical question.

Reviewing What Could Have Been Done with a Baseline Survey—Double Difference Estimates

The analyst next inquires whether there would be another method besides randomization that is robust to these troublesome unobservables. This would require baseline data for both the participants and nonparticipants, collected before PROSCOL started. The idea is that data are collected on outcomes and their determinants both before and after the program is introduced, along with data for an untreated comparison group as well as the treatment group. It is then possible to just subtract the difference between the schooling of participants and the comparison group before the program is introduced from the difference after the program.

Box 3. 2 Sources of Bias in Naïve Estimates of PROSCOL's Impact

The bias described by the statistician is the expected difference in schooling without PROSCOL between families selected for the program and those not chosen. This can be broken down into two sources of bias:

- Bias due to differences in observable characteristics. This can come about in two ways. First, there may not be common support. The "support" is the set of values of the control variables for which outcomes and program participation are observed. If the support is different between the treatment sample and the comparison group then this will bias the results. In effect, one is not comparing like with like. Second, even with common support the distribution of observable characteristics may be different within the region of common support; in effect the comparison group data is misweighted. Careful selection of the comparison group can eliminate this source of bias.
- Bias due to differences in unobservables. The term selection bias is sometimes confined solely to this component (though some authors use that term for the total bias in a nonexperimental evaluation). This source of bias arises when, for given values of X, there is a systematic relationship between program participation and outcomes in the absence of the program. In other words, there are unobserved variables that jointly influence schooling and program participation conditional on the observed variables in the data.

There is nothing to guarantee that these two sources of bias will work in the same direction. So eliminating either one of them on its own does not mean that the total bias is reduced in absolute value. That is an empirical question. In one of the few studies to address this question, the true impact, as measured by a well-designed experiment, was compared with various nonexperimental estimates (Heckman and others 1998). The bias in the naïve estimate was huge, but careful matching of the comparison group based on observables greatly reduced the bias.

This is called the "double difference" estimate, or "difference in differences." This will deal with the troublesome unobserved variables provided they do not vary over time.

This can be explained by adding subscripts to the earlier equation so that the schooling after the program is introduced:

$$S_{ia} = a + bP_i + cX_{ia} + \varepsilon_{ia}.$$

Before the program, in the baseline survey, school attainment is instead

$$S_{ib} = a + cX_{ib} + \varepsilon_{ib}.$$

(Of course $P = 0$ before the program is introduced.) The error terms include an additive time invariant effect, so we can write them as

$$\varepsilon_{it} = \eta_i + \mu_{it} \text{ (for } t = a,b)$$

where η_i is the time invariant effect, which is allowed to be correlated with P_i, and μ_{it} is an innovation error, which is not correlated with P_i (or X_i).

The essential idea here is to use the baseline data to reveal those problematic unobservables. Notice that since the baseline survey is for the same households as we have now, the ith household in the equation for S_{ia} is the same household as the ith in the equation for S_{ib}. We can then take the difference between the "after" equation and the "before" equation:

$$S_{ia} - S_{ib} = bP_i + c(X_{ia} - X_{ib}) + \mu_{ia} - \mu_{ib}.$$

It is now possible to regress the change in schooling on program participation and the changes in X. OLS will give you an unbiased estimate of the program's impact. The unobservables—the ones correlated with program participation—have been eliminated.

Given this, if the program placement was based only on variables, both observed and unobserved, that were known at the time of the baseline survey, then it would be reasonable to assume that the η's do not change between the two surveys. This would hold as long as the problematic unobservables are time invariant. The changes in schooling over time for the comparison group will reveal what would have happened to the treatment group without the program.

This would require knowing the program well and being able to time the evaluation surveys so as to coordinate with the program. Otherwise there are bound to be unobserved changes after the baseline survey that influence who gets the program. This would create 0′s that changed between the two surveys.

This last equation can be interpreted as meaning that the child and household characteristics in X are irrelevant to the change in schooling if those characteristics do not change over time. But the gain in schooling may depend on parents′ education (and not just any change in their education), and possibly on where the household lives, because this will determine the access to schools. There can also be situations in which the changes over time in the outcome indicator are influenced by the initial conditions. Then one will also want to control for differences in initial conditions. This can be done by simply adding X_a and X_b in the regression separately so that the regression takes the form

$$S_{ia} - S_{ib} = bP_i + c_aX_{ia} + c_bX_{ib} + \mu_{ia} - \mu_{ib}.$$

Even if some (or all) variables in X do not vary over time one can still allow X to affect the changes over time in schooling.

The propensity score matching method discussed above can help ensure that the comparison group is similar to the treatment group before doing the double difference. In an interesting study of an American employment program, it was found that failure to ensure that comparisons were made in a region of common support was a major source of bias in the double-difference estimate in comparison with a randomized control group. Within the region of common support, however, the bias conditional on X did not vary much over time. Thus taking the double difference makes sense, after the matching is done (see Heckman and others (1998).

However, in practice, following up on households in surveys can be difficult. It may not be easy to find all those households that were originally included in the baseline survey. Some people in the baseline survey may not want to be interviewed again, or they may have moved to an unknown location.

If dropouts from the sample are purely random, then the follow-up survey will still be representative of the same population in the baseline survey. However, if there is some systematic tendency for people with certain characteristics to drop out of the sample, then there will be a problem. This is called "attrition bias." For example, PROSCOL might help some poor families move into better housing. And even when participant selection was solely based on information available at or about the baseline date (the time-invariant effect 0_i), selected participants may well drop out voluntarily on the basis of changes after that date. Such attrition from the treatment group will clearly bias a double-difference estimate of the program′s impact. Box 3.3 outlines the steps to form a double-difference estimate.

Box 3.3 Doing a Double Difference

The double-difference method entails comparing a treatment group with a comparison group (as might ideally be determined by the matching method in box 3.2) 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 nonparticipants 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, and so forth). 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.

This is the simplest version of double difference. You may also want to control for differences in exogenous initial conditions or changes in exogenous variables, possibly allowing for interaction effects with the program (so that the gain from the intervention is some function of observable variables). A suitable regression model can allow these variations.

Using Instrumental Variables

Given that there is no baseline survey of the same households to do the double-difference method, the professor recommends another methodology to get an estimate that is robust to the troublesome unobservables—an "instrumental variable."

An instrumental variable is the classic solution for the problem of an endogenous regressor. An instrumental variable is an observable source of exogenous variation in program participation. In other words, it is correlated with P but is not already in the regression for schooling and is not correlated with the error term in the schooling equation, ε. So one must have to have at least one variable in Z that is not in X and is not correlated with ε. Then the instrumental variables estimate of the program's impact is obtained by replacing P with its predicted value conditional on Z. Because this predicted value depends solely on Z (which is exogenous) and Z is uncorrelated with ε, it is now reasonable to apply OLS to this new regression.

Since the predicted values depend only on the exogenous variation due to the instrumental variable and the other exogenous variables, the unobservables are no longer troublesome because they will be uncorrelated with the error term in the schooling regression. This also suggests another, more efficient, way to deal with the problem. Remember that the source of bias in the earlier estimate of the program's impact was the correlation between the error term in the schooling equation and that in the participation equation. This is what creates the correlation between participation and the error term in the schooling equation. Thus a natural way to get rid of the problem when one has an instrumental variable is to add the residuals from the first-stage equation for participation to the equation for schooling but keeping actual participation in the schooling regression. However, since we have now added to the schooling regression the estimated value of the error term from the participation equation, it is possible to treat participation as exogenous and run OLS. This only works if there is a valid instrument. If not, the regression will not estimate because the participation residual will be perfectly predictable from actual participation and X, in a linear model.

An instrumental variable can also help if there is appreciable measurement error in the program participation data, another possible source of bias. Measurement error means that there is the possibility that program participation varies more than it actually does. This overestimation in the variance of P leads naturally to an underestimation of its coefficient b. This is called attenuation bias because this bias attenuates the estimated regression coefficient.

Although an instrumental variable can be extremely useful, in practice caution is necessary. When the actual participation is just replaced with its predicted value and OLS is run, this will not give the correct standard errors because the computer will not know that previously estimated parameters to obtain the predicted values had to be used. A correction to the OLS standard errors is required, though there are statistical packages that allow one to do this easily, at least for linear models.

If there was a dependent variable, however, that could only take two possible values, at school or not at school for instance, then one should use a nonlinear binary response model, such as logit or probit. The principle of testing for exogeneity of program participation is similar in this case. There is a paper by Rivers and Vuong (1988) that discusses the problem for such models; Blundell and Smith (1993) provide a useful overview of various nonlinear models in which there is an endogenous regressor.

Testing the Methodologies

When the analyst begins to think about identifying an instrumental variable she realizes that this is not a straightforward process. Every possibility she has come up with could also be put in with the variables in *X*. The problem is finding a valid "exclusion restriction" that justifies putting some variable in the equation for participation but not in the equation for schooling.

The analyst decides to try the propensity score matching method. The logit model of participation looks quite sensible and suggests that PROSCOL is well targeted. Virtually all of the variables that one would expect to be associated with poverty have positive, and significant, coefficients. The analyst then does the propensity score matching. In a comparison of the mean school enrollment rates, the results show that children of the matched-comparison group had an enrollment rate of 60 percent compared with 80 percent for PROSCOL families.

To account for the issue of forgone income, the analyst draws on an existing survey of child labor that asked about earnings. (In this developing country, there is an official ban on children working before they are 16 years of age, but the government has a hard time enforcing it; nonetheless, child wages are a sensitive issue.) From this survey, the earnings that a child would have had if he or she had not gone to school can be determined.

It is then possible to subtract from PROSCOL's cash payment to participants the amount of forgone income and thus work out the net income transfer. Subtracting this net transfer from total income, it is possible to work out where the PROSCOL participants come from in the distribution of preintervention income. They are not quite as poor as first thought (ignoring forgone income) but they are still poor; for example, two-thirds of them are below country's official poverty line.

Having calculated the net income gain to all participants, it is now possible to calculate the poverty rate with and without PROSCOL. The postintervention poverty rate (with the program) is, simply stated, the proportion of the population living in households with an income per person below the poverty line, where "income" is the observed income (including the gross transfer receipts from PROSCOL). This can be calculated directly from the household survey. By subtracting the net income gain (cash transfer from PROSCOL minus forgone income from children's work) attributed to PROSCOL from all the observed incomes, the results show a new distribution of preintervention incomes. The poverty rate without the program is then the proportion of people living in poor households, based on this new distribution. The analyst finds that the observed poverty rate in the Northwest of 32 percent would have been 36 percent if PROSCOL had not existed. The program allows 4 percent of the population to escape poverty now. The schooling gains mean that there will also be both pecuniary and nonpecuniary gains to the poor in the future. In the process of measuring poverty, the analyst remembers learning that the proportion of people below the poverty line is only a basic measure because it tells you nothing about changes below the line (see Box 3.4). She then calculates both the poverty gap index and the squared poverty gap index, and the results suggest that these have also fallen as a result of PROSCOL.

Box 3.4 Poverty Measures

The simplest and most common poverty measure is the headcount index. In this case, it is the proportion of the population living in households with income per person below the poverty line. (In other countries, it is a consumption-based measure, which has some advantages; for discussion and references see Ravallion 1994.)

The headcount index does not tell us anything about income distribution below the poverty line: a poor person may be worse off but the headcount index will not change, nor will it reflect gains among the poor unless they cross the poverty line.

A widely used alternative to the headcount index is the poverty gap (PG) index. The poverty gap for each household is the difference between the poverty line and the household's income; for those above the poverty line the gap is zero. When the poverty gap is normalized by the poverty line, and one calculates its mean over all households (whether poor or not), one obtains the poverty gap index.

The poverty gap index will tell you how much impact the program has had on the depth of poverty, but it will not reflect any changes in distribution among the poor caused by the program. For example, if the program entails a small gain to a poor person who is above the mean income of the poor, at the expense of an equal loss to someone below that mean, then PG will not change.

(Box continues on the following page.)

Box 3.4 *(continued)*

There are various "distribution-sensitive" measures that will reflect such changes in distribution among the poor. One such measure is the "squared poverty gap" (Foster, Greer, and Thorbecke 1984). This is calculated the same way as PG except that the individual poverty gaps as a proportion of the poverty line are squared before taking the mean (again over both poor and nonpoor). Another example of a distribution-sensitive poverty measure is the Watts index. This is the mean of the log of the ratio of the poverty line to income, where that ratio is set to one for the nonpoor. Atkinson (1987) describes other examples in the literature.

In this calculation, the analyst also recognizes that there is some uncertainty about the country's poverty line. To test the results, she repeats the calculation over a wide range of poverty lines, finding that at a poverty line for which 50 percent of the population are poor based on the observed postintervention incomes, the proportion would have been 52 percent without PROSCOL. At a poverty line that 15 percent fail to reach with the program, the proportion would have been 19 percent without it. By repeating these calculations over the whole range of incomes, the entire "poverty incidence curves" have been traced, with and without the program. This is also called the "cumulative distribution function" (see Box 3.5).

Box 3.5 Comparing Poverty with and without the Program

Using the methods described in the main text and earlier boxes, one obtains an estimate of the gain to each household. In the simplest evaluations this is just one number. But it is better to allow it to vary with household characteristics. One can then summarize this information in the form of poverty incidence curves (PICs), with and without the program.

Step 1: The postintervention income (or other welfare indicator) for each household in the whole sample (comprising both participants and nonparticipants) should already exist; 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 postintervention PIC.

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

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 head-count index for that poverty line (box 3.4). Alternatively, looking horizontally gives you the income gain at that percentile. If none of the gains are negative then the postintervention PIC must lie below the preintervention 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. (For further discussion see Ravallion 1994.) 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.)

(Box continues on the following page.)

Box 3.5 *(continued)*

If the PDCs do not intersect then 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 described in box 3.4. 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 impact on any of the distribution-sensitive poverty measures in box 3.4 is unambiguous.

Incorporating Input from the Field

In the implementation of every program, there is insight from beneficiaries and program administrators that may or may not be reflected in program data. For example, in this case the perception of those working in the field is that the majority of PROSCOL families are poor and that the program indeed provides assistance. When the analyst discusses this with a sociologist working with the program, she learns of some uncertainty in the reality of forgone income and the issue of work. The sociologist discusses that in the field one observes many children from poor families who work as well as go to school, and that some of the younger children not at school do not seem to be working. The analyst realizes that this requires some checking on whether there is any difference in the amount of child labor done by PROSCOL children versus that done by a matched-comparison group. This data, however, is not available in the household survey, though it would be possible to present the results with and without the deduction for forgone income.

The sociologist also has noticed that for a poor family to get on PROSCOL it matters a lot which school-board area (SBA) the family lives in. All SBAs get a PROSCOL allocation from the center, even SBAs that have very few poor families. If one is poor but living in a well-to-do SBA, they are more likely to get help from PROSCOL than if they live in a poor SBA. What really matters then, is relative poverty—relative to others in the area in which one lives—which matters much more than the absolute level of living.

This allocation would influence participation in PROSCOL, but one would not expect it to matter to school attendance, which would depend more on one's absolute level of living, family circumstances, and characteristics of the school. Thus the PROSCOL budget allocation across SBAs can be used as instrumental variables to remove the bias in the estimates of program impact.

There is information on which SBA each household belongs to in the household survey, the rules used by the center in allocating PROSCOL funds across SBAs, and how much the center has allocated to each SBA. Allocations are based on the number of school-age children, with an "adjustment factor" for how poor the SBA is thought to be. However, the rule is somewhat vague.

The analyst attempts to take these points into account, and reruns the regression for schooling, but replacing the actual PROSCOL participation by its predicted value (the propensity score) from the regression for participation, which now includes the budget allocation to the SBA. It helps to already have as many school characteristics as possible in the regression for attendance. Although school characteristics do not appear to matter officially to how PROSCOL resources are allocated, any omitted school characteristics that jointly influence PROSCOL allocations by SBA and individual schooling outcomes will leave a bias in the analyst's instrumental variable estimates. Although there is always the possibility of bias, with plenty of geographic control variables this method should at least offer a credible comparator to the matching estimate.

From the results it is determined that the budget allocation to the SBA indeed has a significant positive coefficient in the logit regression for PROSCOL participation. Now (predicted) PROSCOL participation is significant in a regression for school enrollment, in which all the same variables from the logit regression are included except the SBA budget allocation. The coefficient implies that the enrollment rate is 15 percentage points higher for PROSCOL participants than would have otherwise been the case. A regression is also run for years of schooling, for boys and girls separately. For either boys or girls of 18 years, the results indicate that they would have dropped out of school almost two years earlier if it had not been for PROSCOL. This regression, however, raises questions—whether the right standard errors are being used and whether linear models should be used.

Planning for Future Work

Finally, the analyst is ready to report the results of the evaluations. They show that PROSCOL is doing quite well, and as a result the policymakers show interest in expanding the program. From the process the analyst has gone through in carrying out the evaluation, she has a few important observations:

- Impact evaluation can be much more difficult than first anticipated;
- It is possible to come up with a worryingly wide range of estimates, depending on the specifics of the methodology used;
- It is good to use alternative methods in the frequent situations of less-than-ideal data, though each method has pitfalls; and
- One has to be eclectic about data.

In addition to the lessons the analyst has learned, she has a few key recommendations for future evaluation work of PROSCOL. First, it would be desirable to randomly exclude some eligible PROSCOL families in the rest of the country and then do a follow-up survey of both the actual participants and those randomly excluded from participating. This would give a more precise estimate of the benefits. It would, however, be politically sensitive to exclude some. Yet if the program does not have enough resources to cover the whole country in one go, and the program will have to make choices about who gets it first, it would indeed be preferable to make that choice randomly, among eligible participants. Alternatively, it would be possible to pick the schools or the school board areas randomly, in the first wave. This would surely make the choice of school or school board area a good instrumental variable for individual program placement.

Second, if this is not feasible, it is advisable to carry out a baseline survey of areas in which there are likely to be high concentrations of PROSCOL participants before the program starts in the South. This could be done at the same time as the next round of the national survey that was used for evaluating the PROSCOL program. It would also be good to add a few questions to the survey, such as whether the children do any paid work.

And third, it would be useful to include qualitative work, to help form hypotheses to be tested and assess the plausibility of key assumptions made in the quantitative analysis.

Note

1. See Heckman, Lalonde, and Smith (1999), and Abadie, Angrist, and Imbens (1998) for discussion on quartile treatment effects.