THE WORLD BANK ECONOMIC REVIEW, VOL. 15, NO. 1 115–140

The Mystery of the Vanishing Benefits: An Introduction to Impact Evaluation

Martin Ravallion

This article provides an introduction to the concepts and methods of impact evaluation. The author provides an intuitive explanation in the context of a concrete application. The article takes the form of a short story about a fictional character's on-the-job training in evaluation. Ms. Speedy Analyst is an economist in the Ministry of Finance in the fictional country of Labas. In the process of figuring out how to evaluate a human resource program targeted to the poor, Ms. Analyst learns the strengths and weaknesses of the main methods of ex post impact evaluation.

The setting for our story is the developing country of Labas. Twelve months ago, its government introduced an antipoverty program in northwest Labas with support from the World Bank. The program aims to provide cash transfers to poor families with school-age children. To be eligible to receive the transfer, households must have certain observable characteristics that suggest they are poor; to continue receiving the transfer they must keep their children in school until 18 years of age. The program is called PROSCOL.

The World Bank's country director for Labas has just asked the government to assess PROSCOL's impact on poverty to help determine whether the program should be expanded to include the rest of the country or be dropped. The Ministry of Social Development (MSD) runs PROSCOL. However, the Bank asked if the Ministry of Finance could do the evaluation, to help ensure independence and help develop capacity for this type of evaluation in a central unit of the government, close to where the budgetary allocations are made. The government agreed to the Bank's request. The Minister of Finance has delegated the task to Mr. Undersecretary, who has called in one of his brightest staff, Ms. Speedy Analyst.

Four years ago, Ms. Analyst graduated from the Labas National University with a master's degree in applied economics. She has worked in the Ministry of Finance since then. Ms. Analyst has a reputation for combining common sense

Martin Ravallion is with the Development Economics Research Group at the World Bank. For their comments and discussions, the author is grateful to Judy Baker, Kene Ezemenari, Emanuela Galasso, Paul Glewwe, Jyotsna Jalan, Emmanuel Jimenez, Aart Kraay, Robert Moffitt, Rinku Murgai, Pedro Olinto, Berk Ozler, Laura Rawlings, Dominique van de Walle, Michael Woolcock, and the journal's anonymous referees.

with an ability to get the most out of imperfect data. Speedy also knows that she is a bit rusty on the stuff she learned at the university.

Mr. Undersecretary gets straight to the point. "Speedy, the government is spending a lot of money on this PROSCOL program, and the minister wants to know whether the poor are benefiting from it, and how much. Could you please make an assessment?"

Ms. Analyst thinks this sounds a bit too vague for her liking. What does he mean by "benefiting?" Greater clarity on the program's objectives would be helpful.

"I will try to do my best, Mr. Undersecretary. What, may I ask, are the objectives of PROSCOL that we should judge it against?"

Mr. Undersecretary answers, "To reduce poverty in Labas, both now and in the future."

Ms. Analyst tries to pin this down some more. "I see. The cash transfers aim to reduce current poverty, and by insisting that transfer recipients keep their children in school, the program aims to reduce future poverty." Her boss nods as Ms. Analyst continues. "So I guess we need to know two things about the program. First, are the cash transfers mainly going to low-income families? Second, how much is the program increasing school enrollment rates?"

"That should do it, Speedy. Here is the file on the program that we got from the Ministry of Social Development."

Thus begins Ms. Analyst's on-the-job training in how to assess the impact of a social program. She makes notes along the way, which can be found in the appendix.

The Mystery Unfolds

Back in her office, Ms. Analyst finds that the file from the MSD includes some useful descriptive material on PROSCOL. She learns that targeting is done on the basis of various "poverty proxies," including the number of people in the household, the education of the household 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.

The file includes a report, "PROSCOL: Participants' Perspectives," commissioned by the MSD and done by a local consultant. The report was based on qualitative interviews with program administrators and focus groups of participants. Ms. Analyst cannot tell whether those interviewed are representative of PROSCOL participants or how poor they are relative to those who were not picked for the program and were not interviewed. The report says that the children went to school, but Ms. Analyst wonders whether they might not have also gone to school if there had been no program.

Ms. Analyst reflects to herself, "This report is a start, but it does not tell me how poor PROSCOL participants are and what impact the program has on school-

ing. I need hard data." Later she prepares a note on alternative data sources (see the appendix).

There is a promising lead in the MSD file. Nine months ago, the Labas Bureau of Statistics (LBS) carried out the first national household survey of Labas. It is called the Living Standards Survey (LSS). The survey was done for a random sample of 10,000 households, and it asked about household income by source, employment, expenditures, health status, educational attainment, and demographic and other attributes of the family. There is a letter in the file from the MSD to the LBS, dated a few months prior to the LSS, asking for a question to be added on whether the sampled household had participated in PROSCOL. The reply from LBS indicates that the listing of income sources in the survey schedule will include a line item for money received from PROSCOL.

"Wow," says Ms. Analyst, and she heads off to LBS.

Ms. Analyst already knows a few things about the LSS, having used tabulations from it produced by LBS. But she worries that she will not be able to do a good evaluation of PROSCOL without access to the raw household-level data. After a formal request from the minister (which Ms. Analyst wrote for him to sign), the secretary of the LBS agrees to give her the complete microdata from the LSS.

Ms. Analyst already knows how to use a statistics package called SAPS. After a long and frustrating day figuring out how to use the raw LSS data, Ms. Analyst starts the real work. She uses SAPS to make a cross-tabulation of the average amount received from PROSCOL by deciles of households, where the deciles are formed by ranking all households in the sample according to their income per person. In calculating the latter, Ms. Analyst decides to subtract any monies received from PROSCOL; this, she reckons, will be a good measure of income in the absence of the program. She hopes to reveal who gained according to their pre-intervention income.

The cross-tabulation suggests that the cash transfers under the program are quite well targeted to the poor. By the official LBS poverty line, about 30 percent of the population of northwest Labas are poor. From her table, she calculates that the poorest 30 percent of the LSS sample receive 70 percent of the PROSCOL transfers. This looks like good news for PROSCOL, Ms. Analyst reflects.

What about the impact on schooling? She makes another table, giving the average school enrollment rates of various age groups for PROSCOL families versus non-PROSCOL families. This suggests almost no difference between them; the average enrollment rate for children aged 6–18 years is about 80 percent in both cases.

Ms. Analyst then calculates average years of schooling at each age and plots the results separately for PROSCOL families and non-PROSCOL families. The two figures are not identical, but they are very close.

"Was there really no impact on schooling, or have I done something wrong?" she asks herself. The question is just the beginning of the story of how Ms. Analyst solves the mystery of the vanishing schooling benefits from PROSCOL.

Ms. Speedy Analyst Visits Mr. Unbiased Statistica

Ms. Analyst decides to show her curious results to a couple of trusted colleagues. First she visits Mr. Unbiased Statistica, a senior statistician at the LBS. Ms. Analyst likes Mr. Statistica and feels comfortable asking him about statistical problems.

"Mr. Statistica, my calculations from the LSS suggest that PROSCOL children are no more likely to be in school than non-PROSCOL children. Have I done something wrong?" she asks.

Mr. Statistica tells her bluntly, "Ms. Analyst, I think you may well have a serious bias here. To know what impact PROSCOL has, you need to know what would have happened without the program. The gain in schooling attributable to the program is just the difference between the actual school attendance rate for participating children and the rate for those same children if the program had not existed.

"What you are doing, Ms. Analyst, is using non-PROSCOL families as the comparison group for inferring what the schooling would be of the PROSCOL participants if the program had not existed. This assumes that the nonparticipants correctly reveal, at least on average, schooling without the program. Some simple algebra might help make this clear."

Mr. Statistica starts writing. "Let P_i denote the PROSCOL participation of the ith child. This can take two possible values, namely $P_i = 1$ when the child participates in PROSCOL and $P_i = 0$ when she does not. When the ith child does not participate, her level of schooling is S_{0i}, which stands for child i's schooling S when $P = 0$. When the child does participate, her schooling is S_{1i}. The expected gain in schooling due to PROSCOL for a child that does in fact participate is

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

"This is the conditional mean impact, conditional on participating in the program. In the evaluation literature, $E(S_{1i} - S_{0i} \mid P_i = 1)$ is sometimes called the treatment effect or the average treatment effect on the treated."

Ms. Analyst thinks to herself that the government would not like to call PROSCOL a "treatment." But she is elated by Mr. Statistica's equation. "Yes, Mr. Statistica, that is exactly what I want to know."

"Ah, but that is not what you have calculated, Ms. Analyst. You have 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

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

"There is a simple identity linking D and G, namely,

$$D = G + B.$$

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

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

"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. You could correct for this bias if you knew $E(S_{0i} \mid P_i = 1)$. But you can't even get a sample estimate of that. You can't observe what the schooling would have been of children who actually participated in PROSCOL had they not participated; that is missing data—it is a counterfactual mean."

Ms. Analyst sees that Mr. Statistica has a legitimate concern. In the absence of the program, PROSCOL parents may well send their children to school less than other parents do. If so, there will be a bias in her calculation. What the Finance minister needs to know is the extra schooling due to PROSCOL. Presumably this only affects those families who actually participate. So the minister needs 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. So the bias arises if there is a difference in mean schooling between PROSCOL and non-PROSCOL families in the absence of the program.

"What can be done to get rid of this bias, Mr. Statistica?"

"Well, in theory at least, the best way is 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 between participants and nonparticipants, not just the means. There will still be a bias due to sampling error, but for large enough samples you can safely assume that any statistically significant difference in the distribution of schooling between participants and nonparticipants is due to the program."

"Why did you say 'in theory at least?'"

"Well, no method is perfect. Randomization can be fraught with problems in practice. Political feasibility is often a problem. Even when selection is randomized, there can still be selective nonparticipation. There is a book by Manski and Garfinkel [1992] that I have somewhere that includes some interesting papers on the potential problems; I will lend you my copy. But the key point is to recognize the ways in which program placement is not in fact random."

On recalling what she read in the PROSCOL file, Ms. Analyst realizes that she need look no further than the design of the program to see 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, who are presumably less likely to send their children to school, would create bias.

Ms. Analyst tells Mr. Statistica about the program's purposive placement.

"So, Ms. Analyst, if PROSCOL is working well, then you 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 your calculation will underestimate the gain from the program. You may find little or no benefit even though the program is actually working well."

Ms. Analyst returns to her office, despondent. She sees now that the magnitude of the bias that Mr. Statistica is worried about could be huge. Her reasoning is as follows. Suppose that poor families send their children to work rather than school; because they are poor and cannot borrow easily, they need the extra cash now. Nonpoor families send their children to school. The program selects poor families, who then send their children to school. One observes negligible difference in mean schooling between PROSCOL families and non-PROSCOL families; indeed, $E(S_{1i} \mid P_i = 1) = E(S_{0i} \mid P_i = 0)$. But the impact of the program is positive and is given by $E(S_{1i} \mid P_i = 1) - E(S_{0i} \mid P_i = 1)$, which equals $E(S_{0i} \mid P_i = 0) - E(S_{0i} \mid P_i = 1)$ in this case. The failure to take account of the program's purposive, pro-poor targeting could well have led to a very substantial underestimation of PROSCOL's benefits in Ms. Analyst's comparison of mean schooling between PROSCOL families and non-PROSCOL families.

A Visit to Ms. Tangential Economiste

Next Ms. Analyst visits a colleague at the Ministry of Finance, Ms. Tangential Economiste. Ms. Economiste specializes in public finance. She has a reputation as a sharp economist, although sometimes she is a little brutal in her comments on her colleagues' work.

Ms. Analyst first shows her the cross-tabulation of amounts received from PROSCOL against income. Ms. Economiste immediately brings up a concern, which she chastises Ms. Analyst for ignoring. "You have clearly overestimated the gains to the poor from PROSCOL because you have ignored forgone income, Speedy. Children have to go to school if the family is to get the PROSCOL transfer. So they will not be able to work, either in the family business or in the labor market. Children aged 15 to 18 can earn two-thirds or more of the adult wage in agriculture and construction, for example. PROSCOL families will lose this income from their children's work. You should take account of this forgone income when you calculate the net income gains from the program. And you should subtract this net income gain, not the gross transfer, to work out preintervention income. Only then will you know how poor the family would have been in the absence of the PROSCOL transfer. I reckon this table might greatly *overstate* the program's gains to the poor."

"But why should I factor out the forgone income from child labor? Less child labor is surely a good thing," Ms. Analyst says in defense.

"You should certainly look at the gains from reducing child labor, of which the main gain is no doubt the extra schooling, and hence higher future incomes of currently poor families. I see your next table is about that. As I see it, you are concerned with the two main ways PROSCOL reduces poverty: One is by increas-

ing the current income of the poor, and the other is by increasing their future income. The impact on child labor matters to *both*, but in opposite directions. So PROSCOL faces a tradeoff."

Ms. Analyst realizes that this is another reason why she needs to get a good estimate of the impact on schooling; only then will she be able to determine the forgone income that Ms. Economiste is worried about. Maybe the extra time at school comes out of nonwork time.

Next, Ms. Analyst tells Ms. Economiste about Mr. Statistica's concerns about her second table.

"I think your main problem here is that you have not allowed for all the other determinants of schooling besides participation in PROSCOL. You should run a regression of years of schooling on a set of control variables as well as whether the child's family was covered by PROSCOL. Why not try this regression?" Ms. Economiste writes, "For the *i*th child in your 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. If the family of the ith child participates in PROSCOL ($P = 1$), then its schooling will be $a + b + cX_i + \varepsilon_i$. If it does not participate, then its schooling will be $a + cX_i + \varepsilon$. The difference between the two is the gain in schooling due to the program, which is just b."

This discussion puts Ms. Analyst in a more hopeful mood, as she returns to her office to try out Ms. Economiste's equation. Ms. Analyst runs the REGRESS command in SAPS on the regression with and without the control variables Ms. Economiste suggested. When she runs it without them, she finds that the estimated value of b is not significantly different from zero (using the standard t-test given by SAPS). This looks suspiciously like the result she first got, taking the difference in means between participants and nonparticipants—suggesting that PROSCOL is not having any impact on schooling. However, when she puts in the control variables suggested by Ms. Economiste, she immediately sees a positive and significant coefficient on PROSCOL participation. She calculates that by the time a participant reaches 18 years of age, the program has added 2 years to schooling.

Ms. Analyst thinks that this is starting to look more convincing. But she feels a little unsure about what she is doing. "Why do these control variables make such a difference? And have I used the right controls? I need more help if I am going to figure out what exactly is going on here, and whether I should believe this regression."

Professor Chisquare Helps Interpret Ms. Analyst's Results

Ms. Analyst decides to visit Professor Chisquare, who was one of her teachers at Labas National University. Professor Chisquare is a funny little man, who wears

old-fashioned suits and ties that don't match too often. "It is just not normal to be so square," Ms. Analyst recalls thinking during his classes in econometrics. She also recalls her dread at asking Professor Chisquare anything because his answers were sometimes very hard to understand. "But he knows more about regressions than anyone else I know," she reflects.

Ms. Analyst arranges a meeting. Having heard on the phone what her problem is, Professor Chisquare greets his former student with a long list of papers to read, mostly with impenetrable titles and published in obscure places. (His reading list is included in the references section herein.)

"Thanks very much, Professor, but I don't think I will have time to read all this before my report is due. Can I tell you my problem and get your reactions now?"

Professor Chisquare agrees. Ms. Analyst shows him Ms. Economiste's equations and the estimated regressions, thinking that he will be pleased that his former student has been running regressions. He asks her a few questions about what she has done and then rests back in his chair, ready, it seems, to pronounce judgment on her efforts so far.

"One concern I have with your 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."

"Yes, I wondered about that," chips in Ms. Analyst. "Parents with more schooling would be more likely to send their children to school, but this effect may well be more pronounced among the poor, so that the impact of PROSCOL will be higher among more educated families."

"Quite possibly, Ms. Analyst. 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 a mean of zero and uncorrelated with X. To estimate this model, all you have to do is add an extra term for the interaction effects between program participation and observed characteristics to the regression you have already run. So 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

$$(a_1 - a_0) + (c_1 - c_0)X_i + E(\varepsilon_{1i} - \varepsilon_{0i} \mid P_i, X_i)$$

is the mean program impact at any given value of X. Notice that you have something you don't know here, $E(\varepsilon_{1i} - \varepsilon_{0i} \mid P_i, X_i)$, which captures any differences in the unobserved variables that influence the schooling of participants, with and without the program. Under the assumption that the unobserved factors are the same on average among participants with and without the program (which is what Ms. Economiste implicitly assumed), you can get an estimate of mean im-

pact just by plugging the same mean X into this formula. If $c_1 = c_0$, then you get your previous specification in which $b=a_1 - a_0$ is the mean impact.

"A second concern, Ms. Analyst, is in how you have estimated your regression:

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

The REGRESS command in SAPS is just ordinary least squares (OLS). You should recall from when you did my econometrics class that OLS estimates of the parameters will be biased even in large samples unless the right-hand-side variables are exogenous. Here this means that the right-hand-side variables must be determined independently of schooling choices so they are uncorrelated with the error term ε in your regression. In other words, you require that $E(\varepsilon_i \mid P_i, X_i) = 0$. Is PROSCOL participation exogenous, Ms. Analyst?"

Ms. Analyst thinks quickly, recalling her conversation with Mr. Statistica. "No. Participation was purposively targeted. How does that affect my calculation of the program's impact?"

"Consider your original equation for years of schooling

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

You used $a + b + cX_i + \varepsilon_i$ as your estimate of the ith household's schooling when it participates in PROSCOL, while you used $a + cX_i + \varepsilon_i$ to estimate schooling if it does not participate. Thus the difference, b, is the gain from the program. However, in making this calculation you implicitly assumed that ε_i was the same either way. In other words, you assumed that ε was independent of P."

Ms. Analyst now sees that the bias due to nonrandom program placement that Mr. Unbiased Statistica was worried about might also be messing up her estimate based on the regression model suggested by Ms. Tangential Economiste. "Does that mean that my results are way off the mark?"

"Not necessarily," Professor Chisquare replies, as he goes to his white board. "Let's write down an explicit equation for P, as, say

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

where Z is a bunch of variables that includes all the observed 'poverty proxies' used for PROSCOL targeting. Of course there will also be some purely random error term that influences participation; these are poverty proxies that are not in your data, and there will also have been 'mistakes' in selecting participants that also end up in this ν term. Notice, too, that this equation is linear, yet P can only take two possible values, zero or one. Predicted values between zero and one are okay, but a linear model cannot rule out the possibility of negative predicted values or values greater than one. There are nonlinear models that can deal with this problem, but to simplify the discussion I will confine attention for now to linear models.

"Now, there is a special case in which your OLS regression of S on P and X will give you 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 your regression for schooling. This is sometimes called 'selection on observables' in the evaluation literature."

"Why does that eliminate the bias?" asks Ms. Analyst.

"Well, think about it. Suppose that the control variables X in your 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). Then you have eliminated any possibility of P being correlated with ε. It will now be exogenous in your regression for schooling.

"To put it another way, Ms. Analyst, the key idea of selection on observables is that there is some observable X such that the bias vanishes conditional on X."

"Why did it make such a difference when I added the control variables to my regression of schooling on PROSCOL participation?"

"Because your 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."

Ms. Analyst Learns about Better Methods of Forming a Comparison Group

Next, Ms. Analyst tells Professor Chisquare about her first attempt at estimating the benefits. "How might I form a better comparison group?"

"You want to compare schooling levels conditional on observed characteristics. Imagine that you find a sample of non-PROSCOL families with the same values of X as the PROSCOL families. If schooling is independent of participation, given X, then this comparison group 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."

Ms. Analyst tries to summarize. "So a better way to select my comparison group, given the data I have, is to use as a control for each participant a nonparticipant with the same observed characteristics. But that would surely be very hard, Professor, because I 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."

"Ah," says Professor Chisquare, "some clever statisticians have figured out how you can simplify the problem greatly. Instead of aiming to ensure that the matched control for each participant has exactly the same value of X, you can get the same result by matching on the probability of participating, given X. You should read the papers by Rosenbaum and Rubin (1983) on the list I prepared for you. Their paper shows that if (in your case) schooling without PROSCOL is independent of participation given X, then it is also independent of participation given the probability of participation given X. Because this probability is just one number, it is far easier to control for it than X, which could be many

variables, as you say. You can estimate the probability using the predicted value of *P* given *X* from a regression. This is called the 'propensity score.'"

"Let me see if I understand you, Professor. I first regress *P* on *X* to get the predicted value of *P* for each possible value of *X*, which I then estimate for my whole sample. For each participant, I then 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."

"That's right, Ms. Analyst. You can then take the mean of all those differences to estimate the impact. Or you can take the mean for different income groups, say. But you have to be careful with how you estimate the model of participation. A linear model for that purpose could give you crazy predicted probabilities, greater than one or negative. It is better to use the LOGIT command in SAPS. This assumes that the error term v in the participation equation has a logistic distribution and estimates the parameters consistent with that assumption by maximum likelihood methods. You remember my class on maximum likelihood estimation of binary response models, don't you?"

"Yes, I do," says Ms. Analyst, as convincingly as she can.

"Another issue you should be aware of, Ms. Analyst, is that some of the nonparticipants may be excluded as potential matches right from the start. There will, of course, be some nonparticipants who are ineligible according to the official eligibility rules, although they may nonetheless participate, so be very careful about choosing a comparison group that way. However, some families will have observable characteristics that make participation very unlikely. Indeed, you might find that some of the nonparticipant sample has a lower propensity score than any of those in the treatment sample. This is a case of what is sometimes called 'lack of common support.' There are recent results in the literature indicating that failure to compare participants and controls at common values of matching variables is a major source of bias in evaluations. See the paper by Heckman and others (1998) on my reading list. In forming your comparison group, you should eliminate those observations from the set of nonparticipants to assure that you are only comparing gains over the same range of propensity scores. You should certainly exclude those nonparticipants for whom the probability of participating is zero. It is probably also a good idea to trim a little, say, 2 percent, of the sample from the top and bottom of the nonparticipant distribution in terms of the propensity scores. Once you have identified participants and nonparticipants over a common matching region, I recommend you take an average of, say, the five or so nearest neighbors in terms of the absolute difference in propensity scores."

"What should I include in *X*?" Ms. Analyst asks.

"Well, clearly you should include all the variables in your data set that are or could proxy for the exogenous poverty indicators that were used by the MSD in selecting PROSCOL participants. So again *X* should include the variables in *Z*. Notice also that you don't want any *X*'s that were affected by the program. You

might focus solely on variables that were measured prior to joining the program or are unlikely to have changed. But that is not always clear; some characteristic might have changed in anticipation of becoming eligible for the program. Also note that, in principle, a different X will yield a different estimate of impact. You should check robustness."

Ms. Analyst prepares a note summarizing the steps she needs to follow in doing propensity score matching.

Troublesome and Not So Troublesome Unobservables

"I now have a much better idea of how to form the comparison group, Professor Chisquare. This should give me a much better estimate of the program's impact."

"Ah, there is no guarantee of that. All these methods I have described to you so far will only eliminate the bias if there is conditional independence, such that the unobservable determinants of schooling—not included in your set of control variables X—are uncorrelated with program placement after conditioning on X. There are two distinct sources of bias, that due to differences in observables, as we have discussed, and that due to differences in unobservables; the latter is often called 'selection bias.'" (See Ms. Analyst's notes in the appendix.)

Professor Chisquare points to his last equation. "Clearly 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 leave a bias. This is sometimes called 'selection on unobservables.'"

Ms. Analyst interjects. "So, really, the conditions required for justifying the method suggested by Ms. Economiste are no more restrictive than those needed to justify a version of my first method based on comparing PROSCOL families with non-PROSCOL families with similar values of X, or at least a similar propensity score of X. Both rest on believing that these unobservables are not jointly influencing schooling and program participation, conditional on X."

"That is not quite right, Ms. Analyst. The regression method of Ms. Economiste makes stronger assumptions about how schooling is determined, in the form of a parametric regression equation. Propensity score matching is a better method of dealing with the differences in observables. Notice, however, that matching does not necessarily reduce the bias. Matching eliminates part of the bias in your first naive estimate of PROSCOL's impact. That leaves the bias due to any troublesome unobservables. These two sources of bias could be offsetting, one positive and the other negative. Heckman and others (1998) make this point. So the matching estimate could well have more bias than the naive estimate. One of the few tests that have been done suggest that with good data propensity score matching can greatly reduce the overall bias and outperforms regression-based methods. See the paper by Dehejia and Wahba (1999), who assess matching against a randomized experiment on a U.S. training program. However, more tests of this sort are needed."

Ms. Analyst Regrets That a Baseline Survey Was Not Done

Ms. Analyst is starting to feel more than a little desperate. "Is there any method besides randomization that is robust to these troublesome unobservables?" she asks the professor.

"There is something you can do if you have 'baseline data' for both the participants and nonparticipants, collected before PROSCOL started. The idea is that you collect data on outcomes and their determinants both before and after the program is introduced, and you collect that data for an untreated comparison group as well as the treatment group. Then you can just subtract the difference between the schooling of participants and the comparison group before the program is introduced from the difference after the program. This is called the 'double difference' estimate, or just 'double diff' by people who like to abbreviate things. This will deal with the troublesome unobserved variables provided they do not vary over time."

Professor Chisquare turns to his white board again, pointing to one of his earlier equations. "To see how this works, let's add time subscripts, so schooling after the program is introduced is

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

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 deal with those troublesome unobservables. Notice that because the baseline survey is for the same households as you have now, the *i*th household in the equation for S_{iA} is the same household as the *i*th in the equation for S_{iB}. You can then take the difference between the 'after' equation and the 'before' equation; you get

$$S_{iA} - S_{iB} = bP_i + c(X_{iA} - X_{iB}) + \mu_{iA} - \mu_{iB}.$$

"So now you can 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 troublesome unobservables—the ones correlated with program participation—have been swept away."

Ms. Analyst reflects: "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."

Professor Chisquare nods. "Yes, as long as the troublesome 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."

Ms. Analyst thinks to herself that this means one needs to know the program well and be 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 η's that change between the two surveys.

Something about Professor Chisquare's last equation is worrying her. "As I understand it, professor, this last equation means 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, as this will determine the access to schools."

"Yes, Ms. Analyst, there can 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. You can do this simply by adding X_A and X_B in the regression separately, so that the regression takes the form

$$S_{iA} - S_{iB} = bP_i + c_a X_{iA} + c_b X_{iB} + \mu_{iA} - \mu_{iB}.$$

So 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 that I told you about can help ensure that the comparison group is similar to the treatment group before you do 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 when compared with a randomized control group. Within the region of common support, however, the bias conditional on X did not vary much over time. So taking the double difference makes sense, after the matching is done. See the paper by Heckman and others (1998) on my reading list."

Ms. Analyst has had some experience doing surveys and is worried about this idea of following up households. "When doing the follow-up survey, it must 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. Is that a problem?"

"If the dropouts 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 around the baseline date (the time-invariant effect η_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."

Later Ms. Analyst writes up some notes about forming a double difference estimate (see the appendix).

Professor Chisquare Reminds Ms. Analyst about Instrumental Variables

"Double difference is neat, Professor Chisquare. But I don't have a baseline survey of the same households. I don't think anyone thought PROSCOL would have to be evaluated when they started the program. Is there anything else I can do to get an estimate that is robust to the troublesome unobservables?"

"What you then need is an instrumental variable (IV)" he tells her. "You surely recall from my classes that this is the classic solution for the problem of an endogenous regressor."

"Can you just remind me, Professor Chisquare?"

"An instrumental variable is really just some 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 ε in the schooling equation. So you must 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 by 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 ordinary least squares to this new regression."

"I see," says Ms. Analyst. "Because 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, as they will be uncorrelated with the error term in the schooling regression."

"You've got it, Ms. Analyst. That also suggests another way you can deal with the problem. Remember that the source of bias in your 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. So a natural way to get rid of the problem when you have an instrumental variable is to add the residuals from the first-stage equation for participation to the equation for schooling. You still leave actual participation in the schooling regression. But since you have now added to the schooling regression the estimated value of the error term from the participation equation, you can treat participation as exogenous and run OLS. Of course, this only works if you have a valid instrument. If you don't, the regression will not estimate, because the participation residual will be perfectly predictable from actual participation and X in a linear model. Unobserved heterogeneity across people in the program's impact can also create a problem

for the IV estimator of the average gains among participants; the paper by Heckman (1997) talks about this.

"An IV can also help if you think there is appreciable measurement error in your program participation data. This is another possible source of bias. Measurement error means that you think that program participation varies more than it actually does. If the measurement error is random—here meaning that it has zero mean and is not correlated with observed or unobserved determinants of schooling—then your estimate of the program's impact will be biased toward zero, and the larger the variance of the measurement error, the greater the bias."

"Yes, you called that 'attenuation bias' in your class, as I recall, because this bias attenuates the estimated regression coefficient, pushing it toward zero."

"However, if you had a dependent variable that could only take two possible values, at school or not at school, say, then you 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. I have written a program that can do a probit with an endogenous regressor, and I can give you a copy."

"Thanks. I guess I will cross that bridge when I get to it. But what should I use as an instrument?" asks Ms. Analyst.

"Ah, that you will have to figure out yourself, Ms. Analyst."

Ms. Analyst Returns to Her Computer

Ms. Analyst is starting to wonder whether this will ever end. "I'm learning a lot, but what am I going to tell my boss?"

She tries to think of an instrumental variable. But every possibility she can think of could just as well be put in with the variables in *X*. She now remembers Professor Chisquare's class; her problem is finding a valid "exclusion restriction" that justifies putting some variable in the equation for participation but not in the equation for schooling.

Ms. Analyst decides to try the "propensity score matching method" suggested by Professor Chisquare. Her logit model of participation looks quite sensible and suggests that PROSCOL is well targeted. (Virtually all of the variables that she would expect to be associated with poverty have positive and significant coefficients.) This is interesting in its own right. She then does the propensity score matching just as Professor Chisquare has advised her. On comparing the mean school enrollment rates, Ms. Analyst finds that children of the matched comparison group had an enrollment rate of 60 percent, compared with the figure of 80 percent for PROSCOL families.

She now thinks back on the comments that Ms. Economiste made about forgone income. Ms. Analyst finds that the LBS did a special survey of child labor that asked about earnings. (There is an official ban on children working before

they are 16 years of age in Labas, but the government has a hard time enforcing it; nonetheless, child wages are a sensitive issue.) From this she can figure out the earnings a child would have had if he or she had not gone to school.

Ms. Analyst can now subtract from PROSCOL's cash payment to participants the amount of forgone income, and so work out the net income transfer. Subtracting this net transfer from total income, she can work out where the PROSCOL participants come from in the distribution of pre-intervention income. They are not quite as poor as she had first thought (ignoring forgone income), but they are still poor; for example, two-thirds of them are below the official poverty line of Labas.

Having calculated the net income gain to all participants, Ms. Analyst can now calculate the poverty rate with and without PROSCOL. The "post-intervention" poverty rate (with the program) is just 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 she calculates directly from the LSS. By subtracting the net income gain (cash transfer from PROSCOL minus forgone income from children's work) attributed to PROSCOL from all the observed income, she gets a new distribution of pre-intervention income. The poverty rate without the program is then the proportion of people living in poor households, based on this new distribution. Ms. Analyst finds that the observed poverty rate in northwest Labas of 32 percent would have been 36 percent if PROSCOL had not existed. The program allows 4 percent of the population to escape poverty. The schooling gains mean that there will also be both pecuniary and nonpecuniary gains to the poor in the future.

Ms. Analyst also recognizes that there is some uncertainty about the LBS poverty line. So she repeats this calculation over a wide range of poverty lines. She finds that at a poverty line for which 50 percent of the population are poor, based on the observed post-intervention incomes, the proportion would have been 52 percent without PROSCOL. At a poverty line that 15 percent fail to reach with the program would have been 19 percent without it. By repeating these calculations over the range of incomes, Ms. Analyst realizes that she has traced out the entire "poverty incidence curves" with and without the program, which are what statisticians call the "cumulative distribution function."

Ms. Analyst makes an appointment with Mr. Undersecretary to present her assessment of PROSCOL.

A Chance Encounter with Ms. Sensible Sociologist

The day before she is due to present her results to her boss, Ms. Analyst accidentally bumps into her old friend Ms. Sensible Sociologist, who now works for one of Labas's largest nongovernmental organizations, the Social Capital for Empowerment Foundation (SCEF). Ms. Analyst tells Ms. Sociologist all the details about what she has been doing on PROSCOL.

Ms. Sociologist's eyes start to roll when Ms. Analyst talks about "unbiased estimates" and "propensity scores." "I am no expert on that stuff, Speedy. But I do know a few things about PROSCOL. I have visited some of the schools in northwest Labas where there are a lot of PROSCOL children, and I meet PROSCOL families all the time in my work for SCEF. I can tell you they are not all poor, but most are. PROSCOL helps.

"However, this story about 'forgone income' that Tangential came up with, I am not so sure about that. Economists have strange ideas sometimes. I have seen plenty of children from poor families who work as well as go to school. Some of the younger ones who are not at school don't seem to be working. Maybe Tangential is right in theory, but I don't know how important it is in reality."

"You may be right, Sense. What I need to do is check whether there is any difference in the amount of child labor done by PROSCOL children versus a matched comparison group," says Ms. Analyst. "The trouble is that the LSS did not ask about child labor. That is in another LBS survey. I think what I will do is present the results with and without the deduction for forgone income."

"That might be wise," says Ms. Sociologist. "Another thing I have noticed, Speedy, is 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 you are poor but living in a well-to-do SBA, you are more likely to get help from PROSCOL than if you live in a poor SBA. The authorities like to let all areas participate for political reasons. As a result, it is relative poverty—relative to others in the area where you live—that matters much more than your absolute level of living."

"No, I did not know that," replies Ms. Analyst, a little embarrassed that she had not thought of talking to Ms. Sociologist earlier, since this could be important.

"That gives me an idea, Sense. I know which school board area each household belongs to in the LBS survey, and I know how much the center has allocated to each SBA. Given what you have told me, that allocation would influence participation in PROSCOL, but one would not expect it to matter for school attendance, which would depend more on one's absolute level of living, family circumstances, and I guess characteristics of the school. So the PROSCOL budget allocation across SBAs can be used as instrumental variables to remove the bias in my estimates of program impact. "

Ms. Sociologist's eyes roll again, as Ms. Analyst says farewell and races back to her office. She first looks into the original file she was given, to see what rules are used by the center in allocating PROSCOL funds across SBAs. A memo from the ministry indicates that 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.

Ms. Analyst reruns her regression for schooling. But now she replaces 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.

However, she is worried about the possibility that the allocation of the PROSCOL budget might be correlated with omitted determinants of schooling in her model, notably characteristics of the school. Although school characteristics do not appear to matter officially to how PROSCOL resources are allocated, she worries that unofficially they may matter.

"Any omitted school characteristics that jointly influence PROSCOL allocations by SBA and individual schooling outcomes will leave a bias in my IV estimates," Ms. Analyst says to herself. She realizes that her only protection against that problem with the data she has is to add as many school characteristics as possible to her regression for attendance. But she also realizes that she can never rule out the possibility of bias. She is still making a conditional independence assumption. An influential school principal, for example, might simultaneously attract PROSCOL and achieve better outcomes for the students in other ways. She cannot measure the political clout of the principal. But with plenty of geographic control variables, Ms. Analyst thinks that this method should at least offer a credible alternative for comparison with her matching estimate.

Soon she has the results. Consistent with Ms. Sociologist's observations, the budget allocation to the SBA 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 she includes all the same variables from the logit regression, 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. She also runs regressions for years of schooling and for boys and girls separately. For either boys or girls 18 years of age, her results indicate that they would have dropped out of school almost 2 years earlier if it had not been for PROSCOL.

Ms. Analyst wonders what Professor Chisquare will think of this. She is sure he will find something questionable about her methods. "I wonder if I am using the right standard errors? And should I be using linear models?" Ms. Analyst decides she will order a new software program, FEM (Fancy Econometric Methods), that she has heard about. But that will have to wait. For now, Ms. Analyst is happy that her results are not very different from those she got using the propensity score matching method. She is reassured somewhat by Ms. Sociologist's comments based on her observations in the field. Ms. Analyst figures, "They can't all be wrong."

Ms. Analyst Reports Back to Her Boss

Ms. Analyst writes up her results and gives the report to Mr. Undersecretary. He seems quite satisfied. "So PROSCOL is doing quite well." Mr. Undersecretary arranges a meeting with the minister, and he asks Ms. Analyst to attend. The minister is interested in Ms. Analyst's results and asks some questions about how she figured out the benefits from PROSCOL. He seems to appreciate her efforts to ensure that the comparison group is similar to the PROSCOL families.

"I think we should expand PROSCOL to include the rest of Labas," the minister concludes. "We will not be able to do it all in one go, but over about two years I think we could cover the whole country. I want you to keep monitoring the program, Ms. Analyst."

"I would like to do that. However, I have learned a few things about these evaluations. I would recommend that you randomly exclude some eligible PROSCOL families in the rest of Labas. We could then do a follow-up survey of both the actual participants and those randomly excluded from participating. That would give us a more precise estimate of the benefits, and pointers for improving the program."

The minister gets a dark look in his eyes, and Mr. Undersecretary starts shifting in his seat uncomfortably. The minister then bursts out laughing. "You must be joking, Ms. Analyst! I can just see the headlines in the *Labas Herald*: 'Government Randomly Denies PROSCOL to Families in Desperate Need.' Do you not want me to get reelected?"

"I see your point, minister. But because you do not have enough money to cover the whole country in one go, you are going to have to make choices about who gets it first. Why not make that choice randomly among eligible participants? What could be fairer?"

The minister thinks it over. "What if we picked the schools or the SBAs randomly in the first wave?"

Ms. Analyst thinks. "Yes, that would surely make the choice of school or SBA a good instrumental variable for individual program placement," she says with evident enthusiasm.

"Instrumental what?" asks the minister, and Mr. Undersecretary shifts in his seat again. "Never mind. If that works for you, then I will try to see if I can do it that way. The Ministry of Social Development will have to agree of course."

"If that does not work, Mr. Minister, could we do something else instead, namely, a baseline survey of areas in which there are likely to be high concentrations of PROSCOL participants before the program starts in the South? I would like to do this at the same time as the next round of the national survey I used for evaluating PROSCOL in northwest Labas. There are also a few questions I would like to add to the survey, such as whether the children do any paid work."

"Yes, that sounds like a reasonable request, Ms. Analyst. I will also talk to the secretary of Statistics."

EPILOGUE

It is three years later. Ms. Analyst is head of the new Social and Economic Evaluation Unit, which reports directly to the minister of Finance. The unit is currently evaluating all of Labas's social programs on a regular basis. Ms. Analyst has a permanent staff of three assistants. She regularly hires both Professor Chisquare and Ms. Sociologist as consultants. They have a hard time talking to each other. ("Boy, that Chisquare is just not normal," Ms. Sociologist confided

to Ms. Analyst one day.) But Ms. Analyst finds it useful to have both of them around. The qualitative field trips and interviews with stakeholders that Ms. Sociologist favors help a lot in forming hypotheses to be tested and in assessing the plausibility of key assumptions made in the quantitative analysis that Professor Chisquare favors. Ms. Analyst reflects that the real problem with MSD's "Participants' Perspectives" report on PROSCOL was not what it did but what it did not do; casual interviews can help in understanding how a program works on the ground, but on their own they cannot deliver a credible assessment of impact.

However, Ms. Analyst has also learned that rigorous impact evaluation is much more difficult than she first thought, and one can sometimes obtain a worryingly wide range of estimates, depending on the specifics of the methodology used. Professor Chisquare's advice remains valuable in suggesting alternative methods in the frequent situations of less than ideal data and in pointing out the pitfalls. Ms. Analyst has also learned to take an eclectic approach to data.

The Finance minister eventually convinced the minister of Social Development to randomize the first tranche allocation of PROSCOL II across SBAS in the rest of Labas, and this helped Ms. Analyst identify the program's impact. Her analysis of the new question on child labor added to the LBS survey revealed that there was some forgone income from PROSCOL, although not quite as much as she had first thought.

Ms. Economiste made a further comment on Ms. Analyst's first report on PROSCOL to the effect that Ms. Analyst could also measure the future income gains from PROSCOL, using recent work by labor economists on the returns to schooling in Labas. When she factored this into her calculations, PROSCOL was found to have quite a reasonable economic rate of return, on top of the fact that the benefits were reaching the poor. She found that this calculation was somewhat sensitive to the discount rate used for working out the present value of the future income gains.

One big difference compared with her first PROSCOL evaluation is that Ms. Analyst now spends a lot more time understanding how each program works before doing any number crunching. She spreads the evaluation over a much longer period, often including baseline and multiple follow-up surveys of the same households.

However, not everything has gone smoothly. At first she had a lot of trouble getting the relevant ministries to cooperate with her. It is often hard to get them to define the objectives of each program she is evaluating. Ms. Analyst sometimes thinks that getting the relevant ministry to define the objectives of its public spending is an important contribution in its own right. Eventually the ministries realized that they can learn a lot from these evaluations and that they were being taken seriously by the Finance minister.

Internal politics within the government is often a problem. Thankfully, the data side is now working well. The minister had the good idea of making the secretary of Statistics an adviser to the unit, and Mr. Statistica is his representa-

tive. Ms. Analyst often commissions new surveys from LBS and advises them on questionnaire design and sampling.

Ms. Analyst has also started giving advice to other countries and international agencies (including the World Bank) embarking on impact evaluations of social programs. She has also found that swapping notes with other program analysts can be valuable. Ms. Analyst reckons there are policy mysteries galore in Labas and elsewhere—and that the tools she has learned to use might well shed light on them.

Appendix. Ms. Speedy Analyst's Notebook

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, and so on.
- 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 counterfactual 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, say, whether one is educated. 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. These instrumental variables can be valuable in sorting out the likely causal effects of nonrandom 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, and communities).
- The variables on which one has data 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 data on outcomes and their determinants, including program participation, typically come from survey data of some sort. The observation unit could be the household, firm, or 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 database) or setting (such as from geographic databases).

Sources of Bias

A naive estimate of a program's impact is to compare the relevant outcome indicators between participants and nonparticipants. This estimate will be biased if there is a difference between these two groups in outcomes without the intervention. 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, 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 are misweighted. Careful selection of the comparison group can eliminate this source of bias by choosing a comparison group with the same distribution of observed characteristics as the treatment group.
- *Bias due to differences in unobservables*. The term "selection bias" is sometimes confined solely to this component (although 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 no 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.

Methods for Evaluating Impact

Various methods exist to reduce the bias in a naive estimate. The essential problem these methods address is that we do not observe the outcomes for participants if they had not participated. So evaluation is essentially a problem of missing data.

A *comparison group* is used to identify the counterfactual of what would have happened without the program. The comparison group is designed to be very similar to the *treatment group* of participants with one key difference: the comparison group did not participate. The main methods available are as follows:

- *Randomization*. The selection into the treatment and comparison groups is random in some well-defined set of people. Then there will be no difference on average between the two groups besides the fact that the treatment group got the program.
- *Matching*. Here one tries to pick an ideal comparison group from a larger survey. The comparison group is matched to the treatment group on the basis of a set of observed characteristics, or using the predicted probability of participation given observed characteristics ("propensity score"). A good comparison group comes from the same economic environment as the treatment group and was administered the same questionnaire by similarly trained interviewers.
- *Double difference (or difference in difference) methods*. Here one compares a treatment and comparison group (first difference) before and after a program (second difference). Comparators should be dropped if they have propensity scores outside the range observed for the treatment group. (A special case is a "reflexive comparison" that only compares the treatment group before and after the intervention; because there is no control group, this method can be deceptive as a basis for assessing impact.)
- *Instrumental variables methods*. Instrumental variables are variables that matter to participation but not to outcomes given participation. If such variables exist then they identify a source of exogenous variation in outcomes attributable to the program—recognizing that its placement is not random but purposive. The instrumental variables are first used to predict program participation, then one sees how the outcome indicator varies with the predicted values, conditional on other characteristics.

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. 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 nonparticipants as well as one for the participants. The larger the sample of eligible nonparticipants, the better it will 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 the propensity scores. You will have a propensity score for every sampled participant and nonparticipant.

Step 4. Some of 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 only allow 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 will get more precise estimates if you use, say, the nearest five neighbors. Or you can instead use all the nonparticipants as potential matches, but weight them differently according to how close they are (Heckman and others 1998).

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 nonparticipant sample.

Doing a Double Difference

The double difference method entails comparing a treatment group with a comparison group (as might ideally be determined by the propensity 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 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 survey (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.

Suggested Reading

Blundell, Richard W., and R. J. Smith. 1993. "Simultaneous Microeconometric Models with Censoring or Qualitative Dependent Variables." In G. S. Maddala, C. R. Rao, and H. D. Vinod, eds., *Handbook of Statistics Volume 11*. Amsterdam: North Holland.

Dehejia, Rajeev H., and Sadek Wahba. 1999. "Causal Effects in Non-Experimental Studies: Re-Evaluating the Evaluation of Training Programs." *Journal of the American Statistical Association* 94:1053–62.

Grossman, Jean Baldwin. 1994. "Evaluating Social Policies: Principles and U.S. Experience." *World Bank Research Observer* 9(2):159–80.

Heckman, James. 1997. "Instrumental Variables. A Study of Implicit Behavioral Assumptions Used in Making Program Evaluations." *Journal of Human Resources* 32(3):441–61.

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

Heckman, J., H. Ichimura, J. Smith, and P. Todd. 1998. "Characterizing Selection Bias Using Experimental Data." *Econometrica* 66:1017–99.

Meyer, Bruce D. 1995. "Natural and Quasi-Experiments in Economics." *Journal of Business and Economic Statistics* 13:151–62.

Manski, Charles, and Irwin Garfinkel, eds. 1992. *Evaluating Welfare and Training Programs*. Cambridge, Mass.: Harvard University Press.

Manski, Charles, and Steven Lerman. 1977. "The Estimation of Choice Probabilities from Choice-Based Samples." *Econometrica* 45:1977–88.

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

Rivers, Douglas, and Quang H. Vuong. 1988. "Limited Information Estimators and Exogeneity Tests for Simultaneous Probit Models." *Journal of Econometrics* 39:347–66.

Rosenbaum, P., and D. Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70:41–55.

Rosenbaum, P., and D. Rubin. 1985. "Constructing a Control Group using Multivariate Matched Sampling Methods that Incorporate the Propensity Score." *American Statistician* 39:35–39.