POLICY RESEARCH WORKING PAPER 2153

The Mystery of the Vanishing Benefits

Ms. Speedy Analyst's Introduction to Evaluation

Martin Ravallion

This entertaining introduction to the concepts and methods of impact evaluation — as seen through the eyes of Ms. Speedy Analyst — assumes readers are familiar with basic statistics up to regression analysis (as covered in an introductory text on econometrics).

The World Bank
Development Research Group
Poverty and Human Resources
July 1999

POLICY RESEARCH WORKING PAPER 2153

Summary findings

The setting for this good-natured training guide for impact evaluation is the fictional developing country Labas. Twelve months ago the 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 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 government wants to assess PROSCOL's impact on poverty, to help decide whether the program should be expanded or dropped. The Finance Minister asks the undersecretary, and the undersecretary calls in Ms. Speedy Analyst.

Ms. Speedy Analyst's on-the-job training in how to assess the impact of a social program provides the vehicle through which this paper explains:

- Methods of evaluating a program's impact — randomizing, matching, reflexive comparisons, double difference (or "difference in difference") methods, and instrumental variables methods.
- The types of data used for impact evaluation, typical problems with and uses of data, control variables, instrumental variables, regressions, and so on.
- How to form and match comparison groups.
- Sources of bias.
- The value of baseline surveys.
- Measures of poverty (headcount index, poverty gap index, and squared poverty gap).
- How to compare poverty with and without the program.

This paper — a product of Poverty and Human Resources, Development Research Group — is part of a larger effort in the group to provide useful training tools for Bank staff. Copies of the paper are available free from the World Bank, 1818 H Street NW, Washington, DC 20433. Please contact Patricia Sader, room MC4-773, telephone 202-473-3902, fax 202-522-1153, Internet address psader@worldbank.org. Policy Research Working Papers are also posted on the Web at http://www.worldbank.org/html/dec/Publications/Workpapers/home.html. The author may be contacted at mravallion@worldbank.org. July 1999. (40 pages)

The Mystery of the Vanishing Benefits:

Ms Speedy Analyst's Introduction to Evaluation

Martin Ravallion*

World Bank

* This paper aims to provide an introduction to the concepts and methods of impact evaluation. The paper assumes that readers are familiar with basic statistics up to regression analysis (as would be covered in an introductory text on econometrics). For their comments and discussions I am 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, and Michael Woolcock.

The setting for our story is the developing country, Labas. 12 months ago, its Government introduced an anti-poverty 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 kids in school until 18 years of age. The program is called PROSCOL.

The 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 Finance Ministry could do the evaluation, to help assure independence, 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. 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, Speedy Analyst graduated from the Labas National University, where she did a Masters in Applied Economics. She has worked in the Finance Ministry since then. Speedy has a reputation for combining good common sense with an ability to get the most out of imperfect data. Speedy also knows that she is a bit rusty on the stuff she learnt at LNU.

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."

Speedy thinks this sounds a bit too vague for her liking; what does he mean by "benefiting", she thinks to herself. 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 does not seem entirely comfortable with such a direct question. He answers: "To reduce poverty in Labas, both now and in the future."

Speedy tries to pin this down further. "I see. The cash transfers aim to reduce current poverty, while by insisting that transfer recipients keep their kids in school, the program aims to reduce future poverty".

"Yes, that's right, Speedy".

"So I guess we need to know two things about the program. Firstly, are the cash transfers mainly going to low-income families? Secondly, 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 began Speedy Analyst's on-the-job training in how to assess the impact of a social program. Note 1 summarizes the methods she will learn about over the following days.

The mystery unfolds

Back in her office, Speedy finds that the file from 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 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 kids attend 85% of their school classes, which has to be verified by a note from the school.

Note 1: Methods for evaluating program impact

The essential problem of impact evaluation 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 counter-factual of what would have happened without the program. The comparison group is designed to be representative of the "treatment group" of participants with one key difference: the comparison group did not participate. The main methods available are as follows:

- Randomization, in which the selection into the treatment and comparison groups is random in some well-defined set of people. Then there will be no difference (in expectation) between the two groups besides the fact that the treatment group got the program. (There can still be differences due to sampling error; the larger the size of the treatment and comparison samples the less the error.)
- 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 "propensity score" (predicted probability of participation given observed characteristics); the closer the propensity score, the better the match. A good comparison group comes from the same economic environment and was administered the same questionnaire by similarly trained interviewers as the treatment group.
- Reflexive comparisons, in which a "baseline" survey of participants is done before the intervention, and a follow-up survey done after. The baseline provides the comparison group, and impact is measured by the change in outcome indicators before and after the intervention.
- 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.
- 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.

No method is perfect. Randomization is fraught with problems in practice. Political feasibility is often a problem. And even when selection is randomized, there can still be selective non-participation. Matching methods only deal with observable differences; there will still be a problem of latent heterogeneity, leading to a possible bias in estimating program impact. Selective attrition plagues both randomization and double-difference estimates. It is always desirable to triangulate methods.

The file includes a report, "PROSCOL: Participants' Perspectives", commissioned by MSD and done by a local consultant. The report was based on qualitative interviews with program administrators and focus groups of participants. Speedy cannot tell whether those interviewed were 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 kids went to school, but Speedy wonders whether they might not have also gone to school if the program had not existed.

Speedy 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 schooling. I need hard data." Later Speedy prepares Note 2, summarizing the types of data typically used in impact evaluations.

Note 2: 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 also helps a lot to have a reasonably firm grip on the relevant "stylized facts" about the setting. The relevant facts might include the poverty map, the way the labor market works, the major ethnic divisions, other relevant public programs, etc.
- Be eclectic about data. Sources can embrace both informal, unstructured, interviews with participants in the program as well as quantitative data from representative samples.
- However, it is extremely difficult to ask counter-factual questions in interviews or focus groups; try asking someone who is currently participating in a public program: "what would you be doing now if this program did not exist?" Talking to program participants can be valuable, but it is unlikely to provide a credible evaluation on its own.
- One also needs data on the outcome indicators and relevant explanatory variables. You need the latter to deal with heterogeneity in outcomes conditional on program participation. Outcomes can differ depending on whether one is educated, say. It may not be possible to see the impact of the program unless one controls for that heterogeneity.
- Depending on the methods used (Note 1), 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 non-random programs (Note 1).
- The data on outcomes and other relevant explanatory variables can be either quantitative or qualitative. But it has to be possible to organize it in some sort of systematic data structure. A simple and common example is that one has values of various variables including one or more outcome indicators for various observation units (individuals, households, firms, communities).
- The variables one has data on and the observation units one uses are often chosen as part of the evaluation method. These choices should be anchored to the prior knowledge about the program (its objectives of course, but also how it is run) and the setting in which it is introduced.
- The specific source of the data on outcomes and their determinants, including program participation, typically comes from survey data of some sort. The observation unit could be the household, firm, geographic area, depending on the type of program one is studying.
- Survey data can often be supplemented with useful other data on the program (such as from the project monitoring data base) or setting (such as from geographic data bases).

There is a promising lead in the MSD file. Nine months ago the first national household survey of Labas was done by the Labas Bureau of Statistics (LBS). 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 incomes by source, employment, expenditures, health status, education attainments, and

demographic and other attributes of the family. There is a letter in the file from MSD to LBS, a few months prior to the LSS, asking for a question to be added on whether or not 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 Speedy, and she heads off to LBS.

Speedy Analyst already knows a few things about the LSS, having used tabulations from it produced by LBS. But Speedy worries that she will not be able to do a good evaluation of PROSCOL without access to the raw household-level data. But after a protracted and unsuccessful discussion with the head of the LBS unit in charge of the survey, and seemingly endless follow-up phone calls, Speedy starts to worry whether she will get the data, and have anything on outcomes worth showing her boss.

However, after a formal request from the Minister (which Speedy wrote for him to sign), the Secretary of Statistics finally agrees to give her the micro data. Then, after a few more phone calls, LBS also gives her a copy of the documentation she needs to read the data.

Speedy already knows how to use a statistics package called SAPS. After a long and painful day figuring out how to use the raw LSS data, Speedy starts the real work. She uses SAPS to make a cross-tab 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, Speedy decides to subtract any monies received from PROSCOL; this, she reckons, will be a good measure of income in the absence of the program. So she hopes to reveal who gained according to their pre-intervention income.

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

What about the impact on schooling? She makes another table, giving 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 6-18 is about 80% in both cases.

Speedy 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 Speedy Analyst solves the mystery of the vanishing schooling benefits from PROSCOL.

Speedy Analyst visits Mr. Unbiased Statistica

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

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

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

"What you are doing Speedy 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 non-participants 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 PROSCOL participation of the i'th 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 i'th 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 gain in schooling due to PROSCOL for a child that does in fact participate is:

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

"Why do you need this $\mid$"? asks Speedy.

"That stands for 'given that' or 'conditional on' if you prefer. The ' $\mid$ ' is needed to make it clear that we are calculating the gain for a child who actually participated. If we want to know the average gain we simply take the mean of all the G's. This will give you the sample mean gain in schooling amongst all those who participated in PROSCOL. As long as you have calculated this mean correctly (using the appropriate sample weights from your survey) it will provide an unbiased estimate of the true mean gain. The latter is what statisticians often call the "expected value" of G, and it can be written as:

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

You can think of this as another way of saying 'mean'. However, it need not be exactly equal to the mean you calculate from your sample data, given that there will be some sampling error. In the evaluation literature, $\mathrm{E}(S_{1i}\text{-}S_{0i} \mid P_i=1)$ is sometimes called the 'treatment effect' or the 'average treatment effect on the treated'."

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

"Ah, but that is not what you have calculated Speedy. You have not calculated G but rather the difference in mean schooling between kids 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 the 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 kids who actually participated in PROSCOL had they not participated; that is missing data – it is called a 'counter-factual' mean."

Speedy sees that Statistica has a legitimate concern. In the absence of the program, PROSCOL parents may well send their kids to school less than do other parents. If so, then 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 parents and non-PROSCOL in the absence of the program.

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

"Well, the best way is to assign the program randomly. Then participants and non-participants will have the same expected schooling in the absence of the program, i.e., $E(S_{0i} \mid P_i=1) = E(S_{0i} \mid P_i=0)$. The schooling of non-participating families will then correctly reveal the counter-

factual, i.e., 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 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 non-participants is due to the program."

On recalling what she read in the PROSCOL file, Speedy 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 kids to school, would create bias.

She tells Mr. Statistica about the program's purposive placement.

"So, Speedy, 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."

Speedy returns to her office, despondent. She sees now that the magnitude of this bias that Mr. Statistica is worried about could be huge. Her reasoning is as follows: 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. Non-poor 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, $E(S_{1i} \mid P_i=1) = E(S_{0\,i} \mid P_i=0)$ 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 very substantial under-estimation of PROSCOL's benefits from her comparison of mean schooling between PROSCOL families and non-PROSCOL families.

A visit to Ms Tangential Economiste

Next Speedy visits a colleague at the Ministry of Finance, Tangential Economiste. Tangential specializes in public finance. She has a reputation as a sharp economist in the Ministry, though sometimes a little brutal in her comments on her colleague's work.

Speedy first shows her the cross-tab of amounts received from PROSCOL against income. Tangential immediately brings up a concern, which she chastises Speedy for ignoring. "You have clearly overestimated the gains to the poor from PROSCOL because you have ignored foregone income, Speedy. Kids have to go to school if the family is to get the PROSCOL transfer. So they will not be able to work, either on the family business or in the labor market. Kids aged 15-18 can earn two-thirds or more of the adult wage in agriculture and construction, for example. PROSCOL families will lose this income from their kids' work. You should take account of this foregone 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 pre-intervention income. Only then will you know how poor the family would have been is 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 foregone income from child labor? Less child labor is surely a good thing," Speedy says in defense.

"You should certainly look at the gains from reducing child labor Speedy, 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 increasing the current incomes of the poor, and the other is by increasing their future incomes. The impact on child labor matters to both, but in opposite directions. So PROSCOL faces a trade off."

Speedy 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 foregone income that Tangential is so worried about. Maybe the extra time at school comes out of non-work time.

Next, Speedy tells Tangential about Mr. Statistica's concerns about her second table, to see what she thinks.

"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 or not the child's family was covered by PROSCOL. Why not try this regression?" Tangential 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, while ε is a residual that includes other determinants of schooling, and measurement errors. You can see Speedy that the estimated value of b gives you the impact of PROSCOL on schooling."

"No, I don't see that," Speedy interjects.

"Well, 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."

This discussion puts Speedy in a more hopeful mood, as she returns to her office to try out Tangential's equation. She runs the REGRESS command in SAPS on the regression with and without the control variables Tangential 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 a bunch of control variables in the regression, she immediately sees a positive and significant coefficient on PROSCOL participation. She calculates that by 18 years of age, the program has added two years to schooling.

Speedy 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 Speedy's results

Speedy decides to visit Professor Chisquare, who was one of her teachers at LNU. 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", she recalls thinking during his classes in econometrics. Speedy also recalls her dread at asking Chisquare anything, because his answers were sometimes very hard to understand. "But he knows more about regressions than anyone else I know", Speedy reflects.

She arranges a meeting. Having heard on the phone what her problem is, Chisquare greets his ex-student with a long list of papers to read, mostly with rather impenetrable titles, and published in seemingly obscure places.

"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?"

Chisquare agrees. Speedy shows him the regressions, thinking that he will be pleased that his ex-student has been running regressions. He asks her a few questions about what she had done, and then rests back in his chair, ready, it seems, to pronounce judgement 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 Speedy. "Parents with more schooling would be more likely to send their kids to school, so the gains to them from PROSCOL will be lower".

"Quite possibly, Ms Analyst. To allow the gains to vary with X, let mean schooling of non-participants 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, 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 $(a_1 - a_0) + (c_1 - c_0)X$ is the mean program impact at any given value of X. If you use the mean X in your sample of participants then you will have their mean gain from the program.

"A second concern Ms Analyst is in how you have estimated your regression. The REGRESS command in SAPS is just Ordinary Least Squares. 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."

"Yes, I think I do recall that; but can you remind me what 'exogenous' means?"

"It 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. Is PROSCOL participation exogenous Ms Analyst?"

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

"Your equation for years of schooling is:

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

You used $a + b + cX_i + \varepsilon_i$ as your estimate of the i'th 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.

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

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

$$P_i = d + eZ_i + v_i$$

where Z is a bunch of variables that include 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 v term. Notice too that this equation is linear, yet P can only take two possible values, 0 and 1. Predicted values between zero and one are OK, 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 I will confine attention 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 v 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 Speedy.

"Well, you 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 amongst the poverty proxies used for targeting, or were correlated with them, and they are variables that also influenced schooling.

"However, Ms Analyst, all this only works if the assumptions are valid. There are two problems you should be aware of. Firstly, 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; you can 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 your case. The error term ε in the schooling regression probably contains variables that are not found in the LSS, but might well influence participation in the program, i.e., they might be correlated with the error term ν in the participation equation. If that is the case then the error term ε will not have zero mean given X and P,

and so ordinary regression methods will still be biased when estimating your regressions for schooling. So 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."

Speedy learns about better methods of forming a comparison group

Next Speedy tells 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 divide the sample into groups of families with the same or similar values of X and you then compare 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."

Speedy 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 non-participant with the same observed characteristics. But that would surely be very hard Professor, since I could have a lot of those variables. There may be nobody amongst the non-participants with exactly the same values of all the observed characteristics for any one of the PROSCOL participants"

"Ah", says Chisquare, "some clever statisticians have figured out how you can simplify the problem greatly. Instead of aiming to assure that the matched control for each participant has exactly the same value of X, you can get the same result by matching on the predicted value of P, given X, which is called the propensity score of X. You should read the papers by Rosenbaum and Rubin on the list I prepared for you. Their *Biometrika* 1983 paper shows that if (in your case) schooling without PROSCOL is independent of participation given X then they 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 as you say. And yet propensity score matching is sufficient to eliminate the bias provided there is conditional independence given X."

"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 non-participant with the closest value of this predicted probability. The difference in schooling is then the estimated gain from the program for that participant."

"That is basically 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 could give you crazy predicted probabilities, above 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 Ms Analyst?"

"Yes, I do", says Speedy, as convincingly as she can.

"Another issue you should be aware of Ms Analyst is that some of the non-participants may have to be excluded as potential matches right from the start. Some will be ineligible according to the eligibility rules. Others will be eligible, but have observable characteristics that make participation unlikely. In fact there are important 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 Heckman et al. (1998) paper on my reading list.

"The intuition is that you want the comparison group to be as similar as possible to the treatment group in terms of their likelihood of participating in the program, as summarized by the propensity score. You might find that some of the non-participant sample has a lower propensity score than any of those in the treatment sample. This is sometimes called 'lack of common support'. In forming your comparison group, you should eliminate those observations from the set of non-

participants to assure that you are only comparing gains over the same range of propensity scores. You should certainly exclude those non-participants for whom the probability of participating is zero. It is probably also a good idea to trim a little, say 2%, of the sample from the top and bottom of the non-participant distribution in terms of the propensity scores. Once you have identified participants and non-participants 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?" Speedy asks.

"Well clearly you should include all the variables in your data set that are, or could proxy for, the poverty indicators that were used by MSD in selecting PROSCOL participants. So again X should include the variables in Z.

"However, you have touched on a weak spot of propensity score matching. With randomization, the results do not depend on what X you choose. With matching, a different X will yield a different estimate of impact. Nor does randomization require that you specify some model for participation, whether a logit or something else."

"Yes, Professor, I am convinced that a random experiment is the ideal. Alas, that is clearly not the case with PROSCOL."

Speedy prepares Note 3, summarizing the steps she needs to follow in doing propsensity score matching.

Note 3: Steps in matching

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

Step 1: You need a representative sample survey of eligible non-participants as well as one for the participants. The larger the sample of eligible non-participants the better, to facilitate good matching. If the two samples come from different surveys, then they should be highly comparable surveys (same questionnaire, same interviewers or interviewer training, same survey period and so on).
Step 2: Pool the two samples and estimate a logit model of program participation as a function of all the variables in the data that are likely to determine participation.
Step 3: Create the predicted values of the probability of participation from the logit regression; these are called the "propensity scores". You will have a propensity score for every sampled participant and non-participant.
Step 4: Some of the non-participant sample may have to be excluded at the outset because they have a propensity score which is outside the range (typically too low) found for the treatment sample. The range of propensity scores estimated for the treatment group should correspond closely to that for the retained sub-sample of non-participants. You may also want to restrict potential matches in other ways, depending on the setting. For example, you may want to only allow matches within the same geographic area to help assure that the matches come from the same economic environment.
Step 5: For each individual in the treatment sample, you now want to find the observation in the non-participant sample that has the closest propensity score, as measured by the absolute difference in scores. This is called the "nearest neighbor". You 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 incomes in the non-participant sample.

This is the simplest form of propensity score matching. Complications can arise in practice. For example, if there is over-sampling of participants then you can use choice-based sampling methods to correct for this (Manski and Lerman, 1978); 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 non-participants as potential matches but weight them differently, according to how close they are (Heckman et al., 1998).

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 programs' impact."

"Ah, there is no guarantee of that. Recall my warning 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. 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'."

Speedy's Note 4 elaborates on this difference.

Note 4: Sources of bias in naïve estimates of PROSOL's impact

The bias described by Mr. Statistica 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. Firstly 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. Secondly, 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 miss-weighted. 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 non-experimental 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 to various non-experimental estimates (Heckman et al., 1998). The bias in the naïve estimate was huge, but careful matching of the comparison group based on observables greatly reduced the bias.

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 still leave a bias. This is sometimes called 'selection on unobservables'."

Speedy interjects. "So really the conditions required for justifying the method suggested by Ms Economiste are no less restrictive than those needed to justify a version of my 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."

"That's right, Ms Analyst. Intuitively, one might think that careful matching reduces the bias, but that is no necessarily so. Matching eliminates part of the bias in your 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 et al. (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. That is an empirical question."

Speedy regrets that a baseline survey was not done

Speedy 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 non-participants, 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."

Chisquare turns to his whiteboard 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 reveal those troublesome unobservables. Notice that since 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 conveniently swept away."

Speedy 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."

Speedy 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 changed between the two surveys.

Something about 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 assure 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 assure that comparisons were made in a region of common support was a major source of bias in the double difference estimate when compared to 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 et al., in Econometrica 1998 on my reading list."

Speedy 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 who were originally included in the baseline survey. Some people in the baseline survey may not want to be interviewed again, or they have moved to an unknown location. Is that a problem?"

"If the drop outs 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 Speedy writes up Note 5, on the steps to form a double-difference estimate.

Note 5: 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 Note 3) both before and after the intervention. The main steps are as follows:
Step 1: You need a "baseline" survey before the intervention is in place, and the survey must cover both non-participants and participants. If you do not know who will participate, you have to make an informed guess. Talk to the program administrators.
Step 2: You then need one or more follow-up surveys, after the program is put in place. These should be highly comparable to the baseline surveys (in terms of the questionnaire, the interviewing, etc). Ideally the follow-up surveys should be of the same sampled observations as the baseline survey. If this is not possible then they should be the same geographic clusters, or strata in terms of some other variable.
Step 3: Calculate the mean difference between the "after" and "before" values of the outcome indicator for each of the treatment and comparison groups.
Step 4: Calculate the difference between these two mean differences. That is your estimate of the impact of the program.
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.

Chisquare reminds Speedy 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 Speedy. "You must 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 exogeneous 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 you 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 by its predicted value conditional on Z. Since 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 Speedy. "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, since they will be uncorrelated with the error term in the schooling regression."

"You've got it Ms Analyst. That also suggests another, more efficient, 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, since the participation residual will be perfectly predictable from actual participation and X, in a linear model.

"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. This overestimation in the variance of P leads naturally to an underestimation of its coefficient b."

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

"That's right. So you can see how useful an instrumental variable can be. However, you do have to be a little careful in practice. When you just replace the actual participation with its predicted value and run OLS you will not give the correct standard errors since the computer will not know that you had to use previously estimated parameters to obtain the predicted values. A correction to the OLS standard errors is required, though there are statistical packages that allow you to do this easily, at least for linear models.

"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 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, in the programming language called Gauss, 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 Speedy.

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

Speedy later summarizes what she has leant about alternative methods, as in Note 1.

Speedy returns to her computer

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

Speedy 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", which justifies putting some variable in the equation for participation, but not in the equation for schooling.

Speedy decides to try the "propensity score matching method" suggested by 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 had advised her. On comparing the mean school enrollment rates, Speedy finds that kids of the matched comparison group had an enrollment rate of 60%, as compared to the figure of 80% for PROSCOL families.

She now thinks back on those comments that Ms Tangential Economiste had made about foregone income. She finds that the Bureau of Statistics did a special survey of child labor that asked about earnings. (There is an official ban on kids 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 what earnings a child would have had if she had not gone to school.

So Speedy can now subtract from PROSCOL's cash payment to participants the amount of foregone income, and so work out the net income transfer. Subtracting this net transfer from total income, she can now 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 foregone income) but they are still poor; for example, two-thirds of them are below Labas' official poverty line.

Having calculated the net income gain to all participants, Speedy 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 foregone income from kids' work) attributed to PROSCOL from all the observed incomes she gets a new distribution of pre-intervention incomes. The poverty rate without the program is then the proportion of people living in poor households, based on this new distribution. Speedy finds that the observed poverty rate in Northwest Labas of 32% would have been 36% if PROSCOL had not existed. The program allows 4% of the population to escape poverty now. The schooling gains mean that there will also be both pecuniary and non-pecuniary gains to the poor in the future.

Speedy recalls a class Chisquare gave on poverty measurement, in which he pointed out that the proportion of people below the poverty line is a rather crude measure, since it tells you nothing about changes below the line. Note 6 reproduces (after some tidying up) Speedy's class notes. When Speedy calculates both the poverty gap index and the squared poverty gap index the results suggest that these have also fallen as a result of PROSCOL.

Note 6: Poverty measures

The simplest and most common poverty measure is the headcount index. In Labas this 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; not will it reflect gains amongst the poor, unless they cross the poverty line.

A widely used alternative to the headcount index is the poverty gap index (PG). 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 amongst the poor due to 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.

There are various "distribution-sensitive" measures that will reflect such changes in distribution amongst the poor. One such measure is the squared poverty gap (Foster et al., 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 non-poor.) 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 non-poor. Atkinson (1987) describes other examples in the literature.

Speedy 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% of the population are poor based on the observed post-intervention incomes, the proportion would have been 52% without PROSCOL. At a poverty line which 15% fail to reach with the program, the proportion would have been 19% without it. By repeating these calculations over the whole range of incomes, Speedy realizes that she has traced out the entire "poverty incidence curves" with and without the program, which are just the same thing statisticians call the "cumulative distribution function".

Note 7 summarizes the steps Speedy takes in making comparisons of poverty with and without PROSCOL.

Note 7: Comparing poverty with and without the program

Using the methods described in the main text and earlier Notes 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. You can then summarize this information in the form of poverty incidence curves (PICs), with and without the program.

Step 1: You should already have the post-intervention income (or other welfare indicator) for each household in the whole sample (comprising both participants and non-participants); this is data. You also know how many people are in each household. And, of course, you know the total number of people in the sample (N; or this might be the estimated population size, if inverse sampling rates have been used to "expend up" each sample observation).
Step 2: You can plot this information in the form of a PIC. This gives (on the vertical axis) the percentage of the population living in households with an income less than or equal to that value on the horizontal axis. To make this graph, you can start with the poorest household, mark its income on the horizontal axis, and then count up on the vertical axis by 100 times the number of people in that household divided by N. The next point is the proportion living in the two poorest households, and so on. This gives the post-intervention PIC.
Step 3: Now calculate the distribution of income pre-intervention. To get this you subtract the estimated gain for each household from its post-intervention income. You then have a list of post-intervention incomes, one for each sampled household. Then repeat Step 2. You will then have the pre-intervention 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 headcount index for that poverty line (Note 5). Alternatively, looking horizontally gives you the income gain at that percentile. If none of the gains are negative then the post-intervention PIC must lie below the pre-intervention on. 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.)

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 Note 5. 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 Note 5 is unambiguous.

Speedy 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, Speedy accidentally bumps into her old friend, Sensible Sociologist, who now works for one of Labas' largest NGOs, SCEF (the Social Capital for Empowerment Foundation). Speedy tells Sense all the details about what she has been doing on PROSCOL.

Sensible Sociologist's eyes start to roll when Speedy talks about "unbiased estimates" and "propensity scores". "I don't know much about 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 kids, 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 'foregone income' that Tangential came up with, I am not so sure about that. Economists have strange ideas sometimes. I have seen plenty of kids from poor families who work as well as go to school. And some of the younger ones not at school don't seem to be working either. 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 kids versus a matched comparison group," says Speedy. "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 foregone income."

"That might be wise" says Sensible 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 the family lives in. All school areas (SBA) 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. I guess they like to let all areas participate for political

reasons. As a result, it is relative poverty—relative to others in the area you live—that matters much more than your absolute level of living."

"No I did not know that", replies Speedy, a little embarrassed that she had not thought of talking to Sensible 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 to 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 SBA's can be used as instrumental variables to remove the bias in my estimates of program impact. "

Sensible Sociologist's eyes roll again, as Speedy 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.

Speedy re-runs 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 SBS. She realizes that 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, Speedy realizes that any omitted school characteristics that jointly influence PROSCOL allocations by SBA and individual schooling outcomes will leave a bias in her IV estimates. She realizes that she will never rule out the possibility of bias, but with plenty of geographic control variables, this method should at least offer a credible comparator to her matching estimate.

Soon she has the results. Consistent with Sense'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 enrolment, 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, for boys and girls separately. For either boys or girls of 18 years, her results indicate that they would have dropped out of school almost two years earlier if it had not been for PROSCOL.

Speedy 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?" Speedy decides she will order that new program FEM (Fancy Econometric Methods) that she has heard about. But that will have to wait. For now, Speedy is happy that her results are not very different from those she got using the propensity-score matching method. And she is re-assured somewhat by Sense's comments based on her observations in the field. "They can't all be wrong".

Speedy reports back to her boss

Speedy 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 Speedy to attend. The Minister is interested in Speedy's results, and asks some questions about how she figured out the benefits from PROSCOL. He seems to appreciate Speedy's efforts to assure that the comparison group is similar to 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. But I want you to keep monitoring the program Speedy."

"I would like to do that, Minister. However, I have learnt 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".

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 re-elected?"

"I see your point, Minister. But since 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, amongst eligible participants? What could be fairer?"

The Minister thinks it over. "What about if we picked the schools or the school board areas randomly, in the first wave?"

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

"Instrumental what?", says the Minister, while 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 north 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, Speedy. I will also talk to the Secretary of Statistics".

Epilogue

It is three years later. Speedy 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' social programs on a regular basis. Speedy has a permanent staff of three assistants. She regularly hires both Professor Chisquare and Sensible Sociologist as consultants. They have a hard time talking to each other. ("Boy, that Chisquare is just not normal" Sense confided to Speedy one day.) But Speedy finds it useful to have both of them around. The qualitative field trips, and interviews with stake-holders, that Sense favors help a lot in forming hypotheses to be tested and assessing the plausibility of key assumptions made in the quantitative analysis that Chisquare favors. Speedy 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, Speedy has also leant 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. Chisquare's advice remains valuable in suggesting alternative methods in the frequent situations of less than ideal data, and pointing out the pitfalls. Speedy has also learnt to be eclectic about data.

The Finance Minister did eventually convince the Minister of Social Development to randomize the first tranche allocation of PROSCOL II across school board areas in the rest of Labas, and this helped Speedy identify the program's impact. Her analysis of the new question on child labor added to the LBS survey revealed that there was some foregone income from PROSCOL, though not quite as much as she had first thought.

Tangential Economiste made a further comment on Speedy's first report on PROSCOL, to the effect that Speedy could also measure the future income gains from PROSCOL, using recent

work by labor economists on the returns to schooling in Labas. When Speedy 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.

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

However, everything has not gone smoothly. At first she had a lot of trouble getting the relevant line ministries to cooperate with her. It is often hard to get them to define the objectives of each program she is evaluating; Speedy sometimes thinks that getting the relevant line ministry to define the objectives of its public spending is an important contribution in its own right. But eventually the line Ministries realize 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 Advisor to the unit, and Mr. Statistica is his representative. Speedy often commissions new surveys from LSB and advises them on questionnaire design and sampling.

Speedy has also started giving advice to other countries and international agencies (including the World Bank) embarking on impact evaluations of social programs. And she has found that swapping notes with other program analysts can be valuable. For example, I learnt about Speedy's interesting experience with PROSCOL on a recent mission to Labas, and I also told here about recent work evaluating a World Bank supported anti-poverty program in Argentina (Jalan and Ravallion, 1999; Note 8 summarizes the methods and results of that study). Speedy reckons there are policy mysteries galore in Labas and elsewhere — mysteries that the tools she has learnt to use might well throw light on.

Note 8: An example for another anti-poverty program

With support from the World Bank, Argentina introduced the Trabajar Program in 1997, in response to a sharp increase in unemployment, and evidence that this was especially hurting the poor. The program aimed to provide useful work on community projects in poor areas work for unemployed workers from poor families. Jalan and Ravallion (1999) assessed the income gains to the families of participating workers and examined how well targeted the gains were. A survey was done of a random sample of participating families, at the same time, and using the same survey instrument and interviewers, as a pre-planned large national sample survey. A logit model of program participation was first estimated on the pooled sample and the propensity scores were then calculated. The matching methods described in Note 3 were then used to draw a control group from the larger cross-sectional survey. The participants sample had a mean propensity score of 0.40, while it was 0.075 for the national sample. So the national sample is clearly unrepresentative of Trabajar participants. After matching, however, the comparison group drawn from the national sample also had a score of 0.40. The results indicated that income gains were about half of the gross wage on the program (the difference being due to lost income from work that had to be given up to join the program). About 80% of the families of participating workers came from the poorest 20% of all families in Argentina, in terms of (pre-intervention) income per person. A test for selection bias in the resulting matching estimator was also done using instrumental variables. The bias in the matching estimates was negligible.

References (including Professor Chisquare's Reading List for Speedy)

Atkinson, Anthony, 1987, "On the Measurement of Poverty", *Econometrica*, 55: 749-64.

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.

Foster, James, J. Greer, and Erik Thorbecke, 1984, "A Class of Decomposable Poverty Measures", *Econometrica*, 52: 761-765.

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): 441-461.

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: 1017-1099.

Jalan, Jyotsna and Martin Ravallion, 1999, "Income Gains from Workfare and their Distribution", Policy Research Working Paper, World Bank, Washington DC.

Meyer, Bruce D., 1995, "Natural and Quasi-Experiments in Economics", *Journal of Business and Economic Statistics*, April.

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.

Ravallion, Martin, 1994, *Poverty Comparisons*, Fundamentals in Pure and Applied Economics Volume 56, Harwood Academic Publishers.

Rivers, Douglas and Quang H. Vuong, 1988, "Limited Information Estimators and Exogeneity Tests for Simultaneous Probit Models", *Journal of Econometrics*, 39: 347-366.

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.

Policy Research Working Paper Series

	Title	Author	Date	Contact for paper
WPS2136	An Empirical Analysis of Competition, Privatization, and Regulation in Telecommunications Markets in Africa and Latin America	Scott J. Wallsten	June 1999	P. Sintim-Aboagye 38526
WPS2137	Globalization and National Development at the End of the 20th Century: Tensions and Challenges	Andrés Solimano	June 1999	D. Cortijo 84005
WPS2138	Multilateral Disciplines for Investment-Related Policies	Bernard Hoekman Kamal Saggi	June 1999	L. Tabada 36896
WPS2139	Small States, Small Problems?	William Easterly Aart Kraay	June 1999	K. Labrie 31001
WPS2140	Gender Bias in China, the Republic Of Korea, and India 1920–90: Effects of War, Famine, and Fertility Decline	Monica Das Gupta Li Shuzhuo	June 1999	M. Das Gupta 31983
WPS2141	Capital Flows, Macroeconomic Management, and the Financial System: Turkey, 1989–97	Oya Celasun Cevdet Denizer Dong He	July 1999	L. Nathaniel 89569
WPS2142	Adjusting to Trade Policy Reform	Steven J. Matusz David Tarr	July 1999	L. Tabada 36896
WPS2143	Bank-Based and Market-Based Financial Systems: Cross-Country Comparisons	Asli Demirgüç-Kunt Ross Levine	July 1999	K. Labrie 31001
WPS2144	Aid Dependence Reconsidered	Jean-Paul Azam Shantayanan Devarajan Stephen A. O'Connell	July 1999	H. Sladovich 37698
WPS2145	Assessing the Impact of Micro-credit on Poverty and Vulnerability in Bangladesh	Hassan Zaman	July 1999	B. Mekuria 82756
WPS2146	A New Database on Financial Development and Structure	Thorsten Beck Asl1 Demirgüç-Kunt Ross Levine	July 1999	K. Labrie 31001
WPS2147	Developing Country Goals and Strategies for the Millennium Round	Constantine Michalopoulos	July 1999	L. Tabada 36896
WPS2148	Social Capital, Household Welfare, And Poverty in Indonesia	Christiaan Grootaert	July 1999	G. Ochieng 31123

Policy Research Working Paper Series

	Title	Author	Date	Contact for paper
WPS2149	Income Gains to the Poor from Workfare: Estimates for Argentina's Trabajar Program	Jyotsna Jalan Martin Ravallion	July 1999	P. Sader 33902
WPS2150	Who Wants to Redistribute? Russia's Tunnel Effect in the 1990s	Martin Ravallion Michael Lokshin	July 1999	P. Sader 33902
WPS2151	A Few Things Transport Regulators Should Know about Risk and the Cost Of Capital	Ian Alexander Antonio Estache Adele Oliveri	July 1999	G. Chenet-Smith 36370
WPS2152	Comparing the Performance of Public and Private Water Companies in the Asia and Pacific Region: What a Stochastic Costs Frontier Shows	Antonio Estache Martin A. Rossi	July 1999	G. Chenet-Smith 36370