Chapter 12
Impact Evaluation

12.1 Introduction

A government sets up a scheme for extending microcredit to farmers; or builds an irrigation canal; or provides free textbooks to 10-year-olds; or introduces supplemental nutrition for pregnant mothers; or strengthens the social security net with a food-for-work program.

All of these activities sound potentially promising. But do they really work? Increasingly, governments and donors want clear and rigorous answers before channeling funds into such schemes. And that calls for an impact evaluation.

Generally, an impact evaluation seeks to measure the changes in well-being that can be attributed to a particular project or policy (an "intervention" or "treatment"). The results of an impact evaluation can show which interventions have been effective, and so inform decisions on whether they should be eliminated, modified or expanded, and what priority they should be accorded. Impact evaluations are also essential pre-requisites for *ex post* cost–benefit or cost–effectiveness analyses, which weigh program costs against the benefits they deliver.

Impact evaluations are expensive and can be technically complex. Baker (2000, p.79) lists the expenses of undertaking seven high-quality impact evaluation studies: the average cost was $433,000, representing 0.56% of total project costs, with half of the expense going to data collection. It thus makes sense to undertake an impact evaluation only if (a) the policy or program is of strategic importance, or is innovative, and (b) the information from the evaluation is likely to fill gaps in current knowledge, and (c) someone might act on the basis of the results. This latter point is worth emphasizing, because not everyone welcomes impact evaluations. Every evaluation carries the risk that it will find that a program was ineffective or an organization might no longer be justified. In 2006 we were approached to undertake an impact evaluation of a large project in a middle-income country, but were then made to understand that such an evaluation would only be acceptable if it made the sponsoring agency look good (and ultimately the evaluation was not funded).

D. Haughton and J. Haughton, *Living Standards Analytics*, Statistics for Social and Behavioral Sciences, DOI 10.1007/978-1-4614-0385-2_12,

In this chapter, we set out and appraise the main approaches to impact evaluation. In addition to using illustrations from the academic literature, we draw extensively on a study of the impact of a major microcredit scheme in Thailand in order to show how the techniques of impact evaluation are applied. For a recent book-length treatment, see Khandker et al. (2010).

12.2 General Principles

The central idea of impact evaluation is straightforward: we need to compare the actual outcome of the intervention with our evaluation of what would have happened in the absence of the intervention (the *counterfactual*). The central challenge of impact assessment is constructing a plausible counterfactual.

The challenge is a difficult one. Consider the case of a program that provides additional food – maize, milk powder – to poor mothers with infants. Now suppose that the data show that the mothers and infants covered by the program are less well nourished than those who are not covered. Are we to conclude that the project is a failure?

Perhaps; but then again, it is likely that the project targeted poor mothers with malnourished infants – that was probably the whole point of the program! – so it is not surprising that households with underweight children are getting additional food. The problem here is one of estimating how malnourished the mothers and infants covered by the program would have been in the absence of the program, in other words, establishing an appropriate counterfactual (Ravallion 1999).

12.2.1 Case: The Thailand Village Fund

We illustrate many of the ideas of this chapter using as an example the evaluation of the impact of a major microcredit scheme in Thailand. In this section we provide the relevant background for this example, which is based on Boonperm et al. (2009).

In 2001 the newly elected government of Thaksin Shinawatra and the Thai Rak Thai Party established the Thailand Village and Urban Revolving Fund (VRF) program, which proposed to provide a million baht (about $22,500) to every village and urban community in Thailand as working capital for locally run rotating credit associations. Since there are almost 74,000 villages and over 4,500 urban communities in the country, this represented an injection of about 75 billion baht, equivalent to about $1.75 billion. The program was put into place rapidly, reaching over 90% of villages by 2004. By the end of May 2005 the TVF committees had lent a total of 258 billion baht ($6.9 billion) in 17.8 million loans, representing an average loan of $387. Total repayment of principal came to 168 million baht, leaving outstanding principal of 91 billion baht.

The question to be asked is whether the VRF had an impact on household expenditures, income, and asset accumulation. It is not self-evident that there

Table 12.1 Summary of use of Thailand Village Revolving Credit Fund, 2004

	All	Poorest fifth	Rural	Female
Number of observations (adults)	80,950	13,180	30,892	43,916
Expenditure per capita (baht/month)	3,398	1,060	2,578	3,427
Adult obtained $\geq$ 1 VRF loan since 2002 (%)	17	20	22	16
Reason for *not* borrowing from VRF				
No need (%)	29	16	25	29
Did not like to be in debt (%)	30	38	33	30
Amount borrowed (baht)	16,183	17,312	16,462	15,322
Annualized interest rate (%)	6.0	5.8	5.9	6.1
Main objective for obtaining loan				
Agricultural equipment/inputs (%)	40	45	42	35
Buy animals	10	12	10	8
Borrowed elsewhere to repay VRF loan (%)	16	19	17	17

Source: Boonperm et al. (2009)

would be any effect: If financial markets operate well – information is cheap and readily available, there are no policy distortions – then households should already have access to as much credit as they can productively use, and they would just substitute VRF credit for other sources of credit. On the other hand, it is not unreasonable to think that there are imperfections in the market for credit: credit markets have well-known informational asymmetries that village-level credit associations may be able to attenuate, given their (presumably) better knowledge about the ability of villager households to service loans.

The data for the impact evaluation come from the Thailand Socioeconomic Surveys of 2002 and 2004. The 2004 survey interviewed 34,843 households (representing 116,444 people) throughout the country. The data were collected in four rounds, spread throughout the year, and the survey used stratified random sampling (by province) with clustering. The 2002 survey used substantially the same questionnaire and covered 34,785 households. Both surveys collected information on income and expenditure, as well as an array of other socioeconomic variables. An effort was made in 2004 to resurvey all 6,309 households that had been surveyed in rural areas in rounds 2 and 3 of the 2002 survey; of these, 5,755 households were actually resurveyed, representing an annual attrition rate of 4.5%.

A selection of summary information on the VRF is shown in Table 12.1, and come from a special module that was included in the 2004 socioeconomic survey and that asked all adult members of households about their experiences with the VRF. By 2004 a sixth of all adults had borrowed at least once from the VRF, with higher proportions of borrowers among the poor and in rural areas. Adults in 31% of households had borrowed from the VRF by 2004, with an average loan of 16,183 baht ($390). Among those who did not borrow, 29% said they did not need a

loan and 30% said that they did not want to be in debt. Proponents of the VRF had hoped and expected that it would mainly stimulate non-farm business, but in over half of all cases borrowers reported that their main objective for obtaining the loan was to purchase agricultural inputs, animals, or farm land.

It is striking that VRF borrowers had significantly lower incomes (3,209 baht per person per month) than the full sample (4,987 baht). But clearly one cannot conclude that the fund made people poorer. It follows that a more sophisticated method is needed to try to measure the impact of the VRF than a simple comparison of outcomes between the treated group (i.e., borrowers) and the comparators (i.e., nonborrowers).

12.2.2 A More Formal Treatment

It is helpful to treat the problem somewhat more formally. Let us suppose that we are interested in the impact of a program on some outcome variable Y_i. This will often be a standard monetary measure of well-being such as income or expenditure per capita, but there are many other possibilities, depending on the issue at hand, such as school performance, household assets, nutritional levels, and the like. We have observations of Y_i for each unit i (e.g., individual, household) from a sample of size n.

Some of the units have been subject to the intervention ("treated"), in which case we let $T_i = 1$; the remainder are untreated, in which case $T_i = 0$. Following the notation favored by Ravallion (2008), let Y_i^{T} be the value of the outcome for unit i under treatment and Y_i^{C} be the outcome for unit i if not treated (i.e., under the counterfactual). The gain from the treatment for any unit i is defined as

$$G_i \equiv Y_i^{\mathrm{T}} - Y_i^{\mathrm{C}}. \tag{12.1}$$

This is the impact (or "causal effect") of the program that we want to measure. But we cannot do this directly, because an individual is either in the treatment group (so we observe Y_i^{T}) or the comparison group (so we observe Y_i^{C}), but never in both. Thus we are faced with a problem of missing data.

In practice we are usually interested in estimating the average impact of a program or project. There is more than one way to construct an average based on (12.1). The commonest measure is the average treatment effect on the treated, given by the expected gain

$$G^{\mathrm{TT}} = E(Y_i^{\mathrm{T}} - Y_i^{\mathrm{C}} \mid T_i = 1), \tag{12.2}$$

where $E(\cdot)$ is the expectations operator. The G^{TT} measure averages the impact over those who are actually treated, for whom $T_i = 1$. In these cases we observe Y_i^{T} but have to figure out a way to estimate Y_i^{C}. Analysts and politicians are most

often interested in knowing whether a program benefited those for whom it was intended, in which case G^{TT} is the appropriate measure. Occasionally researchers are interested in the average treatment effect on the untreated:

$$G^{\mathrm{TU}} = E(Y_i^{\mathrm{T}} - Y_i^{\mathrm{C}} \mid T_i = 0), \tag{12.3}$$

in which case we observe Y_i^{C} but not Y_i^{T}. The combined average treatment effect is a weighted average of these two effects, given by

$$G^{\mathrm{ATE}} = G^{\mathrm{TT}} \Pr(T = 1) + G^{\mathrm{TU}} \Pr(T = 0), \tag{12.4}$$

and is also widely used.

A number of methods ("evaluation designs") have been developed to measure the average impacts, and we examine them in more detail below. But a natural place to start would be to try to measure the impact of a program by taking the difference in the outcome variable between the treated and the untreated. This unconditional single difference estimate is given by

$$D = E(Y_i^{\mathrm{T}} \mid T_i = 1) - E(Y_i^{\mathrm{C}} \mid T_i = 0). \tag{12.5}$$

In the case of our example of the Thailand Village Fund we have, for per capita income (in baht per month) in 2004

$$D = 3,209 - 6,088 = -2,879$$

and for per capita expenditure

$$D = 2,549 - 4,286 = -1,737.$$

Taken at face value, this would imply that borrowing from the Village Fund left households worse off, which is hardly credible.

The problem is that this simple difference is typically subject to bias. Quite generally,

$$D = G^{\mathrm{TT}} + B \tag{12.6}$$

where the selection bias (B) is given by

$$B = E(Y_i^{\mathrm{C}} \mid T_i = 1) - E(Y_i^{\mathrm{C}} \mid T_i = 0). \tag{12.7}$$

The bias is given by the difference in outcomes, without the treatment, between those who are treated and those who are not. We note in passing that the first term in (12.7) is not observed, so the bias cannot be measured directly.

Consider the case of an anti-poverty program that is targeted at raising the incomes of poor households. Then, by design, the treated are likely to be poorer than the untreated, so

$$E(Y_i^{\mathrm{C}} \mid T_i = 1) < E(Y_i^{\mathrm{C}} \mid T_i = 0), \tag{12.8}$$

which means that the bias is negative. It follows from (12.6) that in such cases the simple difference in outcomes (such as income) between the treated and untreated group, given by D, will underestimate the impact G^{TT}.

The bias disappears if we can assume that the assignment of treatment, conditional on a set of covariates X, is independent of the value of the outcomes. This is the key assumption on which all impact evaluation rests. Imbens (2004, p. 7) formalizes it as the assumption of *unconfoundedness*:

$$(Y_i^{\mathrm{T}}, Y_i^{\mathrm{C}}) \perp T_i \mid X_i, \tag{12.9}$$

where $\perp$ is the independence operator. In other words, we need to assume that the treatments are not assigned in a way that is systematically related to the outcome variable, once we have controlled for the effects of the X covariates. Depending on the author and the literature, (12.9) is also referred to as the assumption of ignorable treatment assignment; or the conditional independence assumption; or the approach of selection on observables.

Every impact assessment has to make the case that unconfoundedness (or a slightly weaker version such as conditional exogeneity of treatment, addressed below) is plausible, because otherwise it is impossible to identify the effects of the treatment.

12.3 Experimental Design

One elegant way to ensure unconfoundedness, and thereby to solve the problem of bias, to assign treatments randomly. This *experimental design* approach ensures, by construction, that the expected value of the outcome variables can be assumed to be the same for the treatment group and the *control group* in the absence of treatment, which means that in this case

$$B = E(Y_i^{\mathrm{C}} \mid T_i = 1) - E(Y_i^{\mathrm{C}} \mid T_i = 0) = 0. \tag{12.10}$$

Having eliminated bias, the single difference in the mean values of the outcome variables between the control and treatment groups (12.5) can be attributed to the effects of the intervention, give or take some sampling error. Equivalently, the impact may be measured by the estimated coefficient $\hat{b}$ from the following regression:

$$Y_i = a + bT_i + \varepsilon_i \tag{12.11}$$

Table 12.2 Examination results (% of correct answers), practice exams for the Kenya Certificate of Primary Education, 8th grade, Busia and Teso districts

	School got flip charts	School did not get flip charts	Difference	SD of difference
July 1997	45.5	46.0	−0.5	12.5
July 1998	42.7	42.9	−0.3	11.2

Source: Glewwe et al. (2000)

where ε_i is an error term that is assumed to be normally distributed with variance σ^2, Y_i is defined as

$$Y_i \equiv Y_i^{\mathrm{T}} T_i + Y_i^{\mathrm{C}}(1 - T_i) \tag{12.12}$$

and T_i takes on a value of one if the unit is treated, and of zero otherwise. Estimating (12.11) makes it particularly easy to obtain a confidence interval for the treatment effect.

12.3.1 Case Study: Flip Charts in Kenya

Glewwe et al. (2000) present an interesting, if rare, example of pure randomization. The question that they address is whether flip charts – large, spiral-bound wall charts that can be used in high school classrooms – improve student learning, as measured by test scores. In 1997, a Dutch NGO provided funding for flip charts in 89 schools in the relatively poor Busia and Teso districts of western Kenya. The schools were chosen randomly from a total of 178 schools in these districts and the charts were distributed in early 1997.

The essential results of the study are shown in Table 12.2, and represent the percent of correct responses on standardized national tests given to eighth-graders. As one would expect with random assignment, in July 1997 the test scores did not differ between the schools that received flip charts and those that did not; this test was administered shortly after the flip charts were distributed and represents a benchmark position. Interestingly, there was no statistical difference in test scores between treatment and control schools in the July 1998 tests either, suggesting that the flip charts did not affect academic performance, even though Glewwe et al. found that teachers knew about the flip charts and used them regularly, and none of them had been lost. More complete regression results confirm the essential conclusion: the flip charts had no discernible effect.

The Dutch NGO also provided flip charts to a hundred schools elsewhere in Kenya, but the schools were not chosen randomly (although how they were selected is not entirely clear). Glewwe et al. compared the examination scores

of schools that were given flipcharts with a comparison group of those that were not, and obtained the following regression results:

$$\begin{aligned}\text{Test scores} &= 0.192\ \text{Number of flipcharts} + \text{School random effects},\\ &\quad \text{SE} = 0.080\\ &\quad + \text{Subject and grade fixed effects}.\end{aligned}$$

The key point here is that flip charts in this case are associated with substantially higher test scores. Indeed the effect is twice as strong as that of providing textbooks, a measure that would cost ten times as much! Glewwe et al. argue that these latter results do not provide evidence that flip charts work; it is entirely possible that schools that received flipcharts were different in some systematic (but unobserved) way – perhaps they were more accessible, or richer – and these characteristics cannot be disentangled from the effects of the flip charts. This is a classic case of omitted variables bias, and underscores the importance of random assignment when conducting social experiments.

Another good example of a study that uses randomization is the research by Angrist et al. (2002) on school vouchers in Colombia. In 1991, the government of Colombia established the PACES (Programa de Ampliacion de Cobertura de la Educacion Secundaria[1]) program, which provided vouchers (i.e., scholarships) to students who had applied to and been accepted into private secondary schools. The vouchers were awarded based on a lottery; this provided the randomization that allowed the authors to compare the outcomes for applicants who received vouchers with the outcomes for those who did not.

One of the more interesting findings of this study is that voucher winners were 15–16 percentage points more likely to be in a private school when they were surveyed in 1998. It also appears that the program had a positive and significant effect on the number of years of schooling completed: Those who received vouchers in 1995 in the capital (Bogotá) completed 0.12–0.16 more years than those who did not. Furthermore, repetition rates fell significantly as a result of the project: In the 1995 Bogotá sample, the probability of repetition was reduced by 5–6 percentage points for lottery winners.

12.3.2 Partial Randomization

Pure randomization, which would be required for (12.11) to be appropriate, "is virtually inconceivable for anti-poverty programs" (Ravallion 2008, p. 19); after all, the point is that such programs should be geared to helping the poor, and so they are unlikely to be relevant or appropriate for a significant segment of society.

Thus in practice it is more common to find *partial randomization*, under which the treatment and control samples are chosen randomly, conditional on some

[1] "Program for the Expansion of Educational Coverage."

observable variables, X, that might include measures such as location, or age of the head of the household.

This *conditional exogeneity of program placement* may allow one to estimate the impact of a treatment using a parametric model with controls. Suppose that we may assume

$$Y_i^{\mathrm{T}} = \alpha^{\mathrm{T}} + X_i\beta^{\mathrm{T}} + v_i^{\mathrm{T}}, \quad i = 1, ..., n \tag{12.13a}$$

$$Y_i^{\mathrm{C}} = \alpha^{\mathrm{C}} + X_i\beta^{\mathrm{C}} + v_i^{\mathrm{C}}, \quad i = 1, ..., n \tag{12.13b}$$

where the error terms are normally distributed with zero means and constant variances. These two equations are often estimated together in the form of a switching regression using the pooled data from both the treatment and control samples, giving

$$Y_i = \alpha^{\mathrm{C}} + (\alpha^{\mathrm{T}} - \alpha^{\mathrm{C}})T_i + X_i\beta^{\mathrm{C}} + X_i(\beta^{\mathrm{T}} - \beta^{\mathrm{C}})T_i + \varepsilon_i^{\mathrm{T}}, \quad i = 1, ..., n \tag{12.14}$$

where T_i takes on the values of one or zero and the error term now takes the form $\varepsilon_i = T_i(v_i^{\mathrm{T}} - v_i^{\mathrm{C}}) + v_i^{\mathrm{C}}$.[2] If we may assume that the error term (the "latent effects") has zero mean conditional on the X covariates and treatment – a reasonable assumption if there is even partial randomization – then we have $E(v_i^{\mathrm{T}}|X, T = t) = E(v_i^{\mathrm{C}}|X, T = t) = 0,\ t = 0, 1,$ and we can get consistent estimates of the average treatment effects by applying OLS to (12.14), and noting that[3]

$$G^{\mathrm{ATE}} = E[\alpha^{\mathrm{T}} - \alpha^{\mathrm{C}} + X_i(\beta^{\mathrm{T}} - \beta^{\mathrm{C}})]. \tag{12.15}$$

If we are also willing to assume (more problematically) that $\beta^{\mathrm{T}} = \beta^{\mathrm{C}}$ – the *common-impact model* – then the average treatment effect reduces to $\alpha^{\mathrm{T}} - \alpha^{\mathrm{C}}$.

12.3.3 Randomization Evaluated

Randomized experiments have been called "the gold standard for scientific experimentation" (Murray 2005, p. 17; Rubin and Waterman 2006, p. 210), but this overstates the case. Often the most serious problem with randomized experiments is that the withholding of treatment may be unethical. For instance, if we are trying to determine the effects of providing Vitamin A supplementation, which helps prevent blindness, it is likely to be unethical to withhold this very inexpensive treatment from significant numbers of young children. And once the treatment is applied universally there is no control group.

---

[2] Note that $Y_i^{\mathrm{T}} = Y_i I_{T_i=1}$ where I_{A} denotes the indicator function of an event (1 if A occurs, 0 if not). We also have $Y_i^{\mathrm{C}} = Y_i I_{T_i=0}$.

[3] We have that $G^{\mathrm{ATE}} = E(Y_i I_{T_i=1}) - E(Y_i I_{T_i=0})$ since $G^{\mathrm{ATE}} = E((Y_i I_{T_i=1} - Y_i I_{T_i=0})I_{T_i=1} + E((Y_i I_{T_i=1} - Y_i I_{T_i=0})I_{T_i=0}$ by the definition of conditional expectations.

The counterargument is that, when resources are limited, random assignment of treatment is fair and, moreover, it is also efficient inasmuch as it allows one to generate information that might steer more resources to a worthy program than would otherwise flow there.

In practice, it is often politically difficult to provide a treatment to one group and not to another. A recent proposal to provide $100 gifts to a random sample of Vietnamese households, with the eventual purpose of estimating a pure income effect on expenditure, was turned down by the national statistics office because it was considered to be invidious.

True random assignment is often difficult in practice, since it is rare to find an up-to-date and definitive list of all households (or individuals) in the population of interest. Moreover, even if eligibility for a program is assigned randomly, actual participation may not be, if those who decline to participate are nonrepresentative. For instance, a study of those eligible to borrow microloans from the Grameen Bank in Bangladesh found that those who did not borrow expressed more worry about their ability to repay loans than those who did borrow. Selective compliance of this kind compromises the internal validity of the impact evaluation.

Even when initial participation in a program is random, there is often selective attrition. There can also be selective uptake over time, for instance if people with sick children move to villages where health clinics were (randomly) established, again contaminating the results.

There are additional problems with randomized (or any) experiments, which sometimes limit their ability to yield worthwhile information for impact evaluation. Some experiments have spillover effects; for instance, a project to treat children against worms is likely to help untreated children too, since they are now less likely to come into contact with worms. But this makes it difficult to find a suitable control group. In other cases, projects could not be scaled up without creating macroeconomic effects – for instance, a small-scale job-training project might not affect overall wage rates, while a large-scale one would – in which case the impact as measured on the pilot project would be a poor guide to the impact of the project replicated on a national scale.

In some cases, the results of an experiment may be warped by the Hawthorne ("expectancy") effect, which occurs when the simple fact of being included in an experiment may affect behavior nonrandomly. And social experiments tend to be expensive, with the consequence that they are usually applied to small samples, which in turn makes inference less precise.

While (randomized) social experiments can be useful, they are no panacea for impact evaluation, and in practice most impact assessments have to rely on quasi-experimental methods, also referred to as "observational studies" or "nonexperimental evaluations," to which we now turn.

12.4 Quasi-Experimental Methods

If households are not assigned randomly to an intervention – such as food stamps, or vaccinations, or irrigation water – then those who benefit are unlikely to be typical of the eligible population. There are two main reasons for this. First, there may be nonrandom program placement, of which the researcher may or may not be aware; for instance, an anti-poverty program may be more likely to be set up in poor villages. This is the problem of *unobserved area heterogeneity*. Second, there may be self-selection into program participation; for instance, more-dynamic individuals may be the first to sign up, or program benefits may flow to those who are politically well-connected, or sick people may move to villages that have been equipped with clinics. Such effects are often hard to detect, and give rise to the problem of *unobserved individual and household heterogeneity*.

The presence of these unobservables immediately brings us back to the problem of selection bias. To see why it arises here, let us return to the case of the Thailand Village Fund. Our interest here is in determining whether this microcredit scheme has any impact on individual incomes.

A reasonable place to start would be to collect data on the outcome indicator (expenditure, for instance, given by Y_i), and on individual and household characteristics (X_i), for a sample of individuals that do ($T_i = 1$), and do not ($T_i = 0$), participate in the scheme, and to use this information to estimate a common-impact equation of the following form:

$$Y_i = \alpha^{\mathrm{C}} + (\alpha^{\mathrm{T}} - \alpha^{\mathrm{C}})T_i + X_i\beta + \varepsilon_i, \quad i = 1, ..., n \tag{12.16}$$

This is the *common-impact model*, and may be derived from (12.14) by setting $\beta^{\mathrm{T}} = \beta^{\mathrm{C}}$. At first sight it would appear that the value of the estimated coefficient on participation (i.e., $\alpha^{\mathrm{T}} - \alpha^{\mathrm{C}}$) would measure the impact of the microcredit scheme on income.

Unfortunately this is unlikely to be the case, because program participation is often related to the other individual, household and village variables, some of which may not be observable. For instance, those who borrow money may be better educated, or younger, or live in villages with a loan office, or be more motivated. The degree of individual motivation is an unobservable; but a more motivated individual is more likely to participate in the program (a higher probability of $T_i = 1$) and to benefit more from it (a higher Y_i). This implies that there is a correlation between T_i and ε_i and so leads to a biased estimate of $\alpha^{\mathrm{T}} - \alpha^{\mathrm{C}}$. As a practical matter there will always be unobservables in such circumstances, and so there will always be some selection bias (which may also be thought of as a form of omitted variable bias).

Table 12.3 Treatment effects for Thailand Village Fund borrowing, common-impact equations, 2004

	Coefficient	*t*-statistic	Variables	R^2
1. All data: $N = 34{,}843$	−0.383	−47.42	1	0.060
2. All data: $N = 34{,}843$	−0.213	−30.42	6	0.331
3. All data: $N = 34{,}843$	−0.036	−5.47	28	0.497
4. All data: $N = 34{,}843$	0.023	3.69	103	0.558
5. Common support[a]: $N = 34{,}648$	0.023	3.69	103	0.555
6. P-score 0.2–0.8[b]: $N = 21{,}274$	0.031	4.53	103	0.399

[a]Common support refers to region of common support as determined by propensity score equation
[b]P-score indicates propensity score
Source: Based on data from Thailand Socioeconomic Survey, 2004

This point may be made more forcefully with the help of the numbers in Table 12.3, which are based on the estimation of the common-impact model using data from the Thailand Socioeconomic Survey of 2004. Each row in the table shows the key result from a separate regression, in which the dependent variable is the log of per capita household expenditure. The "coefficient" column reports the estimate of the impact of borrowing from the Village Revolving Fund; it is an estimate of $(\alpha^T - \alpha^C)$ in (12.16), using our terminology. The first regression includes only a dummy variable that indicated whether one borrows or not, and no other variables; in effect this shows that per capita expenditure levels are about 32% (i.e., $1 - e^{-0.382}$) lower for VRF borrowers than for nonborrowers. This comparison would only be a legitimate measure of the impact of borrowing if there were random assignment, which is not the case here.

In the second row of Table 12.3 we re-estimate (12.16), including five additional variables (the X_i), including the age and educational level of the head of the household. This improves the fit of the equation and reduces the measured impact of borrowing. As we add more covariates (rows 3 and 4) the equation fits better, of course, and the measured impact of borrowing changes dramatically. In the fourth equation, which includes dummy variables for each province in Thailand, the measured impact of borrowing appears to be positive – it raises per capita expenditure by about 2.3% – and statistically significant. We still do not know whether this is a correct measure of impact, since there may well be further relevant covariates that we have not observed, or cannot observe, but it is more plausible than the measures in rows 1 through 3.

Rows 5 and 6 of Table 12.3 show the effects of estimating the common-impact model using a subsample of the survey data, in effect retaining only those households that are in the area of "common support"; we define this more carefully below, but the essential idea is to exclude borrowing households whose predicted probability of borrowing is too high for there to be relevant nonborrowing comparators, and to exclude nonborrowing households whose predicted probability of borrowing

is too low for there to be relevant borrowing comparators. The net result is to show a somewhat higher measured impact, on household per capita expenditure, of borrowing from the Village Fund.

At first sight, the presence of selection bias, whether due to unobserved area or household heterogeneity, might appear to doom all efforts to obtain an adequate measure of the impact of a project or program. But there are two possible directions that one can take, if not to solve the problem of selection bias, at least to attenuate it enough to arrive at usable estimates of program impact.

The first tack is to try to make the assumption of nonconfoundedness (or of the weaker assumption of conditional exogeneity of treatment, which says that $E(v_i^{\mathrm{T}} \mid X, T = t) = E(v_i^{\mathrm{C}} \mid X, T = t) = 0,\ t = 0, 1,$ (Ravallion 2008)) more plausible, or at least more palatable. This may be attempted by using matching methods, or by using double or triple differences. An alternative tack is to assume that one can find instrumental variables that affect participation but not the outcome. We address the strengths and weaknesses of each of these approaches below.

12.4.1 Solution 1. Matching Comparisons

Even if treatment (or program participation) has not been assigned randomly, it may be possible to measure the impact of the program by using matched comparisons. In its purest form, the basic idea is to match each participant with an otherwise identical nonparticipant (the comparator) – based on observed pretreatment characteristics – and then to measure the average difference in the outcome variable between the participants and the comparison group.

Units that cannot be matched are usually discarded – this is central to good matching – because they "cannot support causal inferences about missing potential outcomes" (Diamond 2005, p. 9). The hope is that this allows one to mimic the effects of randomization, even though the treatments were not applied randomly in practice. The resulting measure of impact is only compelling to the extent that one believes that the matching has been done well and the treatment assignment is ignorable; in other words, we know that the treatment was not assigned randomly, but we believe that we may proceed as if it were.

The difficult part, of course, is finding the appropriate matches for participants. The ideal would be *exact matching*, which requires that for each unit (household, person) treated, one can find someone who did not receive treatment but who is otherwise identical in every relevant way – for instance, who is also 48 years old, father of two, illiterate, living in a small village, and working in construction. If there are many X_i covariates, or some of them are continuous, then exact matching is all but impossible. The two main solutions, discussed more fully below, are propensity score matching and covariate matching.

12.4.2 Propensity Score Matching

The problem of matching treatment with nontreatment units is much more tractable if we can create a summary measure of similarity in the form of a *propensity score*. Let $p(X_i)$ be the probability that unit i be assigned to the treatment group, conditional on X_i, and define

$$p(X_i) \equiv \Pr(T_i = 1 \mid X_i) = E(T_i \mid X_i). \tag{12.17}$$

This probability of participation – the propensity score – can be estimated using an *assignment model*. Given survey information, the commonest procedure starts by pooling the two samples (i.e., the participants and nonparticipants) and estimating a logit or probit model of program participation as a function of pretreatment variables that might influence participation. Diamond (2005) uses a robust logit estimator that reduces the influence of outliers. Some authors (e.g., Imbens 2004) favor the use of nonparametric binary response models, in order to constrain the assignment model as little as possible. Ironically, if the equation fits too well it is difficult to identify nonparticipants who are otherwise similar to participants.[4]

12.4.2.1 An Illustration

To illustrate, Table 12.4 shows the key elements of a propensity score equation estimated using data from the Thailand Socioeconomic Survey of 2004. In this case the dependent variable is set to 1 if the household had borrowed from the Village Rotating Fund at the time of the survey, so one may also think of this as a participation equation. The estimates are for a probit form and based on information from 34,752 households. Those households that are poor enough to get a subsidized health card, or that have more earners, are more likely to borrow. When the head of the household is older, or more educated, the household is initially more likely to borrow, but the negative coefficients on the squared terms show that these effects are gradually attenuated and eventually go into reverse (at an age of 44, and after 8 years of education).

Every village in Thailand was eligible for an initial injection of one million baht for the VRF, regardless of its population. Thus the probability of borrowing should be inversely proportional to the size of the village, here measured by the number of households. Table 12.4 shows that this effect is strong and highly statistically significant.

[4] To see this, consider an extreme case where all men borrow and no women borrow, so that gender perfectly predicts whether one will borrow. But then it will be impossible to match a borrower with an "otherwise identical" nonborrower.

Table 12.4 Propensity score equation for borrowing from Thailand Village Fund, 2004

			Full sample	VRF borrowers
	Coefficient	*p*-value	Mean	Mean
Does household borrow from VRF? (Yes = 1)				
Age of head (in years)	0.017	0.00	49.7	50.4
Age of head squared (in years'00)	−1.935	0.00	26.9	27.1
Educational level of head (in years)	0.100	0.00	7.1	6.1
Educational level of head squared	−0.006	0.00	69.6	47.2
Number of adult males in household	−0.153	0.00	1.1	1.2
Number of adult females in household	−0.136	0.00	1.3	1.3
Size of household	0.100	0.00	3.5	3.8
Household has 30-baht medical card	0.223	0.00	0.83	0.93
Province 1 (metro Bangkok)	−0.660	0.00		
(other provinces)[a]				
1/(number of households per village, block)	29.810	0.00	0.007	0.008
(other variables)[b]				
Constant	0.395	0.57		
Memo items				
Number of observations	34,752			
Consumption (baht/capita/month)			3,622	2,549
Pseudo R^2	0.190			
Region of common support	0.004–0.985			

[a]There are 76 provinces in Thailand (including Bangkok), and dummy variables were included for all but one of these provinces

[b]Eighteen other variables were included; for details, see Boonperm et al. (2009), Table 2

12.4.2.2 Matching with Propensity Scores

The computation of propensity scores is only the first step in the process. Rosenbaum and Rubin (1983) prove that

$$(Y_i^{\mathrm{T}}, Y_i^{\mathrm{C}}) \perp T_i \mid X_i \quad \Rightarrow \quad (Y_i^{\mathrm{T}}, Y_i^{\mathrm{C}}) \perp T_i \mid p(X_i). \tag{12.18}$$

Put plainly, this implies that conditional independence ("unconfoundedness") extends to the propensity score, so that treatment cases may be matched with comparison cases using just the propensity score rather than the entire set of predetermined covariates X_i. In other words, to find the nonparticipant that is most closely matched to the participant, one only needs to find the nonparticipant with the propensity score closest to that of the participant.

Rosenbaum and Rubin (1983) also show that

$$G^{\mathrm{TT}} = E_{p(X)}[G_i|_{T=1,p(X)}|T_i = 1], \tag{12.19}$$

where $G_i|_{T=1,p(x)}$ is the difference between the treatment outcome Y_i^{T} for treated unit i and the (control) outcome for the nontreated unit closest in propensity score to i.

This says that the average treatment effect (the gain for the treated) may be obtained by computing the expected value of the difference in the outcome variable between each treated household and the perfectly matched comparison household (as matched using the propensity score).

Perfect matching is not possible in reality, so in practice one needs to compute

$$\hat{G}_i \mid_{T=1} = \frac{1}{|N|} \sum_{i \in N} \left(Y_i - \frac{1}{|J_i|} \sum_{j \in J_i} Y_j \right), \tag{12.20}$$

where Y_i is the observed outcome for the ith individual who is treated and J_i is the set of comparators for i, and N is the set of units for which the set of comparators is nonempty (the "common support," discussed in more detail below). The comparators are typically chosen with replacement – which means that they may be used in more than one matching – in which case the bias is lower but the standard error higher than without replacement. The key point here is that unmatched observations are simply dropped; without such pruning there would be little point to the exercise! This effort to create a useful dataset is not a method of estimation per se; it preprocesses the data so that we may draw causal inferences more satisfactorily (Ho et al. 2006, p. 12).

With nearest-neighbor matching one chooses the m closest comparators. It is common to use $m = 1$, but practice varies: some researchers prefer higher values of m (e.g., Abadie and Imbens 2002 favor $m = 4$), and others have used caliper matching (which uses all comparators within a given distance from the treatment), kernel matching, or Gaussian matching (both of which put more weight on closer comparators than those that are more distant). Dehejia and Wahba (2002) argue that the choice of matching mechanism is not as crucial as the proper estimation of the propensity scores, but this is not a settled issue.

In practice, the plausibility of propensity score matching depends on ensuring "common support" and "balancing."

In our example of the Thailand Village Fund, the highest propensity score (i.e., predicted value from the probit equation in Table 12.4) for a nonborrower in Thailand in 2004 was 0.985 while the lowest value for a borrower was 0.004. In between these extremes is the area of *common support*, where it is possible in principle to match borrowers with nonborrowers that have similar propensity scores.

Only in the area of common support is it possible to make comparisons that allow us to make inferences about causality (Rubin and Waterman 2006), so our comparisons need to be confined to this area, and an impact evaluation is not possible unless there is an area of common support (Imbens 2004, p. 7). Identifying the region of common support can be difficult in small samples. This is at the core of the debate about the extent to which nonexperimental methods can identify treatment effects (see for instance Dehejia and Wahba (2002) and Diamond (2005) who try to reproduce the results of an impact evaluation based on the randomized US National Supported Work Program using nonexperimental methods applied to fewer than 500 observations).

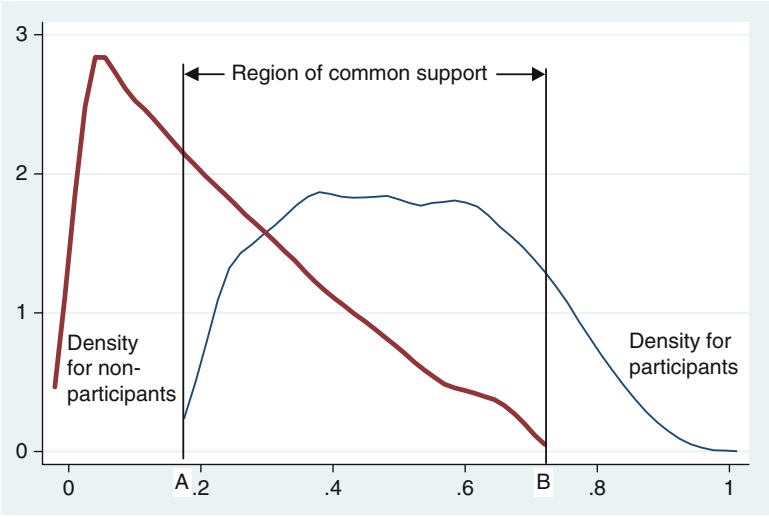

Fig. 12.1 The region of common support. This is the zone where the densities of the propensity scores for participants and nonparticipants overlap

The area of common support occurs where the densities of the estimated propensity scores for participants and for nonparticipants overlap, as shown in Fig. 12.1, which is based on a version of the propensity scores for the Thailand Village Fund study. Above point B there are no comparators for borrowers, so matching is not possible in this zone; below point A there are no borrowers that need to be matched.

In addition, for propensity score matching to work, the treatment and comparison groups need to be "balanced." A treated unit and matched comparator might both have essentially the same propensity scores, but this does not guarantee that they are similar: one household might be agricultural but young while the other might be urban but old. It is not necessary for every individual match to be close, but it is important for the distributions of covariates for the treated and the comparators to be similar, and this is what is meant by balance. More formally, in order to verify balance we need to check whether

$$\hat{p}(X \mid T = 1) = \hat{p}(X \mid T = 0) \tag{12.21}$$

where $\hat{p}$ gives the empirical (rather than population) density of the data.

Theoretically, the true (as opposed to estimated) propensity score ensures balance automatically. Unfortunately, we only have an estimate of the propensity score, and we do not know if the assignment model that generates the propensity scores is in fact the correct one, or even whether it is a consistent estimator of the true propensity scores. Thus we cannot invoke any theoretical results to help guide the choice of assignment model. However, Ho et al. (2006) argue that, paradoxically, we can make use of what they call the "propensity score tautology:" the estimated propensity score

achieves balance when it provides a consistent estimate of the true score, but we only have a consistent estimate of the propensity score when matching balances the X_i covariates. One commonly used algorithm is to estimate a propensity score, match treated with nontreated units, and check for balance; if balance is not achieved, revise the assignment model. Repeat this process until balance is achieved.

It is common to use a formal test for balance. Since our interest is in comparing an entire distribution, one approach is to divide the data into ten or more strata ("blocks"), based on the estimated propensity score, and then use a series of *t*-tests or chi-square tests to check that, within each stratum, the values of each covariate (height, age, hair color) are on average the same. This allows one to "check the adequacy of the statistical model" of assignment (Imbens 2004, p. 18); if balance has not been achieved, the assignment model needs to be revisited and revised.

Ho et al. (2006) are critical of the use of formal statistical tests of balance, on the grounds that balance is a characteristic of the sample, not some underlying population. They also argue that it is likely to be important to achieve balance in some variables (which have large effects on the outcome, *Y*) while it is less essential for others. On the other hand, they also suggest that one compare each covariate between the treatment and comparison samples, and apply a rule of thumb that the difference should be no more than half a standard deviation! They conclude, and here we agree, that "evaluating balance should always be done in multiple ways" (p. 20).

12.4.2.3 Propensity Score Matching Illustrated

Given estimates of the propensity score, standard practice is to simply compare the outcomes of interest (such as expenditure or income) between the treatment group and the matched comparators; any difference between the two may reasonably be inferred to have been caused by the treatment, again provided that we believe that differences between the treated and matched groups are not contaminated by the effects of unobservables. Imbens (2004, p. 16) argues that the simple difference is not unbiased, and the outcomes should first be weighted by the inverse of the propensity scores. A more elaborate procedure is to use a "blocking-on-the-propensity-score" estimator (Rosenbaum and Rubin 1983), which estimates the treatment effect within each "block" and then obtains a weighted average for the overall effect.

An important advantage of these procedures is said to be that they are nonparametric, but of course this is only true conditional on the model used to generate the propensity scores.

At this point it is useful to return to our example of the Thailand Village Revolving Fund (VRF) to illustrate the application of propensity score matching. The estimations of the propensity score equation shown in Table 12.4 generate scores that yielded a wide region of common support (from 0.004 to 0.985).

To check for balance, the propensity scores were divided into 17 bins such that the estimated propensity score within each group was the same for borrowers and

Table 12.5 The effect of Thailand Village Fund borrowing on household income, expenditure and durable assets, using propensity score and covariate matching

	Sample means			Matched comparisons	
	Whole sample	VRF borrowers	Not VRF borrowers	VRF – non-VRF	t
Complete sample, provincial dummy variables					
Ln(expenditure/capita)[a]		2,549	4,286	0.033	2.67
Ln(income/capita)[b]				0.019	1.27
HH has VCR	0.60	0.61	0.60	0.036	4.04
HH has fridge	0.80	0.82	0.78	0.045	6.56
HH has washing machine	0.36	0.33	0.39	0.049	5.36
Rural households, regional dummy variables					
ln(exp/capita): propensity score matching				0.076	4.39
ln(exp/capita): covariate matching				0.013	1.02

[a]Means show levels, not logs

Source: From Boonperm et al. (2009), based on data from Thailand Socioeconomic Survey 2004

nonborrowers. Then within each bin we tested for significant differences in the values of the covariates between the borrowers and nonborrowers. In a total of 43 cases there were significant differences (at the 1% level), which is somewhat more than the 17 cases that might be expected randomly ($= 1\% \times 17$ bins $\times$ 102 variables), but not so far out of line as to make the results implausible.[5]

Once satisfied that the balancing property is met, one measures the impact by taking each borrower, finding the nonborrower with the closest propensity score, and recording their outcome variables such as expenditure per capita. The difference, if any, between the average outcomes for these two matched groups measures the impact. The computation of the propensity score, and the tests for balance, can be done easily enough with the `pscore` command in Stata; and actual matching then uses `attnd` (for single nearest-neighbor matching) or a related command such as `attr` (for caliper or radius matching) or `attk` (for kernel matching).

A selection of results of this technique, applied to the Thailand VRF data for 2004, are shown in Table 12.5, along with the (unmatched) sample mean values of the outcomes. These results show, for instance, that borrower households had an average monthly expenditure per capita of 2,549 baht compared to 4,286 baht for nonborrowers. However, when borrowers are matched using the propensity score, with a single nearest neighbor, their expenditure per capita is 3.3% higher

[5] An earlier version of the model had used regional, rather than provincial, dummy variables in the propensity score equation; when it did not show adequate balance we revised the model, mainly by using the (more numerous) provincial dummy variables.

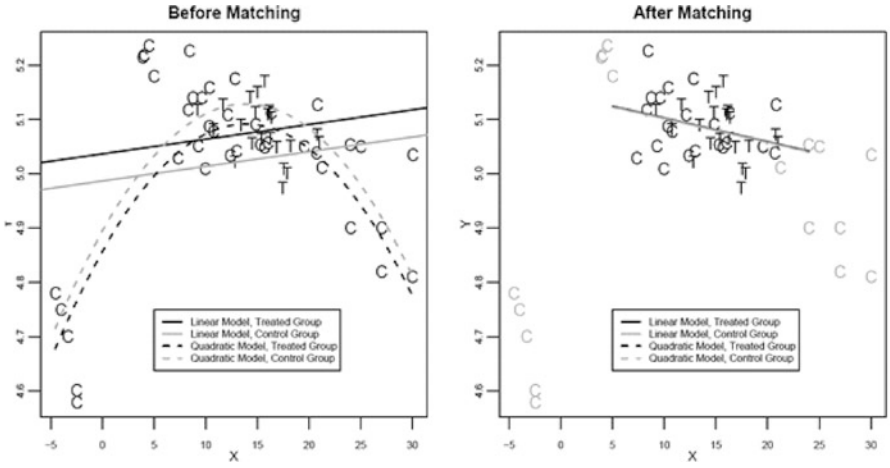

Fig. 12.2 Illustration of the effect of preprocessing, based on the propensity score, on the relationship between outcome *Y* and covariate *X*. The data come from a hypothetical example constructed by Ho et al. (2006); reproduced with permission. Treated units are marked T, comparison units are marked C. The best-fit lines are either *solid* (if linear) or *dashed* (if quadratic), *black* (if based on treated group), or *grey* (for the control group). The *right-hand panel* confines the estimation to observations based on the region of common support

($t = 2.67$) than for nonborrowers. Since we have, in effect, controlled for other observables, this difference may indeed be attributed to the microcredit program.

It is also clear from Table 12.5 that VRF borrowers were on average less likely than a typical household to have a phone or washing machine, but when matched with otherwise similar nonborrowers they had higher assets, which suggests that VRF borrowing enabled them to increase their assets.

Ho et al. (2006) make a strong case that the real value of propensity score matching lies in trimming ("preprocessing") the original dataset so that it is more appropriate for the usual type of parametric analysis. Thus, having matched and balanced the data, one could apply OLS or quantile regression to estimate the effect of the treatment on the outcome of interest, along the lines of (12.16). The results of such an approach are shown in the final two row of Table 12.3 and give results close to those generated by propensity score matching.

The importance of trimming the data is nicely illustrated in Fig. 12.2, which comes from Ho et al. (2006, Fig. 1) and is based on an artificial data set. In each of the two panels the values of the outcome variable (*Y*) are shown on the vertical axis and of a covariate (*X*) on the horizontal axis. Each observation is marked with either a T (for a treated unit) or C (for a comparison unit). The graphs show linear and quadratic fitted curves for treated cases (black) and comparison cases (grey). The left-hand panel fits the curves to all the data, while the right-hand panel fits the curves only to matched data in the zone of common support. The appropriate trimming of this data set leads to a very different conclusion about the relationship between *X* and *Y*.

12.4.2.4 Propensity Score Matching Cases

Propensity score matching has been used in a number of interesting impact evaluations. Jalan and Ravallion (1999) examined the effects of the *Trabajar II* program in Argentina, which was introduced in 1997 in response to a sharp rise in unemployment. The program provided low-wage work on community projects, and was intended to raise the incomes of the poor.

To analyze the impact of this "workfare" program, they used the results of the 1997 *Encuesta de Desarrollo Social* (Social Development Survey), coupled with a similar survey of participants in the Trabajar program. They estimated a logit model of program participation, using variables such as gender, schooling, housing, and subjective perceptions of welfare, and used it to derive propensity scores for participants and nonparticipants. They limited the sample of nonparticipants to those with common support.

Their key findings were that the program raised incomes by about half of the gross wages paid out, and that four-fifths of the participating workers came from the poorest quintile of the population.

The 1997 *Encuesta*, which surveyed 40,000 urban households, has also been used to assess the impact of Argentina's efforts to privatize the provision of water. By comparing data from the *Encuesta* with earlier data from the census, and comparing municipalities where the water supply was, and was not, privatized, Galiani et al. (2005) found that privatization increased access to water by 11.6%. Using data on child deaths, and applying propensity score matching to municipalities (rather than households), they also found that the privatization of water supply reduced child mortality by 6.7% on average, and by 24% in poor municipalities.

12.4.3 Covariate Matching

It is possible to match treated units with otherwise similar untreated units in a way that is more realistic than full matching but does not use propensity scores. One of the simplest forms of *nearest-neighbor matching* first normalizes all of the covariates (household size, age of household head, and so on) so that they have mean zero and unit variance. If each variable is given a weight of one, then one can match any treated unit with the closest untreated unit, where closeness may be defined as the minimum sum of squared differences across all covariates.

More formally, let

$$U_i = \frac{X_i - \bar{X}}{s_i} \tag{12.22}$$

be the normalized $k \times 1$ vector of covariates for unit i. Define the distance between normalized vectors U_i^{T} (for a treated unit) and U_i^{C} (for a comparison unit) to be

$$\left\|U_i^{\mathrm{T}} - U_i^{\mathrm{C}}\right\|_V \equiv \sqrt{(U_i^{\mathrm{T}} - U_i^{\mathrm{C}})' V (U_i^{\mathrm{T}} - U_i^{\mathrm{C}})}, \tag{12.23}$$

where V is a positive definite weight matrix. In the simplest case, V is the identity matrix I_k, which gives equal weight to the distance between each covariate.

This simple matching scheme may be used to check the robustness of the estimates of the impact of the Thailand Village Fund, with the results that are shown on the bottom rows of Table 12.5. It appears that VRF borrowing in rural Thailand raises expenditure per capita by 1.3%, compared to 7.6% using propensity score matching (and regional dummy variables). In this case, the measured effect was smaller with covariate matching and was not statistically significant at conventional levels.

A key issue in the use of direct matching of this type is the appropriate choice of the weight matrix, V. Let Σ_X be the covariance matrix of the covariates. Then it is common to use, as a weight matrix applied to the original (i.e., not normalized) magnitudes, the reciprocals of the variances, which is given by $\operatorname{diag} \Sigma_X^{-1}$. This gives the results shown in Table 12.5. Some researchers prefer use the full inverse of the covariance matrix, Σ_X^{-1}, which gives the Mahalanobis distance, although there is no consensus that this represents an improvement over the simpler distance (Imbens 2004, footnote 6).

Diamond and Sekhon (2005) proposes the use of *genetic matching*. The basic idea is to start with a weight matrix V_0 and to adjust it iteratively until the best possible balance is achieved (Sekhon 2006). As usual, balance is achieved when the covariates (such as household size, age of head, and so on) do not differ, on average, between the treatment group and the sample with which they are matched; Sekhon recommends requiring every p-value associated with a t-test of the difference in the means of covariates to be 0.15 or higher, and likewise with Kolmogorov–Smirnov tests for the distributions of continuous variables. Whether genetic matching represents an improvement is not yet clear, but this is an active area of current research.

An interesting recent example of the creative use of matching may be found in a recent study by Cattaneo et al. (2007) of the impact of the *piso firme* program in Mexico. This is a large program, funded by the federal government but implemented by the states, that provides homeowners with ready-to-pour concrete to replace dirt floors. According to the 2000 census, three million Mexican households had dirt floors; by 2005 the government had provided concrete for over 300,000 of these.

A major justification given for the program is that it reduces the transmission of parasites to children, thereby improving their health, including their cognitive development. Cattaneo et al. set out to test this assertion, by surveying treatment and control households and comparing the outcomes. In doing this they are able to exploit a geographic discontinuity: the twin cities of Gómez Palacio/Lerdo and Torreón face each other across the border between states of Coahuila and Durango, and form part of a single urban area, but as of 2005 the *piso firme* program had only been implemented in Coahuila and not in Durango.

In order to construct a sample of households, Cattaneo et al. used data from the census blocks of the 2000 Census. They first chose census blocks in the treatment area (i.e., Gómez Palacio/Lerdo) and then matched these with similar census blocks in Torreón; similarity is measured by the smallest distance, which in turn is the maximum difference between any of four variables (including household size, and proportion of households with dirt floors). Having identified matching blocks, the researchers then sampled households that owned their house, had lived there since 2000, had at least one dirt floor in 2000, and had at least one child aged under six at the time of the survey. A total of 2,783 households were surveyed, more or less evenly divided between treatment and control areas, and information was collected on sociodemographic variables, anthropometric measurements, cognitive ability (for instance, using the Picture Peabody Vocabulary Test for children aged 36–71 months), and blood and stool quality.

To measure the impact of the program, Cattaneo et al. regressed the relevant outcomes, such as the parasite count in the blood, on the share of the cement floors (CF), and a large number of control variables.[6] Here is a typical finding:

$$\begin{aligned}\text{Parasite count} = -0.371\,CF + \ &\text{other variables}\\ [\text{SE} \ = 0.229]&\end{aligned}$$

The mean parasite count was 0.613, and this measures the number of different parasites found in a child's stool sample. The coefficient is significant at the 10% level, and is large, indicating that cement floors have a substantial effect in reducing parasitic infection in urban areas in Mexico.

In 2000, the proportion of rooms that had cement floors was 33% both for the treatment and the control sample; by 2005 this proportion had risen to 73% for the control sample and almost 100% for the treatment sample. Thus the *piso firme* program increased the proportion of cement floors by 27% (Cattaneo et al. 2007, p. 14). Thus it should be no surprise to find that when one regresses outcomes on a summary variable that measures whether an area has been treated, the result is about a quarter of the effect of a concrete floor. Thus

$$\begin{aligned}\text{Parasite count} = -0.078\ T + \ &\text{other variables},\\ [\text{SE} \ = 0.049]&\end{aligned}$$

where T is set equal to one if the area is covered by the program (i.e., if the household is in Coahuila) and zero otherwise.

[6] This variable was instrumented using a dummy variable that indicated whether the area was covered by the *piso firme* program; the rationale for and use of instrumental variables is discussed further below.

It is not surprising that the installation of concrete floors would reduce infection, but what is less obvious is the effect on the cognitive development of children. Here is another typical result from the Cattaneo et al. study:

$$\text{Picture Peabody Vocabulary Test} = 2.956\,T + \text{ other variables.} \quad [\text{SE } = 1.477]$$

The mean value of the test, which gives a percentile score, was 30.7, so this result indicates that the program raised test scores by about a tenth. This is a large effect, and implies that the program was cost-effective even relative to programs that are aimed more directly at boosting educational performance (Cattaneo et al. 2007, p. 20).

12.4.4 Solution 2. Double Differences

To recapitulate, the problem we are addressing is how to measure the impact of a program ("treatment") when the units of interest (household, individuals) have not been randomly assigned to the program. The matching methods set out in the previous section were designed to reduce bias by selecting comparison units based on observable covariates ("selection on observables"). They are typically implemented in practice by using survey data collected after the program has been operating for some time.

More powerful measures of program impact are possible if we have panel data both from a baseline survey before the intervention occurs and a follow-up survey after the program is operating. Both surveys should be comparable in the questions used and the survey methods applied, and they must be administered both to participants and nonparticipants.

These requirements are not often met. An impact evaluation of, say, a single irrigation project may well be able to draw on baseline information from a national survey, but it is quite possible that the sample of those potentially affected by the irrigation project may be too small to serve as a useful baseline.

Given that we have the data, the simplest version of the double difference estimator consists of first computing the difference between the outcome variable after ($Y_{i,\text{after}}$) and before ($Y_{i,\text{before}}$) the intervention, both for the treatment and comparison samples. The difference between these two differences (the "double difference") gives an estimate of the impact of the program.

Figure 12.3 (from Khandker 2007) helps clarify the situation. When we have random assignment, we may compare the outcome for the beneficiaries ($Y = 30$ in Fig. 12.3) directly with the outcome for those who did not receive treatment ($Y = 17$ in the left-hand panel of Fig. 12.3). However, if the comparison group was not chosen randomly, and its outcome rose from 14 to 21 over time (right-hand panel of Fig. 12.3), then we might plausibly assume that the outcome for the treated group would also have risen by 7 over time, creating a counterfactual

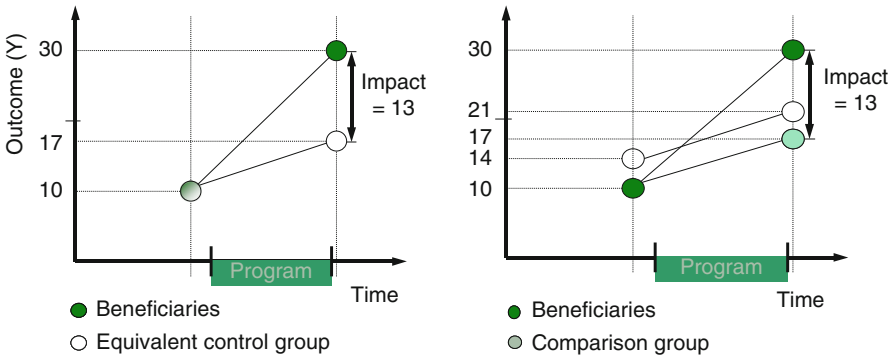

Fig. 12.3 Measuring program impact under randomization (*left*) and using double differences (*right*). The treatment sample starts with an outcome of 10 and finishes with an outcome of 30. With random assignment, the final level may compared with the control group (*left*); with a comparison group a counterfactual must be inferred

output of 17, which may be compared with the observed output of 30. This latter may be computed from the double difference

$$(30 - 10) - (21 - 14) = 13.$$

A fundamental advantage of panel data is that they allow one to eliminate unobserved variable bias, provided that this bias is linear and does not vary over time. These are not trivial conditions, but they are weaker than assuming that bias can be ignored (even conditional on the covariates). Consider the following model:

$$Y_{i,\text{before}} = a + cX_{ib} + \varepsilon_{ib} \tag{12.24}$$

and

$$Y_{i,\text{after}} = a + bT_i + cX_{ia} + \varepsilon_{ia}, \tag{12.25}$$

where the errors consist of a time-invariant component and an innovation error, so for time $t = a,b$ we have

$$\varepsilon_{it} = \eta_i + \mu_{it}. \tag{12.26}$$

With panel data we can take the difference between (12.25) and (12.24), to get

$$Y_{i,\text{after}} - Y_{i,\text{before}} = bT_i + c(X_{ia} - X_{ib}) + \mu_{ia} - \mu_{ib}. \tag{12.27}$$

Our double-difference measure of the impact of the treatment is given by the estimate of coefficient b, and we have swept away the effect of any unobservable

Table 12.6 Double difference estimates of the impact borrowing from the Thailand Village Revolving Fund, panel data for 2002 and 2004

	Expenditure per capita	Income per capita	Farm income per capita	Non-farm income per capita
	Means, 2002, baht per person per month			
Households that borrow from VRF	2,002	2,519	784	362
Households that do not borrow from VRF	2,205	2,984	551	403
	Impacts (in log form)			
Impact	0.020	−0.002	0.020	0.097
t-statistic	1.14	−0.11	0.25	1.36
Number of observations	6,966	6,966	992	2,681

Source: Based on panel component of Thailand Socioeconomic Surveys of 2002 and 2004 (which covers rural areas only). Each observation refers to one adult

(and observable) variables that do not vary over time – such as, for instance, the innate drive or ability of a farmer, or the social capital in a village, or the effects of nonrandom program placement – that would otherwise appear in the η_i terms.

The double difference method may be refined in a number of ways. One hybrid, which appears to perform well, first uses propensity score matching with data from the baseline survey to preprocess the data in order to ensure that the comparison group is similar to the treatment group, and then applies double differences; this helps deal with unobservable heterogeneity in the initial conditions. And more complex specifications of (12.27) could be used, for instance by including the levels of the X covariates (Ravallion 2008).

Double differencing is often very helpful in measuring impact effects, but it will give biased results if there is selective attrition of the treatment group – in other words, if some of the treatment group cannot be resurveyed a second time, and if those who drop out are not a random sample of the treatment group (for instance, if they are older or richer than their peers in the treatment group). The double difference method is typically relatively expensive to implement, inasmuch as it usually requires at least two rounds of survey data.

Returning to our example of microlending by the Thailand Village Revolving Fund, it was possible to apply double differences using data from the Thailand Socioeconomic Surveys of 2002 and 2004. The 2004 survey resurveyed 5,755 rural households, which created the necessary panel data.

Before computing the double differences, we first estimated the propensity scores using the 2002 data – as described above – and then confined the double differencing to the area of common support. We weighted the differences for each treated case (i.e., each adult who borrowed in 2004) by 1, and each comparison case by $p/(1-p)$ – where p is the propensity score – as recommended by Imbens (2004; see also Ravallion 2008). The results are summarized in Table 12.6, and show that the VRF had no statistically significant effect on income or farm

Table 12.7 Log(wage) and education in 1995 by age cohort and program intensity, Sekolah Dasar INPRES program, Indonesia

	Log(wages)			Years of education
	Level of program in region of birth			
	High	Low	Difference	Difference[a]
Aged 2–6 in 1974	6.61 (0.008)	6.73 (0.006)	−0.12 (0.010)	−1.27 (0.057)
Aged 12–17 in 1974	6.87 (0.009)	7.02 (0.007)	−0.15 (0.011)	−1.39 (0.067)
Difference	−0.26 (0.011)	−0.29 (0.010)	0.026 (0.015)	0.12 (0.089)

Source: Duflo (2001)

Notes: Terms in parentheses are standard errors

[a]Difference is in educational level attained by individuals between high- and low-intensity program areas

income; the effects on per capita expenditure and on non-farm income are only marginally stronger, and certainly not compelling.

12.4.4.1 Illustrations: Schools in Indonesia, Subsidies in Mexico

It is sometimes possible to apply double differences even without panel data, particularly where there is a "natural experiment" that has the effect of applying a treatment unexpectedly, or to a well-defined group. A good example is the study by Duflo (2001) of the impact on educational achievement and wages of Indonesia's massive push to build more schools in the mid-1970s under the Sekolah Dasar INPRES program. Between 1973–1974 and 1978–1979, Indonesia built and staffed 61,000 additional three-teacher primary schools, each capable of serving 120 children. This was equivalent to about two additional schools per thousand children of primary school age, and was associated with a rise in the gross primary school enrollment rate from 69% in 1973 to 83% in 1978.

One of the questions addressed by Duflo was whether the program had an impact on wages. Using data from the 1995 Intercensal Survey, she was able to compile information on wages, birthdates, and birthplaces for 60,633 individuals. The first cohort of people in a position to benefit from the Sekolah Dasar program were those aged 2–6 in 1974; on the other hand, anyone aged 12–17 in 1974 was too old to benefit from it. One can also distinguish those regions where the program was pursued with high intensity from those areas where it was less intensive. This allows one to construct a simple difference estimator, along the lines set out in Table 12.7.

Those aged 2–6 in 1974 in low-intensity regions had *higher* wages (as of 1995) than those regions where the Sekolah Dasar program was pursued intensively – not surprisingly, because the program was explicitly structured to build more schools in under-served parts of the country. But this gap in log wages was smaller for the cohort that benefited from the school building program (−0.12) than for those that did not (−0.15), as Table 12.7 shows. The difference between these differences, which comes to 0.026 (i.e., about 2.6%), is a measure of the impact of the program, although it should also be noted that the standard error of this double difference was

Table 12.8 Notation for enrollment rates, Progresa project, Mexico

	Poor households, eligible for Progresa grants	Nonpoor households, not eligible for Progresa grants
Progresa localities	$S_{1,t}$	$S_{3,t}$
Non-Progresa localities	$S_{2,t}$	$S_{4,t}$

Source: Schultz (2001, Fig. 1)

relatively high, at 0.015. It is worth emphasizing that this double difference estimator rests on the identification assumption that any increase in wages would not have differed systematically between regions in the absence of the Sekolah Dasar program; only if this seems plausible can we have much faith in the estimates. Duflo refines her estimates using regressions, but the identification strategy, and basic results, do not differ markedly from those summarized here in Table 12.7.

One of the most widely cited, and useful, studies to use differences to measure the impact of a program is Paul Schultz's analysis of the Progresa subsidies in Mexico. First implemented in 1998, the Progresa program made payments to poor mothers in rural Mexico provided that their children continued to attend school. The payments, which were indexed to inflation, initially varied from 70 pesos per month (about $7) for a child in third grade to 255 pesos per month for a girl in ninth grade. These amounts were substantial: the monthly wage for an adult male day-laborer was about 580 pesos, and for a child worker approximately 380 pesos.

The challenge here is to measure the impact, if any, that this program had on school enrollment rates. It helped that it was possible, at the start of the program, to set up a large-scale social experiment. First, 495 poor localities were identified in rural Mexico. Based on a census of households in these areas, conducted in October 1997, two-thirds of these households were deemed to be poor, and therefore eligible for Progresa grants. In the summer of 1998 the program was introduced in just 314 of the districts, chosen randomly, with a promise that the program would apply to the remaining districts 2 years later. A sample of households in all 495 localities were surveyed on four other occasions through November 1999. Table 12.9 shows that just prior to the program, although 95.1% of those who had completed fifth grade were still at school in the subsequent year, only 57.7% of those who had just completed the primary cycle (i.e., sixth grade) stayed on for the next grade.

In measuring the impact of the program it is useful to refer to the notation used by Schultz, reproduced here in Table 12.8. Let $S_{1,t}$ be the school enrollment rate for poor children in Progresa localities and $S_{2,t}$ the rate in non-Progresa localities. Then $\mathrm{D1}_t \equiv S_{1,t} - S_{2,t}$ is the difference in outcome between the treatment and control samples in time t. When this is compared over time we have the double difference $\mathrm{DD1}_t \equiv \mathrm{D1}_2 - \mathrm{D1}_1$. Measures of these single- and double-differences are shown in Table 12.9 and show that while enrollment rates for poor children did not differ significantly between Progresa and non-Progresa areas prior to the program, after the program was put in place these first differences became substantial, as did the double differences. Take, for instance, the case of those who

Table 12.9 Changes in enrollment rates, Progresa and Non-Progresa households, Mexico, 1997–1999

	Schooling year just completed		
	5	6	7
Proportion enrolled prior to Progresa program	0.951	0.577	0.956
Progresa – non-Progresa localities			
Difference in enrollment rates of poor before program	0.015	0.024	−0.012
p-value	0.129	0.345	0.894
Difference in enrollment rates of poor after program	0.047	0.111	0.013
p-value	0.001	0.002	0.147
Double difference	0.032	0.087	0.025
p-value	0.146	0.004	0.378
Progresa – non-Progresa localities			
Difference in enrollment gap[a] before program	−0.020	0.042	0.014
p-value	0.293	0.023	0.627
Difference in enrollment gap[a] after program	−0.047	−0.035	0.002
p-value	0.003	0.006	0.910
Triple difference	−0.027	−0.077	−0.012
p-value	0.279	0.001	0.738

[a]The enrollment gap is the enrollment rate of nonpoor children minus the enrollment rate of poor children

Source: Schultz (2001, Tables 2–4)

had just completed sixth grade: poor children in the areas with a Progresa program were more likely to be enrolled before the program (a difference of 2.4 percentage points, but not statistically significant) and much more likely to be enrolled after the program (an 11.1% point difference, and significant). The gap thus grew by 8.7 percentage points, an effect that is substantial and statistically significant.

An alternative approach is to measure the *differential* in enrollment rates between poor and nonpoor children (the "enrollment gap"), between Progresa and Non-Progresa localities. This is defined as $D2_t \equiv (S_{3,t} - S_{1,t}) - (S_{4,t} - S_{2,t})$. We would expect D2 to be close to zero before the program is introduced, and to become negative afterwards (as $S_{1,t}$ rises relative to the other terms). This is indeed what was observed, as the numbers in the lower half of Table 12.9 show. One can difference again, creating what is really the triple difference $DD2_t \equiv D2_2 - D2_1$. For the important case of those who have just finished sixth grade we have $DD2_t = -0.077$ and statistically significant; this indicates that the Progresa program increased the proportion of poor children who, having just completed primary school, continued on to lower secondary school, by 7.7 percentage points.

Although the key results of the study emerge clearly from the analysis of differences, Schultz also uses a probit regression to control for a limited number of covariates, but this does not alter the findings. He finds that the expected cumulative effect of the Progresa program is to increase the average length of

time that children stay in school by 0.66 years, a 10% improvement. If viewed purely as a program to increase investment in human capital, Schultz estimates that it yields a real rate of return of about 8%.

12.4.5 Solution 3. Instrumental Variables

To repeat, our interest is in finding unbiased coefficients for the treatment term in the outcome regression, typically specified as in (12.16) (reproduced here):

$$Y_i = \alpha^{\mathrm{C}} + (\alpha^{\mathrm{T}} - \alpha^{\mathrm{C}})T_i + X_i\beta + \varepsilon_i, \quad i = 1, ..., n \tag{12.16}$$

The problem is that for one of a number of reasons – an omitted, mis-measured, or endogenous explanatory variable – T_i may be correlated with ε_i. For instance, a dynamic individual might be more likely to participate in a program (a high T_i) and to benefit from it ($\varepsilon_i > 0$). Murray (2005) refers to T_i in this context as the "troublesome explanator;" without further adjustments, OLS estimation of (12.16) will yield biased estimates of the impact coefficients.

The idea behind instrumental variables (IV) estimation – also sometimes referred to as the statistical control method – is to try to find variables Z that are correlated with T_i but not with ε_i. In the jargon of instrumental variables, we ideally want strong instruments Z that have high "instrument relevance" so that they are closely correlated with T_i, but at the same time satisfy the "exclusion restriction" (or "instrument exogeneity") so that they play no direct role in the outcome regression (thus $\mathrm{cov}(Z_k,\varepsilon) = 0$ for all k instruments).

Given such instruments, it is common to estimate, as a first stage, a separate *participation equation* (or "assignment equation") of the form

$$T_i = Z_i\gamma + X_i\varphi + u_i. \tag{12.28}$$

and then to use the estimated values of participation from (12.28) (i.e., $\hat{T}_i$) instead of T_i in estimating (12.16). In practice, (12.28) is typically estimated in logit or probit form, given the binary nature of the dependent variable.

To see why this technique works, return to the case of a dynamic individual who is both more likely to participate (so $u_i > 0$) and more likely to benefit from the program ($\varepsilon_i > 0$). By using $\hat{T}_i$ instead of T_i, the forces that influence ε_i and T_i now only affect ε_i, but not $\hat{T}_i$, so the correlation disappears, along with the bias. However, this is only true if there are influences on T_i that do not influence Y_i. The idea is to create variation in $\hat{T}_i$ so that we have some people in the sample who, even if they have the same X_i, may have different T_i; in effect we now have a source of variation in Y_i that is attributable to the program.

The major practical problem is finding appropriate instruments. They must influence program participation while somehow not influencing the outcome of the program once one is enrolled. This is difficult. In a useful review of IV

estimation, Murray (2005, p. 18) writes, "all instruments arrive on the scene with a dark cloud of invalidity hanging overhead," and states, correctly, that "the credibility of IV estimates rests on the arguments offered for the instruments' validity" (p. 11).

It is not possible to test formally for unconditional instrument exogeneity – i.e., to test whether $\text{cov}(Z_k,\varepsilon) = 0$ – because ε is not known. Thus the case must be made using intuition, economic theory, and logical reasoning.

However, if there is more than one available instrument, it is possible to test, provided that one is already using a given instrument, whether additional instruments are justified (essentially by adding them to the instrumental version of (12.16) and testing for their statistical significance).

It is now considered good practice in applied work to report the results of the first state estimation (12.28), which allows the reader to judge whether the estimates look reasonable. A low value of a test of the joint significance of the coefficients of the instruments in this equation would signal weak instruments, and these in turn compromise the ability of IV to improve on the bias inherent in OLS estimation (Murray 2005; he also discusses the Stock–Yogo test for weak instruments and provides some critical values).

Instrumental variables estimation is widely used by economists, including in impact evaluations, and researchers have been imaginative in their search for suitable instruments. To take just one example: a recent study of the effect of famine relief on child growth in Ethiopia was able to use past climatic variation as an instrument in a model of the impact of the relief (Yamano et al. 2003).

The instrumental variables method is especially helpful if there is measurement error. Suppose that, because of measurement errors, observed program participation is more variable than true participation; this will lead to an underestimation of the impact of the program ("attenuation bias"), essentially because the noise of measurement error is getting in the way of isolating the effects of program participation. However, the predicted value of program intervention ($\hat{P}_{i1}$) is less likely to reflect measurement error, and can reduce the effects of attenuation bias.

12.4.5.1 An Illustration: Thai Microcredit

To illustrate the application of the instrumental variables approach we return again to our example of the Thailand Village Revolving Credit Fund (VRF). A feature of the VRF is that it provided a million baht to each Village Rotating Fund, regardless of the size of the village. Thus the probability of obtaining a VRF loan is approximately in inverse proportion to the size of the village ("nhinv"). Our measure of the size of the village is the number of households, which is likely to be closely correlated with the theoretically ideal measure (the number of people eligible for VRF loans, which is the number of adults aged 20 and above).

Some instrumental variables estimates of the impact of the VRF are summarized in Table 12.10. In each case the first-step equation is probit and the instruments are highly statistically significant at that point. The second-stage

Table 12.10 Instrumental variables estimates of the impact of borrowing from the Thailand Village Revolving Fund, 2004

	Expenditure per capita	Income per capita
2004 data		
	Mean, baht per household per month	
Borrowed from VRF in 2004	2,549	3,209
Did not borrow from VRF in 2004	4,286	6,088
	Impact (in log form)	
Instrument: nhinv[a]	0.016	0.017
z-statistic	0.36	0.33
Instruments: nhinv, anydebt[b]	0.196	0.163
z-statistic	15.6	10.8
Instruments: nhinv, non-VRF debt[b]	0.464	
z-statistic	24.2	
Panel Data		
	Mean, baht per household per month	
Borrow from VRF in 2004 only	2,376	3,179
Borrow from VRF in neither 2002 nor 2004	2,632	3,413
	Impact (in log form)	
VRF borrowing in 2004 vs. no VRF borrowing		
Instrument: nhinv[a], household fixed effects	0.179	0.152
z-statistic	3.19	2.24

Source: Based on data from the Thailand Socioeconomic Surveys of 2002 and 2004
Notes: nhinv is the inverse of the number of households per village.; anydebt is equal to 1 if a household has debt from any source, and to zero otherwise; non-VRF debt is 1 if a household has debt from any source other than the VRF, and is otherwise zero
[a]Used two-step estimator because maximum likelihood estimator did not converge
[b]Used maximum likelihood estimator

equation is linear. When the only instrument is *nhinv*, the measured impact is not statistically significant. However, when one also uses additional instruments – whether a household has any outstanding debt ("anydebt"), or whether it has any debt other than from the VRF ("non-VRF debt") – the measured impact becomes large and statistically significant. This underlines the general point that instrumental variables estimates are often highly sensitive to the choice of instruments, which can be disconcerting.

The bottom panel of Table 12.10 shows the results of applying the instrumental variables approach to the panel data that are available for some rural areas. The sample is confined to those households who either borrowed from the VRF both in 2002 and 2004, or borrowed in neither year. The second-stage equation uses household fixed effects – equivalent to a separate intercept for each household – which in principle should sweep away the effects even of unobserved differences between households ("household-level heterogeneity") provided that such effects do not vary over time. These estimates show the VRF having a large effect both on incomes and on expenditures in rural areas.

12.4.6 Other Solutions

Although matching, double differencing and instrumental variables estimation are the most widely used techniques in impact evaluation, a number of other techniques have been used, with more or less success.

Reflexive comparisons. In this approach, one first undertakes a baseline survey of the treatment group before the intervention, with a follow-up survey afterwards. The impact of the intervention is measured by comparing the before and after data; in effect the baseline provides the comparison group.

Such comparisons are rarely satisfactory. The problem in this case is that we really want a "with" and "without" comparison, not a "before" and "after." Put another way, in the reflexive comparison method there is no proper counterfactual against which the outcomes of the project may be compared. There is also a problem if attrition occurs, so that some of those surveyed before the project drop out in some systematic way. On the other hand, this may be the only option in trying to determine the impact of full-coverage interventions, such as universal vaccinations, where there is no possibility of a comparison or control group.

Discontinuity designs. The idea here is to make use of a sharp structural discontinuity in the data that is not caused by the outcome of interest. The approach is typically applied to time-series data – for instance, to measure the effect of a profit announcement on the market valuation of a firm. In this case, the data need to be available for small time units (e.g., weeks, days), and one typically restricts the data sample to a small neighborhood of the discontinuity – in this example, the period just before and immediately after the major exogenous event. The impact is then typically measured using a regression equation with appropriate dummy variables.

The approach can be applied in other contexts. For instance, Esther Duflo (2000) wanted to measure the effect of newly expanded old age pensions on child height and weight in the Republic of South Africa. She used the fact that men are eligible for a pension at 65, and women at 60, to compare the stature of children in households with members slightly below and slightly above pensionable age. She found that pensions received by women had a measurable positive effect on the anthropometric status of girls, but not boys; and pensions received by men had no such effects.

Qualitative methods. Some evaluations rely largely on qualitative information, which comes from focus groups, unstructured interviews, survey data on perceptions, and a variety of other sources. Such information complements, but does not supplant, the more quantitative impact evaluations, because qualitative methods are based on subjective evaluations, do not generate a control or comparison group, and lack statistical robustness.

12.5 Impact Evaluation: Macro Projects

It is much harder to evaluate the impact of an economy-wide shock (e.g., a devaluation) or macroeconomic policy change (e.g., increase in the money supply) than a project or program change, because the universal nature of the change makes it almost impossible to construct an appropriate counterfactual. But this challenge has not deterred researchers, so to finish this chapter we summarize some of the techniques that have been applied to this problem.

12.5.1 Time-Series Data Analysis: Deviations from Trend

One of the simplest, and commonest, methods used to measure the assumed effect of a shock is to use time-series data to extrapolate the outcome of interest – GDP growth, for instance – to create a counterfactual, and then to compare the actual outcome with this counterfactual. This is the approach taken by Kakwani and his co-authors in estimating the effects of the Asian financial crisis of 1997 on poverty and other indicators in South Korea and Thailand.

The first difficulty with this method is arriving at a robust counterfactual; for instance, how far back in time should one go when developing an equation that is used for the projections. And the second problem is linking the shock or program with the observed deviation from trend, because there are likely to be many other potentially plausible explanations. A good illustration of this is Datt and Hoogeveen's paper on the post-1997 slowdown of economic growth in the Philippines, which is titled "El Niño or El Peso?" Many observers claimed that the Asian financial crisis was largely to blame ("El Peso"), but they argued that more probably the slowdown was mainly due to drought ("El Niño") (Datt and Hoogeveen 1999).

12.5.2 CGE and Simulation Models

A computable general equilibrium (CGE) model of an economy is a set of equations that aims to quantify the main inter-relationships between households, firms and government in an economy. CGE models range from just a few to many hundreds of equations. In principle they may be used to simulate the effects of many types of policy interventions. Unfortunately, CGE models are technically difficult to build, are typically highly aggregated (which makes it difficult to identify the effects of policies on income distribution and poverty with much precision), require considerable data to construct the underlying social accounting matrix, and produce results that are sensitive to the assumptions made about the parameters. However, they have been used with some success to evaluate

the economic and distributional effects of such interventions as programs to reduce HIV/AIDS, change food subsidies, alter taxes, or liberalize trade. The International Food Policy Research Institute (IFPRI) has developed a standard CGE model that has been applied with some success to a number of problems in developing countries (Loefgren et al. 2001), and has a comparatively short learning curve.

Heckman et al. (1998) argue that the incorporation of general equilibrium effects can greatly alter the conclusions of an impact evaluation. A partial-equilibrium analysis shows that a $500 per student college tuition subsidy in the USA would be expected to raise university attendance by 5.3%. However, this assumes that the relative wages of graduates to nongraduates remains unchanged, an assumption that is implausible if the subsidy were introduced nationwide. Using a general equilibrium model, Heckman et al. estimate that the $500 subsidy would raise university enrollment by just 0.46%. The smaller effect arises because rising university enrollments would lower the wages of graduates relative to nongraduates, thereby depressing the incentive to attend university, and offsetting the tuition subsidy to a substantial extent.

12.5.3 Household Panel Impact Analysis

If we have panel data on households then we could compare the situation of each household before and after the shock. By including household fixed effects in our estimating equation – equivalent to a separate dummy variable for each household – we can largely eliminate the effects of time-invariant household and area-specific heterogeneity (i.e., of the special or unique features of households, many of which are unobservable – such as whether the head is an alcoholic, or sick, or entrepreneurially inclined).

Again, the main difficulty here is that a before-and-after comparison does not establish an adequate counterfactual. For instance, if the income of a household in the Philippines fell between 1996 and 1998, how do we know that it was due to the 1997 financial crisis? It might have been caused by some other event; perhaps a family member fell ill, or the village suffered from a drought. No survey is ever complete enough to capture every conceivable relevant explanatory variable. Moreover, most household-level economic magnitudes (income, expenditure, even assets) do not follow regular or highly predictable trends from year to year.

12.5.4 Self-Rated Retrospective Evaluation

Another possibility is to ask the household to assess how much it has been affected by the crisis – as was done, for instance, in the APIS survey in the Philippines in 1998.

By definition, self-rated evaluations are subjective, which makes it difficult to assess whether the reported effects are indeed due to the shocks. In Vietnam, households reported higher levels of illness in 1998 than in 1993, despite being much better off in 1998; this is hardly plausible, unless one supposes that the definition of "illness" changes over time or with affluence. Whatever the reason, it makes the subjective evaluations untrustworthy. It is also optimistic to expect that most households have a clear enough grasp of the forces buffeting them to be able themselves to diagnose the root causes of variations in their incomes or expenditure.

A variant on this theme is to ask households whether they were hit by a shock. We then compare the situation of households that reported being affected with those that did not. Since self-reported shocks are highly endogenous – any household that has had a spell of bad luck is likely to report being hit by a shock – researchers often use the shock reported by a cluster (e.g., the village, the city ward) as an instrumental variable, to help resolve this endogeneity.

Even with this latter adjustment, we are left with the problem of unobserved community-level heterogeneity – for instance, for reasons that may not be apparent, some communities or clusters may report a shock more than others, even if objectively the shock hit all areas equally.

12.6 In Conclusion

Three simple points about impact evaluation are worth emphasizing. First, no method of impact evaluation is perfect, even randomization (although it can be helpful). The method used will depend on the problem, and the resources and time available, but will always face the problem of unobservables and hence the need to conjure up and rationalize a satisfactory counterfactual. Constructing a compelling impact evaluation is as much an art as a science; good econometric practice is helpful, but does not substitute for sensible, logical explanation.

Second, impact evaluation is more difficult with economy-wide policy interventions and crises than with micropolicies.

And third, program impact evaluation is important. It serves as a tool for learning whether and how programs matter, and has had a marked effect on public policy in a number of cases; Bamberger (2005) gives some interesting examples. Agencies such as the World Bank earmark as much as 1% of project funds for monitoring and evaluation, and increasingly, impact evaluations are being required in the name of accountability. This may be a passing fad; Ravallion (2008) doubts that impact evaluations will ever suffice for "informing future development projects and policies," because they are so dependent on the specific context of the programs whose impacts they are measuring. This may be unduly pessimistic: the impact of impact evaluations could often be enhanced by paying attention to creating adequate feedback mechanisms, so that policy makers do take the lessons to heart.

References

Abadie, Alberto, and Guido Imbens. 2002. Simple and bias-corrected matching estimators for average treatment effects. NBER Technical Working Paper No. 283.

Angrist, Joshua, Eric Bettinger, Erik Bloom, Elizabeth King, and Michael Kremer. 2002. Vouchers for private schooling in Colombia: Evidence from a randomized natural experiment. *American Economic Review* 92(5): 1535–1558.

Baker, Judy. 2000. *Evaluating the impact of development projects on poverty: A handbook for practitioners*. Washington, DC: World Bank [A useful handbook, with extensive examples.].

Bamberger, Michael. 2005. "Influential evaluations," presentation to the Monitoring and Evaluation Thematic Group, April 26. Washington, DC: World Bank.

Boonperm, Jirawan, Jonathan Haughton, and Shahid Khandker. 2009. Does the village fund matter in Thailand? Policy Research Working Paper 5011. Washington, DC: World Bank.

Cattaneo, Matias D., Sebastian Galiani, Paul J. Gertler, Sebastian Martinez, and Rocio Titiunik. 2007. Housing, health and happiness. Policy Research Working Paper 4214. Washington, DC: World Bank.

Datt, Gaurav, and J.G.M. Hoogeveen. 1999. "El Niño or El Peso? Crisis, poverty and income distribution in the Philippines." Policy Research Working Paper No. 2466. Washington, DC, World Bank.

Dehejia, Rajeev, and Sadek Wahba. 2002. Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and Statistics* 84(1): 151–161.

Diamond, Alexis. 2005. Reliable estimation of average and quantile causal effects in non-experimental settings. Working draft, Harvard University, Cambridge, MA.

Diamond, Alexis, and Jasjeet Sekhon. 2005. Genetic matching for estimating causal effects. Harvard University and University of California Berkeley.

Duflo, Esther. 2001. Schooling and labor market consequences of school construction in Indonesia: Evidence from an unusual policy experiment. *American Economic Review* 91(4): 795–813.

Duflo, Esther. 2000. Grandmothers and granddaughters: Old age pension and intra-household allocation in South Africa. MIT.

Galiani, Sebastian, Paul Gertler, and Ernesto Schargrodsky. 2005. Water for life: The impact of the privatization of water services on child mortality. *Journal of Political Economy* 113: 83–120.

Glewwe, Paul, Michael Kremer, Sylvie Moulin, and Eric Zitzewitz. 2000. Flip charts in Kenya. NBER Working Paper 8018, Cambridge, MA.

Heckman, James J., Lance Lochner, and Christopher Taber. 1998. *General-equilibrium treatment effects: A study of tuition policy*, 381–386. May: American Economic Review.

Ho, Daniel, Kosuke Imai, Gary King, and Elizabeth Stewart. 2006. Matching as non-parametric preprocessing for reducing model dependence in parametric causal inference. http://gking.harvard.edu/files/matchp.pdf [An excellent and up-to-date guide for practitioners.]

Imbens, Guido. 2004. Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics* 86(1): 4–29 [An essential reference for anyone planning to use propensity score matching.].

Jalan, Jyotsna, and Martin Ravallion. 1999. Income gains from workfare and their distribution. Policy Research Working Paper. Washington, DC: World Bank.

Khandker, Shahidur. 2007. Program impact evaluation, PowerPoint presentation. Washington, DC: World Bank.

Khandker, Shahidur, Gayatri Koolwal, and Hussain Samad. 2010. *Handbook on impact evaluation*, World Bank, Washington DC.

Loefgren, H, R.L. Harris, and S. Robinson. 2001. A standard computable general equilibrium model in GAMS. TMD Discussion Paper No. 75, International Food Policy Research Institute, Washington, DC.

Murray, Michael. 2005. The bad, the weak, and the ugly: Avoiding the pitfalls of instrumental variables estimation. Bates College. October.

Ravallion, Martin. 1999. The mystery of the vanishing benefits: Ms Speedy Analyst's introduction to evaluation. Policy Research Working Paper 2153. Washington, DC: World Bank. [A witty and accessible introduction to some of the finer points of impact evaluation.]

Ravallion, Martin. 2008. Evaluating anti-poverty programs. In *Handbook of development economics*, vol. 4, ed. Evenson Robert and T. Paul Schultz. Amsterdam: North Holland 3787–3846.

Rosenbaum, P., and D. Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70(1): 41–55.

Rubin, Donald, and Richard Waterman. 2006. Estimating the causal effects of marketing interventions using propensity score methodology. *Statistical Science* 21(2): 206–222.

Schultz, T. Paul. 2001. School subsidies for the poor: Evaluating the Mexican PROGRESA Poverty Program. Economic Growth Center Discussion Paper No. 834. Yale University, New Haven.

Sekhon, Jasjeet. 2006. Multivariate and propensity score matching software for causal inference. http://sekhon.berkeley.edu/matching Accessed on August 1, 2011.

Yamano, Takashi, Harold Alderman, and Luc Christiaensen. 2003. Child growth, shocks and food aid in rural Ethiopia. Policy Research Working Paper No. 3128. Washington, DC: World Bank.

A growing number of impact evaluations are now available on the Web, and can serve as templates for new evaluations; for a useful list, see http://www.worldbank.org/poverty (and follow links Impact Evaluation and then Selected Evaluations).