IDB
Development Effectiveness

Impact-Evaluation Guidelines
Technical Notes
No. IDB-TN-136
May 2010

Program Evaluation and Spillover Effects

Manuela Angelucci
Vincenzo Di Maro

Office of Strategic Planning and Development Effectiveness, Inter-American Development Bank

Program Evaluation and Spillover Effects

Impact-Evaluation Guidelines

Manuela Angelucci
Vincenzo Di Maro

Inter-American Development Bank
2010

Manuela Angelucci. University of Arizona. angelecm@eller.arizona.edu

Vincenzo Di Maro. Inter-American Development Bank. vincenzod@iadb.org

Program Evaluation and Spillover Effects

Abstract

Manuela Angelucci* and Vincenzo Di Maro**

This note defines what spillover effects are, why it is important to measure them, and how to design a field experiment that will enable researchers to measure the average effects of the treatment in the presence of spillover effects on subjects both eligible and ineligible for the program. In addition, it discusses how to use nonexperimental methods for estimating spillover effects when the experimental design is not a viable option. Several practical examples are provided to show how spillover effects can be estimated.

Evaluations that account for spillover effects should be designed in such a way that they explain both the cause of these effects and who is affected by them. Failure to have such an evaluation design can result in wrong policy recommendations and in the neglect of important mechanisms through which the program operates. To estimate the direct and indirect effect of a program, one has to use control groups that are not affected by the program either directly or indirectly. This often means selecting the control groups from different geographic units (e.g. the village or school). In order to understand the mechanisms that cause spillover effects one has to think about competing explanations and collect data on relevant outcomes. In many cases, unveiling the mechanisms behind the spillover effects results in a better understanding of how the program works in general.

JEL Classification: **C93, C81, D62**

Keywords: Impact Evaluation, Spillover Effects, Field Experiments, Data Collection, Indirect Treatment Effect, Program Mechanisms

* Assistant Professor, University of Arizona and Research Fellow, IZA. Correspondence address: University of Arizona, Economics Department, 1130 E. Helen St, McClelland Hall 401, Tucson, AZ, 85721.
E-mail: **angelecm@eller.arizona.edu**. Phone: +1 520-621-4281

** Research Fellow, Office of Strategic Planning and Development Effectiveness, Inter-American Development Bank. Correspondence address: IADB, Mail stop E0805, 1300 New York Ave NW, Washington, Dc, 20577.
E-mail: **vincenzod@iadb.org**. Phone +1 202-623-2126

We would like to thank Paul Winters for his extremely useful comments. We also thank Sarah Strickland and Jorge Olave for their help with the editing. All remaining errors are our own.

Table of Contents

1. Introduction

This note defines what spillover effects are, why it is important to measure them, and how to design an evaluation that will enable researchers to measure the average effects of the treatment in the presence of spillover effects on subjects both eligible and ineligible for the program.

Welfare-enhancing interventions often have a specific target population. For example, conditional cash transfer (CCT) programs in Latin America and beyond target the poor; other interventions provide incentives to increase schooling for indigent children, from providing equipment to distributing deworming drugs to infected children. However, the target population is often a subset of the "local economy," intended in a loose sense as the geographic unit or the local institution within which the target population lives and operates. In this sense, we can consider the village, the neighborhood, the city, the municipality, the school district, the church, or the extended family, as the relevant local economy.

The local nontarget population may also be indirectly affected by the treatment through social and economic interaction with the target population. For example, the recipients of CCTs may share resources with or purchase goods and services from ineligible households who live in the locality, as well as affecting the incentives for them to accumulate human capital. Children who receive free textbooks and computers may share them with "untreated" children (such as relatives, friends, and members of the local church), raising enrollment levels in both groups. Supplying deworming drugs to a group of children may benefit untreated children by reducing disease transmission, thus lowering infection rates for both groups. If parasites affect school performance (e.g. by reducing emetic iron, causing weakness and inattention) both groups of children may end up learning more.

Examples of other types of interventions are no less important. For example, an intervention which improves the supply and quality of water to only some beneficiary households in a neighborhood is likely to have effects on all the residents of the same neighborhood, ranging from an increase in property values to a decrease in infection rates. In addition, if the intervention includes the provision of hygiene information and education campaigns, we might expect that information to flow from person to person within the community.

Agricultural interventions may be particularly susceptible to spillover effects. For instance, an intervention to help beneficiary farmers switch to nongenetically modified (GM)

crops must take into account the possibility that neighboring farmers may continue to plant GM crops, potentially contaminating the quality of the seeds produced by non-GM farmers. In another example, an intervention in Peru is helping beneficiary farmers to eradicate the *Ceratitis capitata*, a fruit fly capable of wreaking extensive damage to a wide range of fruit crops. Eradication of the fly from one farmer's land would not prevent flies moving to the neighbor's land. This is of relevance both for the program and the evaluation design.

Since these programs end up affecting members of the local economy that are not direct beneficiaries, one should account for these spillover effects in the design of field experiments to evaluate programs. For example, to study the effect of a deworming drug on school performance, we would randomly select half the pupils of a school and treat them with the drug. We would then compare the grades of treated and untreated children. Since the drug is likely to have decreased infection rates among both treated and untreated children, it is possible that both groups' performance may change. That is, suppose deworming children increases the average grades of treated pupils by 10 percentage points and of untreated pupils by 2 percentage points. The Average Treatment on the Treated effect[1] is 10 percentage points. However, if we simply compare the grades of treatment and control pupils, we will observe only an 8 percentage point increase. In sum, the failure to recognize the possibility that the drug may affect untreated pupils also and to design the experiment accordingly, will result in a double underestimate of the treatment's effectiveness. Not only will its effect on the treated be underestimated but its effect on the untreated will go unmeasured. This may result in the wrong policy conclusion (e.g. to discontinue a program because it is not cost-effective).

In some cases, spillover effects are intended. For example, agricultural extension programs encourage participants to adopt a certain technology and hope that this will induce further adoption within the community or in neighboring communities. Immunization campaigns among high-risk populations lower the likelihood of contagion, reducing the infection rate among low-risk populations. Whether intentionally or not, nonparticipants can be affected by programs, and these spillover effects should be taken into account when conducting an impact evaluation. This can be done in two ways. At the very least, the evaluation design must select a control group that is not indirectly affected by the program. This will enable researchers to

[1]The Average Treatment on the Treated effect can be defined as the impact of the program on subjects that are eligible for it. In our case this is the effect of the deworming intervention on the treated pupils.

measure the program effect on eligible subjects. If possible and relevant, the design should also allow for the measurement of spillover effects, as these effects are often important policy parameters in themselves.

In order to design an evaluation that accounts for the presence of spillover effects, one needs to understand why and how these effects occur. This knowledge would help researchers understand which subset of nonparticipants is most likely to be indirectly affected by a particular treatment. For example, in the case of a deworming drug for children, one has to know how people become infected (by being in contact with contaminated matter) to understand which other people are likely to benefit from the deworming (the friends and families of dewormed children, who will be less likely to come in contact with contaminated matter). Importantly, the evaluation must be carefully designed to account for spillover effects before the program is implemented. Spillover effects cannot be detected accurately ex post unless the design considers their existence from the start.

The following section describes different types of spillover effects. Section three explains in more detail why it is important to have an evaluation design that accounts for the presence of these spillover effects. Section four explains how to design an experiment in the presence of spillover effects. Section five discusses how to use nonexperimental methods when an experiment is not a viable option. Section six provides a specific example using the evaluation of the CCT program PROGRESA. Section seven concludes. Some more technical issues are discussed in the appendix.

2. Types of Spillover Effects

In order to discuss different types of spillover effects, we first need to clarify what we mean in this paper when we talk about spillover effects.

As mentioned above, interventions often have a specific target population, commonly only a subset of the larger relevant geographical or institutional unit (the "local economy"). This is the group one looks at to measure the direct effect of the program. For instance, an intervention might target only poor children (the target group) within a locality (the local economy). In many cases, the local nontarget population may also be indirectly affected by the treatment through social and economic interaction with the treated. These possible interactions are what we define as spillover effects in this paper. Given that we are not putting any restrictions on how these interactions work, it would be impossible to review and categorize the variety of ways in which these interactions can operate.

In what follows, we describe four types of spillover effects that are particularly relevant in the development and health economics literature. We call these four types of spillover effects (1) externalities, (2) general equilibrium effects, (3) interactions, and (4) behavioral effects.[2]

Externalities. These effects operate from the treated subjects to the untreated population. A particularly relevant domain is health. Miguel and Kremer (2004) show that deworming drugs can have an additional indirect effect, in addition to the direct effect of ingesting the drug. Supplying deworming drugs to a group of children may benefit untreated children by reducing disease transmission, lowering infection rates for both groups. If parasites affect school performance (e.g. by reducing emetic iron, causing weakness and inattention), both groups of children may end up learning more. Since these programs end up affecting members of the local economy that are not direct beneficiaries, these spillover effects should be taken into account in the design of field experiments to evaluate programs.

General equilibrium effects. These are the effects that an intervention, which targets only part of the 'local economy," can have on the entire population. Active labor market policies or any intervention that can affect equilibrium prices are examples of these types of spillover effects. Angelucci and De Giorgi (2009) discuss how conditional cash transfer (CCT) programs may affect prices, as the program transfers increase demand for goods and services, potentially increasing their prices. It should be noted that while general equilibrium effects stem from the

[2]These labels are somewhat arbitrary, but are a useful way of grouping similar types of spillover effects.

intervention on treated subjects, they can have effects on the total population: both treated and nontreated groups.

Interactions. The local nontarget population may also be indirectly affected by the treatment through any social and economic interaction with the treated. For example, the recipients of CCTs may share resources with, and affect the incentives to accumulate human capital of, ineligible households who live in treated localities (Angelucci and De Giorgi, 2009). Recipients of human-capital-enhancing programs who receive free textbooks and computers may share them with untreated children (such as relatives, friends, and members of the local church), raising enrollment for both groups.

Behavioral effects. These spillover effects stem from an intervention that affects the behavioral or social norms within the contexts (say a locality) in which these interactions are relevant. For example, Avitabile and Di Maro (2007) show that a CCT in rural Mexico can affect the social norm whereby husbands oppose their spouses being screened for cervical cancer by male doctors. The intervention exogenously increases the number of women who screen for cervical cancer in each locality, which is dramatically low in these contexts, and by doing so makes it more costly for husbands to follow the social norm. This produces a spillover effect: the screening rate of women ineligible for the program, but living in the villages covered by the program, goes up because of the increased screening of eligible women complying with the conditionality of the program.

While one can classify the types of spillover effects, the evaluation designs we will discuss do not necessarily enable one to distinguish one from the other. One way to disentangle a specific type of spillover effect is to carefully design survey instruments to elicit information on the relevant type of unintended response. This highlights once more the importance of understanding which type of spillover effects might occur and how, before designing the evaluation (and the program).

3. Motivation and Implications

There are at least two reasons why accounting for spillover effects is important. First, to correctly identify and estimate the effect of a treatment on eligible subjects: the direct or intended treatment effect. Second, to correctly identify and estimate the effect of a treatment on ineligible subjects: the indirect or unintended treatment effect. Measuring these two effects enables one to design successful policies and to learn features of the "local economy" and of human behavior.

Before discussing these issues, it is useful to define the two parameters of interest. Consider a group of subjects (e.g. individuals, households). Some subjects are eligible ($E=1$) for some treatment ($T=1$), while others are not ($E=0$). Eligibility may or may not be random, as we discuss below. Y_0 and Y_1 are potential outcomes in the absence and presence of the treatment, respectively. The Average Treatment Effect on the Eligibles (ATE) is the effect of the treatment on subjects that are eligible for it: [3]

$$ATE = E(Y_1 - Y_0 \mid T=1, E=1)$$

That is, the ATE identifies the effect of the treatment on subjects that are meant to be treated. The Average Indirect Treatment Effect (ITE) is the effect of the treatment on the ineligibles:

$$ITE = E(Y_1 - Y_0 \mid T=1, E=0)$$

That is, the ITE identifies the effect of the treatment on subjects that are *not* meant to be treated.

It is useful to employ a specific example to discuss the importance of an experimental design that accounts for the presence of spillover effects. Consider the case in which the treatment, T, is a deworming drug, the subjects are pupils of a school, and the "local economy" is the school. The outcome of interest, Y, is the rate of infection with intestinal parasites. Pupils in the school study and play together, so providing a group of students with deworming drugs may also affect the infection rate of pupils that are not directly treated with the drug.

---

[3] In the formulas which follow we make use of the expectation operator $E()$, which gives the average value of the object to which it is applied. For example, for a sample of N individuals i we write:

$$E(Y_1 - Y_0 \mid T=1, E=0) = \frac{\sum_i^N (Y_{1,i} - Y_{0,i} \mid T_i=1, E_i=1)}{N}$$

While this example is fictional, the problem of intestinal parasite infection is real. There are high intestinal parasites infection rates in developing countries. Worms may lead to anemia, malnutrition, and pain. Deworming drugs are cheap and are offered through mass school-based deworming campaigns (endorsed by the WHO). Miguel and Kremer (2004), from whose work we drew in devising this example, discuss the issue, its implications, and the effect of deworming campaigns in detail.

The drug affects infection rates through two separate channels. First, it has a direct effect: by ingesting the drug, a child becomes parasite free. Second, there is an indirect effect: because other pupils are taking the drug, the child's re-infection rate drops, as the chances of her coming in contact with contaminated fecal matter decrease. Consider a school in which a group of pupils is offered the deworming drug ($E=1$) and another group is not ($E=0$). The ATE is the average effect of the deworming drug on the infection rate of eligible pupils, while the ITE is the effect of treating the eligible pupils on the infection rates of pupils not offered the drug. In this particular example, the ATE will be some combination of the first and second effects described above (the deworming effect of the drug and the lowered re-infection rates), while the ITE will be only caused by the decreased contact with contaminated feces. That is, offering the deworming drug generates a positive externality: treated children receive a deworming drug, their infection rates drop, and reduce the infection rates of children who do not receive the drug.

Suppose the infection rate in the absence of the treatment were 80% and treating a random group of pupils in a school decreased this rate to 10% for the eligibles and to 60% to the ineligibles. In this case, the ATE and the ITE would be -70 and -20 percentage points. However, the policy evaluator does not know that. To measure this effect, he splits the school pupils into two randomly assigned groups and administers the drug to one group only. The naive evaluator compares the infection rates of the two groups of pupils exploiting the randomization, that is assuming that the infection rate for the control group is the same rate we would observe in the absence of the treatment, or that

$$60\% = E(Y \mid T=1, E=0) = E(Y_0 \mid T=1, E=1).$$

As just mentioned, we know this is not true. The naive evaluator would estimate an ATE of -50 percentage points, given by

$$E(Y \mid T=1, E=1) - E(Y \mid T=1, E=0) = 10\% - 60\%$$

thus underestimating the true ATE (in absolute value) by 20 percentage points. Moreover, he would fail to notice that the ineligibles are also indirectly benefiting from the deworming drug, as their infection rate has dropped.

To understand the magnitude of this double underestimation of the program effect and its negative policy implications, suppose the school has 100 pupils and the treatment and control groups have an equal size of 50 students each. Before the drug is offered to the treatment group, 80 pupils are infected. After treating 50 students, only 5 treated pupils ($50 \times 10\%$) and 30 control pupils ($50 \times 60\%$) are infected, for a total of 65 healthy children compared to the initial 20. However, the naive evaluator compares the infection rates in the two groups after administering the drug, and erroneously believes that the drug was successful for 25 pupils in the treatment group and none in the control group. That is, first he fails to notice that 10 other students in the treatment group have benefited from the drug; then he fails to notice that the drug has also reduced the infection rates for the control group, with only 30 infected pupils rather than the original 40. In total, 20 more children are not infected compared to his estimate.

Suppose the drug cost US$1 per child and the expected benefits of not having worms were 1.5. The naive evaluator would conclude that this treatment was not cost effective, because its cost of US$50 was higher than its benefit of US$37.5 ($1.5 \times 25$). In fact, the drug *is* cost effective because, when 50 out of 100 pupils are treated, 45 pupils stop being infected[4], for a total benefit of US$67.50 ($1.5 \times 45$) that exceeds its costs. The drug would also result in being cost effective if the evaluator managed to measure only its ATE properly, as the benefits for the treated would be US$52.50 ($1.5 \times 35$).

To summarize the lessons from this simple example, failing to have an evaluation design that accounts for spillover effects results in a double underestimation of the program's impact: first, because the average effect on the eligible children is underestimated; second, because the effect of the treatment on the ineligible children goes completely unnoticed. This double underestimation has the following negative policy implications: (1) researchers would mistakenly conclude that the treatment is less effective than it actually is (and that reducing

[4]In principle, it is also possible to define a total treatment effect as the sum of ATE and ITE. In this simple example this would be straightforwardly equal to 45 pupils who stopped being infected. However, in practical applications it is not always immediately clear how to define the total treatment effect. See appendix A.1 for a more detailed discussion of this point.

infections is costlier than it is); and (2) the wrong policy recommendations could be made as a result (e.g. to drop the program because it is too costly). Importantly, the experiment gives "wrong" information even if the randomization is performed successfully. That is, one cannot easily infer the existence of spillover effects from data alone. A conceptual framework of how the treatment may indirectly affect the outcomes for ineligible subjects is essential for understanding whether and to what extent there may be spillover effects.

4. Experimental Design

Experimental designs for program evaluations that fail to consider spillover effects may, if the randomization is within the local economy, provide incorrect estimates of the average treatment effect on eligibles and miss potential important indirect effects. Given this is the case one should use a different experimental design in the presence of spillover effects.

The key insight is that, rather than randomizing *within* separate local economies, one should randomize *between* them[5]. For example, consider the case of the deworming drug. If one believes the main indirect effects of the drug occur at the school level, then it is important to ensure the experiment is executed comparing pupils that attend different schools, rather than pupils that attend the same school. While this is the basic insight, different types of experimental designs are possible, and different types of data may be collected, depending on the evaluator's goal, time, and budget.

4.1 Measuring ATEs: basic randomization

The simplest experimental design in the presence of spillover effects consists of using the local economy, rather than the individual subject, as the randomization unit. In the example considered, it consists of selecting a sample of schools and randomly assigning two subsets to the treatment and control groups ($T=1$ and $T=0$).

If the goal of the evaluation is to measure the effect of the treatment on eligible subjects, then one could simply treat all subjects in the treatment group and none of the subjects in the control group. This would mean giving the deworming drug to all pupils at the treatment schools. In this design, therefore, all pupils are eligible (E=1) but only the ones attending the randomly selected group of treatment schools are actually offered the treatment. The school randomization ensures all individual and school characteristics have the same distribution in the two groups. That is, there is no systematic difference between the observable and unobservable characteristics of treatment and control schools, or between treatment and control pupils. In addition, this design ensures there are no spillover effects, as long as it is true that these effects

[5]Here we are abstracting from some practical considerations which might indicate that randomization at the local economy level is not the best way to go. For example, there might be cases, such as small pilot studies, in which one is left with only a very few local economies (localities, neighborhoods, etc.) between which to randomize. In this case, randomization between localities might not create a good counterfactual, that is control localities will not necessarily be comparable to treatment localities because samples are too small.

exist only at the hypothesized level, such as within the school but not across different schools. In other words, the effect of providing deworming drugs to pupils in the treatment schools has no effect on pupils who attend control schools. These two assumptions enable one to evaluate the average effect of the treatment on eligible subjects by comparing the average observed outcomes of pupils in the two sets of schools:

$$ATE = E(Y_1 - Y_0 \mid T = 1, E = 1) = E(Y \mid T = 1) - E(Y \mid T = 0)$$

We suppressed the condition that $E = 1$ because in this design there is no further grouping of students into eligible and ineligible.

4.2 Measuring ATEs and ITEs: double randomization

If one's goal is simply to evaluate the benefit of the treatment on eligible subjects, the above design is adequate. However, by using this design one runs the risk of missing out on valuable information.

First, one may still underestimate the broader effect of the program by failing to measure its effect on ineligible subjects who are in contact with the eligibles. Whether this is an actual issue or not depends on the program specifics. For example, if the policy in question is whether to treat all pupils in all schools in a given country, school attendance is high, and only school-age children have high infection rates, then there are probably no spillover effects. On the other hand, suppose enrollment rates were low. Then it would be possible that the siblings of children going to school also indirectly benefited from their relatives being given the deworming drug, but the above experimental design would not let the evaluator estimate the possible magnitude of such effects.

Second, we may fail to understand the mechanisms that make treatment successful for ineligible subjects.

A type of experimental design that allows one to measure the effect of the treatment on both eligible and ineligible subjects is a "double randomization". The first randomization is the same as before: randomly assign "units" (e.g. schools) to treatment (T=1) and control (T=0) groups. Then, offer the treatment only to a randomly selected group of subjects ($E = 1$) in the treatment units and not to the remaining students in the treatment schools ($E = 0$). In this case one would split the pupils in the treatment schools into two randomly selected groups. There are

thus three distinct groups, defined by treatment and eligibility: eligible and ineligible subjects in treatment units and subjects in control units. If the two randomizations are effective, these three groups are identical ex ante, that is their characteristics do not systematically differ. The three groups are different ex post. The eligible subjects in treatment units receive the treatment; the ineligible subjects in treatment units are indirectly affected by the treatment; if the spillover effects occur only within unit, the average outcome of pupils in control units is equivalent to the average outcome in treatment schools in the absence of the treatment. Thus, the presence of these distinct groups enables one to estimate two different parameters, the ATE and the ITE:

$$ATE = E(Y_1 - Y_0 \mid T = 1, E = 1) = E(Y \mid T = 1, E = 1) - E(Y \mid T = 0)$$
$$ITE = E(Y_1 - Y_0 \mid T = 1, E = 0) = E(Y \mid T = 1, E = 0) - E(Y \mid T = 0)$$

Since the grouping into eligible and ineligible subjects is randomly assigned, the evaluator does not need to form these groups in control units. This is why the second conditional expectation is the same for both the ATE and ITE.

The key assumptions under which these comparisons identify the two average treatment effects are the same as above. First, that the two randomizations "work," i.e. that there are no *ex-ante* systematic differences in the characteristics of the three groups of subjects. Second, that any spillover effect of the treatment does not go beyond the unit boundary.

Going back to the example of the deworming drug, we could now measure the "true" treatment effects on eligible and ineligible subjects. The infection rates in the control school (80%) are the same as those in the treatment schools in the absence of deworming. The infection rates in treatment schools after half the pupils are randomly offered the deworming drugs are 10% for eligible and 60% for ineligible pupils. The estimated ATE and ITE are $ATE = 10\% - 80\% = -70$ percentage points; $ITE = 60\% - 80\% = -20$ percentage points.

4.3 A different design

In many interventions, eligibility for the program within the unit is decided on the basis of predetermined criteria (i.e. poverty score), meaning that randomization within the unit cannot be performed in these settings. However, in this case it is still possible to measure the spillover effects of the treatment on ineligible subjects living (or studying, working, etc.) in treatment units. The only difference in the experimental design is that the second step has some nonrandom

assignment.[6] There are now four distinct groups, depending on unit type and eligibility type: eligible subjects (E=1) in treatment (T=1) and control (T=0) units and ineligible subjects (E=0) in treatment (T=1) and control (T=0) units.

Treating the eligible subjects may affect ineligible subjects' outcomes. Therefore, as long as we observe outcomes for the four groups, we can measure the treatment effect on both the eligible and the ineligible subjects:

$$ATE = E(Y_1 - Y_0 \mid T = 1, E = 1) = E(Y \mid T = 1, E = 1) - E(Y \mid T = 0, E = 1)$$
$$ITE = E(Y_1 - Y_0 \mid T = 1, E = 0) = E(Y \mid T = 1, E = 0) - E(Y \mid T = 0, E = 0)$$

These comparisons identify the two average treatment effects as long as: (1) the unit randomization "works," i.e. there are no *ex-ante* systematic differences in the characteristics of eligible subjects in treatment and control units and of ineligible subjects in treatment and control units; and (2) any spillover effect of the treatment is only within the treated unit.

For example, suppose one offers the deworming drug to girls only, perhaps because the donor is an agency promoting female welfare. The first step would be, as before, to randomly group schools into a treatment and control group. The second step would be to offer the drug to female students in treatment schools. The four groups are girls (E=1) and boys (E=0) in treatment (T=1) schools, and girls (E=1) and boys (E=0) in control (T=0) schools. There may be spillover effects of the deworming drug among the male students attending treatment school. Therefore, one should collect data on both female and male pupils in the two groups of schools. In this way one can measure the effect of offering the drug to girls on both girls' and boys' infection rates.

While establishing subjects' gender is relatively straightforward, in other cases determining eligibility is more complex and involves the collection of multiple variables. For example, for conditional cash transfer (CCT) programs, eligibility is determined by poverty status. In this case the evaluator needs to collect enough data to compute a continuous poverty indicator for both eligible and ineligible subjects in treatment and control units. This process of *ex-ante* data collection for eligible and ineligible subjects, as well as *ex-post* collection of information on the outcomes of interest, makes the estimation of spillover effects costlier and more time consuming.

[6] See Appendix A.1 for more details on this point.

4.4 Understanding the mechanisms

A well-conceived evaluation design should be based on some hypothesis about whether spillovers are present and why they exist. After documenting the existence of spillover effects and measuring their magnitude, it might be useful to understand the mechanisms that generate them and, in doing so, attempt to verify the initial hypothesis. Having a clear idea of such mechanisms may help in the design of effective policies.

Understanding the causes of spillover effects almost always requires collecting more data. For example, consider the case of treating girls to a course of deworming drugs using the experimental design discussed in the previous section. Suppose that, after noticing that infection rates among boys is also lower because of the treatment, the evaluator wants to understand whether this is caused by a reduction in contaminated fecal matter or by a change in basic health behavior, such as washing one's hands. In that case, one would need to collect data on these two additional variables, measuring the contamination rates and the frequency of hand washing. If one were to find that hand-washing frequency does not differ between boys in treatment and control schools, but the level of fecal contamination is lower in treatment schools, this would suggest this latter channel is probably the main cause of the effect.

In this specific example, the additional evidence is at most suggestive. If one wanted to compare the benefits of basic health education versus deworming drugs and the interaction of the two treatments, the ideal experimental setting would be one in which there were four random groups of schools. Girls in these four groups would receive respectively: (1) the drug only, (2) the health education only, (3) both the drug and the health education, and (4) neither. Comparing girls' infection rates in each of the first three groups with the fourth would provide estimates of the effect of each of the three treatments on eligible pupils, while the comparison for boys would provide estimates of the effect of each of the three treatments on the eligible pupils.

Having a clear idea of what is behind the spillover effects can also lead to a better understanding of what ATE and ITE are really capturing, especially in those cases in which different types of spillover effects are coupled with a general equilibrium effect. For example, an intervention which endows a certain group within the local economy with more resources might have both a spillover effect in terms of sharing of resources between treated and untreated subjects (interaction effect) and a general effect on the local economy prices (general equilibrium effect). The setup for estimating the spillover effect will not change, although knowledge that

there is more than one mechanism behind the spillover will help to correctly interpret the estimated ATE and ITE. In this case the ATE would be the sum of the direct effect and the general equilibrium effect and the ITE would be the sum of the indirect (interaction) effect and the general equilibrium effect.

4.5 Important considerations

As discussed above, the key assumptions for the identification and estimation of treatment effects are that (1) the distribution of characteristics of subjects in treatment and control groups is *ex-ante* identical; and (2) any spillover effect of the treatment to ineligible subjects is circumscribed to the "local economy," the unit of the initial randomization.

Neither of these assumptions is directly testable. However, while one may obtain indirect evidence that the randomization "worked" by comparing the means or the distribution of characteristics of subjects in the treatment and control groups, it is often hard to test whether there are no spillover effects between two randomized units. It is therefore extremely important to have a good understanding of how local markets and institutions work before designing the experiment.

Knowledge of the local economy is crucial for a successful experimental design. For example, in understanding at what level the deworming drug externalities may operate, knowledge of sanitary habits and facilities in the school and at home is essential. Suppose schools had high-quality sanitary facilities and monitored pupils' hand-washing, while houses had low-quality facilities. In this case, the risk of contact with contaminated feces may be higher at home than at school. Therefore, it is possible that the siblings, rather than the classmates of treated children, may benefit indirectly from the treatment. In this case, a double randomization at the school level may detect no spillover effect although the drug has actually caused a reduction in the infection rates of treated children's siblings.

In sum, a "theory" regarding what may cause the spillover effects, and knowledge of the local social and economic features and interactions, are both essential to increasing confidence in the experimental design's suitability to control for or measure spillover effects.

5. Spillover Effects in Non-Experimental Designs

As argued in several parts of this paper the experimental design is the most valid way to estimate the direct and indirect (spillovers) impact of an intervention. However, in many cases an experimental design may not be a viable option, typically because of budgetary or ethical reasons. In what follows we focus on how to correctly estimate spillover effects with nonexperimental evaluation designs.

A quite general definition would label nonexperimental designs as those approaches which try to estimate the impact of a program when the assignment to the treatment group was not based on randomization (or experiment). Without randomization, the main challenge in terms of the evaluation design is to find a group (defined as the "comparison group"[7]) which can be compared to the treatment group. In the case of the presence of spillover effects, an added complication is that comparison groups also have to be found for the groups of those who effectively receive the program and those who do not; nonetheless a valid evaluation of direct and indirect effects of an intervention can still be performed as long as the evaluation design takes this into account from the very initial stages.

We follow here the same setting as above, that is we have an intervention which is offered only to a subgroup of subjects (E=1) living in the treatment units (T=1) where the program is active. However, here the assignment of a unit (locality, municipality, etc.) to treatment and the assignment of subjects to the program in treatment units are not random (that is, eligibility is decided according to some criterion, for example a poverty score). This means that, on the one hand a sample of non treatment units (T=0) would not necessarily be comparable to a sample of treatment areas; on the other hand, the group of subjects participating in the program (E=0) cannot be directly compared to those who are not eligible (E=1) as many observable and unobservable characteristics might differ. There are several nonexperimental methods which, under some assumptions, can estimate the program's impacts in this case.

One initial consideration, which would apply to all nonexperimental methods discussed here, refers to the data requirements for the estimation of spillover effects. Suppose we want to estimate the direct impact of the program by simply comparing those who receive the intervention (E=1) and those who do not (E=0) with both groups living in treatment localities

[7]In experimental designs we used the terminology "control group" but in nonexperimental designs it is more correct to speak about a "comparison group" in order to stress the fact that this group, which can be compared to the treatment group, was not selected by a random assignment.

(T=1). As discussed above these two groups are not immediately comparable but several econometric methods, which will be discussed below, can be used to control for these differences. While this simple comparison would estimate the direct impact of the program with minimal data requirements, it must be noted that it has two major drawbacks if spillover effects are present. First, the estimated impact would be biased by the fact that those who do not receive the intervention but live in treatment localities (E=0,T=1) likewise are affected by the program (through the spillover effect). For example, if the program has a positive impact on some relevant outcomes the difference between the E=1 and E=0 groups, both in T=1, is likely to be smaller than the "true" impact because the outcome of the E=0 group is affected by the positive spillover effect. The second drawback is that this simple strategy would not allow one to estimate the spillover effect because a comparison group for the (E=0,T=1) group—which is the group affected by the spillover—is not available.

In order to estimate spillover effects more data is needed: in our setting we would require samples of four different groups: eligibles in treatment units (T=1, E=1), eligibles in comparison units (T=0, E=1), ineligibles in treatment units (T=1, E=0) and ineligibles in comparison areas (T=0, E=0). Typically, areas which were not included in the program are those that can be used for the comparison groups. The complication in a nonexperimental setting is that these areas are by construction different from treatment areas: for instance, if the treatment units are chosen as those localities which are poorer, then those who are left out (T=0) are by definition "less poor" than those who are in (T=1).

One first nonexperimental design is to employ a propensity score matching (PSM) approach[8]. The basic assumption with this method is that one can find a set of observable characteristics which can be used to "match" the treatment groups with the comparison groups so as to make them comparable. In practice, a model in which participation in the program is explained by many pretreatment characteristics is estimated (with a probit or a logit regression) and then predictions of this estimation are used to create a score (the propensity score) from 0 to 1 (which can be thought of as the probability of participating in the program) which is assigned to each subject (individuals, households, localities, etc.). The idea is that then one can compare units which are "close" to each other in terms of the propensity score.

[8]In another paper in the Impact Evaluation Guideline series Hendrick, Maffioli, and Vazquez (2010) give a much more detailed treatment of PSM. See also Caliendo and Kopeining (2008).

Following the example above, while treatment (T=1) and comparison (T=0) areas are different by definition, this would not prevent one from finding households within T=1 which are comparable (i.e. with a similar propensity score) to households in T=0, before the intervention takes place. Intuitively, a PSM approach estimates the program's impact comparing units which have been made "more comparable" because they are matched on the basis of the propensity score. The main difference between this method and a simple linear regression is that the PSM only imposes minimal parametric restrictions. One strong assumption behind PSM is that the observable characteristics which are used to estimate the propensity score explain all the differences between the treatment and comparison group before the intervention. Obviously, this assumption fails if there other observable characteristics which explain these differences but that were not included in the model, or if there are unobservable characteristics which explain the differences.

In our setting, we can estimate the ATE matching those who participate in the program (E=1) in treatment areas (T=1) with those who were eligible to participate (E=1) in comparison areas (T=0) and the ITE matching those who were not eligible for the program (E=0) living in treatment areas (T=1) with those not eligible (E=0) in comparison areas (T=0). Note that in the following equations P(X) is the propensity score:

$$ATE = E(Y_1 - Y_0 \mid T = 1, E = 1, P(X)) = E(Y \mid T = 1, E = 1, P(X)) - E(Y \mid T = 0, E = 1, P(X))$$
$$ITE = E(Y_1 - Y_0 \mid T = 1, E = 0, P(X)) = E(Y \mid T = 1, E = 0, P(X)) - E(Y \mid T = 0, E = 0, P(X))$$

Here, in the absence of randomization, the assumption is that we can estimate ATE and ITE because we are matching treatment and comparison areas on the basis of the propensity score, P(X).

While the PSM is a powerful nonexperimental approach which can estimate the program's impacts in a relatively straightforward manner, it is based on quite stringent assumptions and it requires a large amount of data (large samples of pre and postintervention treatment and comparison groups). A different method which is valid under less strict assumptions and may require less data is a regression discontinuity (RD) approach[9]. The basic assumption behind RD is that typically programs are assigned on the basis of some score (e.g. a

[9]A detailed treatment of regression discontinuity is in another paper in the Impact Evaluation Guidelines series, by Chay, Ibarraran, and Villa (2010). See also Lee and Lemieux (2009).

poverty score or a credit score) and there is a cutoff point above which units (individuals, households, localities, etc.) are eligible for the program and below which they are not. Intuitively, units which are just above this cutoff would not differ greatly from units which are just below the cutoff, with the only difference being eligibility for the program. An RD approach exploits this assumption and it assumes that in small enough intervals around the cutoff point, units do not differ in terms of observable or unobservable characteristics.

The following example shows how RD can work. We have been assuming that assignment of areas to treatment is done in a nonrandom way; in most cases this type of targeting implies that "poorer", or in general "needier", areas are included in the program. Typically, characteristics related to the relevant targeting objectives are used to compute a score which is then employed to rank the areas. A cutoff point is then chosen[10] and only areas which have a score below (or above[11]) this cutoff will be included in the program. In our setting RD can be employed both for estimating the ATE and the ITE. The idea would be to use only areas which are just above and just below the cutoff point under the assumption that these areas have both comparable observable characteristics and, most importantly, comparable unobservable characteristics. Straightforwardly, the only important difference between these areas is that those just above the cutoff (T=1) receive the program and those below do not (T=0).

We can write:

$$
\begin{aligned}
ATE &= E(Y_1 - Y_0 \mid T=1, E=1, P(X)) = \\
&= E(Y \mid T=1, E=1, \text{"just above the cut off"}) + \\
&- E(Y \mid T=0, E=1, \text{"just below the cut off"}) \\
ITE &= E(Y_1 - Y_0 \mid T=1, E=0, P(X)) = \\
&= E(Y \mid T=1, E=0, \text{"just above the cut off"}) + \\
&- E(Y \mid T=0, E=0, \text{"just below the cut off"})
\end{aligned}
$$

[10]The cutoff point is often chosen according to criteria which are related to the program budget. For example, if only a certain number of localities can be included due to budgetary restrictions, then the cutoff point can be chosen accordingly.

[11]Inclusion of below or above the cutoff areas depends on the definition of the score. For example, if a higher value score means that the areas are poorer and the program wants to include poorer areas, then only areas above the cutoff will be included.

RD has two important strengths: it requires a smaller amount of data[12] and it estimates the impact in a very convincing way as it controls both for observable and unobservable characteristics. On the other hand, results from an RD estimation only represent a specific sample, that is units around the cutoff point, but not necessarily the general population of interest. We say that RD has a strong internal validity but weak external validity. In our setting, this would mean that RD is a powerful way to estimate the direct and indirect impact on the areas around the cutoff point (high internal validity) but this result cannot be easily extended to the other areas which are farther from the cutoff point (low external validity). This is even truer if there are reasons to believe that areas around the cutoff are different from areas far from the cut-off point. This would be the case in our example, as areas just above the cutoff (T=1) are those units which are only marginally poor and the other areas, substantially above the cutoff, are by definition poorer.

Another method which could be used to estimate the direct and indirect effects of a program is an instrumental variables (IV) approach[13]. Typically, any nonrandom assignment to a program creates a bias. For instance, program administrators of a health intervention may assign localities to a program because those areas are better equipped to provide the health packages. The comparison of localities with and without the program would then be biased by definition since one would be comparing groups which are constructed differently. The impact of any relevant outcome would then be a sum of the "true" impact and this bias. One could control for some of these differences, including many observed characteristics in the model, but in practice the choice of inclusion into the program would be based on criteria that were not necessarily observable, such as convenience or political, logistic, or budgetary considerations.

The IV approach tries to mitigate this bias by mean of another variable (an instrument, Z) which is correlated with the assignment to treatment variable (T) but is not affected by the bias above. For example, if we want to study the impact on screening for breast cancer of the same health intervention as in the example above, we could just compare groups based on the original nonrandom assignment (T=1 vs. T=0) and we might find a big positive impact. This impact would not be very informative however, as we would not know whether women were really

[12]This depends on which methodology is employed. Some use only the data around the cutoff point, but some others use the whole sample and so would require the same amount of data as other nonexperimental approaches. See the references provided above for more details.

[13]A full treatment of an IV approach can be found in Chapter four of Angrist and Pischke's book (2009).

screening more for breast cancer, or if the difference was just due to the difference in the characteristics of the two groups (among other things, localities with better health infrastructure (T=1) are likely to offer better screening services).

However, there might be other variables which explain the assignment to treatment but do not necessarily drive differences in screening for breast cancer, for example political affiliation of the locality leader or distance of the locality from the municipality center. While these variables can clearly drive the choice of assigning the locality to treatment, it is much less clear why they should explain differences in the rate of screening for breast cancer. In general, an ideal instrumental variable has two main characteristics: it is correlated with the treatment variable (Z is correlated with T) and the only effect it has on the relevant outcome comes from the effect it has on the treatment variable (Z does not explain screening for cervical cancer, but it only affects it because it affects T).

An IV approach exploits these variables to estimate the impact of the program[14]. In our setting we can employ an IV approach to estimate ATE and ITE as follows:

$$ATE = E(Y_1 - Y_0 \mid T = 1, E = 1) = E(Y \mid T(Z) = 1, E = 1, X) - E(Y \mid T(Z) = 0, E = 1, X)$$
$$ITE = E(Y_1 - Y_0 \mid T = 1, E = 0) = E(Y \mid T(Z) = 1, E = 0, X) - E(Y \mid T(Z) = 0, E = 0, X)$$

An intuitive explanation of what IV does in our setting is that it will estimate the impact of the program for those localities which have different values of T because of changes in Z (T[Z]=1 vs. T[Z]=0). As in the RD case this means that the validity of the estimated impact is only specific to this group of localities and it cannot be extended to the total universe of interest (low external validity)[15].

The issue of external validity is a major one and when applying nonexperimental methods one has to consider that, under standard assumptions, PSM would have more external validity than RD and IV.

[14]In practice, the most difficult part is to find valid instrumental variables. It is always possible to come up with arguments against the validity of instrumental variables, including those we are using as an example in this chapter.
[15]Following this interpretation the estimated impacts are only local. We can call them: Local Average Treatment Effect (LATE) and Local Indirect Treatment Effect (LITE).

6. An Example: PROGRESA

Consider the case of the Mexican conditional cash transfer (CCT) program, PROGRESA.[16] The program's objectives are to improve education, health, and nutrition for poor households in rural Mexico. It provides cash transfers to eligible households on condition that they send their children to school, have periodic health checks, and participate in informal health and nutrition classes.

The transfers are a sizeable share of household income. The eligible households in the villages sampled for the policy evaluation receive monthly grants of about 200 *pesos* per household, which is about 143% and 100% of the average food consumption per adult for the eligible and ineligible households in control villages, respectively.

Suppose the goal of the evaluator is to measure the effect of PROGRESA on consumption. As discussed, one of the things to consider when designing the experiment is whether there may be spillover effects. To understand this, it is important to know some of the features of rural Mexico. First, the villages that the program targeted are small and marginalized (a combination of financial and geographic isolation), and the residents' income tends to be very volatile. It is conceivable that households may want to borrow money or buy insurance to ensure their consumption is stable even though their income varies over time. However, there are no formal credit or insurance institutions such as banks. Therefore, it is highly likely that these households will resort to informal activities to stabilize their consumption, consistent with the evidence that consumption is much more stable over time than income.

One such activity may be to share resources with other households in the village. Households may share assets or labor, make each other loans, or give each other gifts. Social scientists have collected abundant evidence that suggests these informal activities are common and frequent among the poor. In particular, we have reason to believe that these informal resource-sharing networks may be important also in the villages in which PROGRESA is implemented. For example, these villages have very low migration rates (in 1999, only about 5% of households had at least one member leave in the previous five years, of which 20% moved within the same village). Subsequently, about 80% of households have relatives in the same village. In these villages, groups of related households (parents' offspring and siblings, as well as more distant relations) live and work in close proximity with each other. It is therefore highly

[16]The following discussion is a synthesis of the work by Angelucci and De Giorgi (2009).

likely that, when one household experiences high earnings, be it a good harvest or a government transfer, it might share part of it with its extended family.

This discussion suggests the following. First, to evaluate the effect of PROGRESA on the eligibles, one should not compare their consumption with that of ineligible households living in the same village as the eligibles, but randomly select a group of treatment and control villages, choosing eligible households in these villages (if one believes these spillover effects occur mainly at the village level), and collecting data on eligible households in both treatment and control villages. Second, to measure spillover effects one should also collect information on ineligible households living in both treatment and control villages.

The design of the experiment is exactly as described. Because of logistical constraints, the program was started in a limited number of localities and gradually expanded to cover all the targeted localities. The evaluator exploited this gradual phasing-in of the program to create the following experiment. Out of a sample of 506 villages, 320 were randomized to receive the program from May 1998. The remaining 186 villages started receiving the program from the end of 1999. A village-level census conducted before the randomization ensured there were data on all households in the 506 villages. Each household was classified into eligible and ineligible based on a poverty score computed from observable characteristics (ranging from education to assets owned, to dwelling characteristics). Thus, we have information on eligible ($E=1$) and ineligible ($E=0$) households in treatment ($T=1$) and control ($T=0$) villages, as shown in Figure 1.

Importantly, there were both eligible and ineligible households in the same extended family. Thus, it is likely that a woman in a household whose income increased by 25% because of the PROGRESA grant shared some of it with her sister's household, for example. This leads to the formulation of our testable hypotheses.

- Hypothesis 1: PROGRESA increases the consumption of eligible households, that is the consumption ATE is positive:

$$ATE^{C} = E(C_1 - C_0 \mid E=1, T=1) > 0 .$$

- Hypothesis 2: PROGRESA has a positive indirect effect on the consumption of ineligible households, that is the consumption ITE is positive:

$$ITE^C = E(C_1 - C_0 \mid E = 0, T = 1) > 0 .$$

Because of the experimental design, to test these hypotheses one can simply compare average consumption of eligibles and ineligibles in treatment and control villages:

$$E(C \mid E = 1, T = 1) - E(C \mid E = 1, T = 0)$$
$$E(C \mid E = 0, T = 1) - E(C \mid E = 0, T = 0)$$

To increase the precision of the estimates, it is common practice to estimate the above parameters using a regression, adding predetermined determinants of consumption. These are variables that affect consumption (e.g. socioeconomic characteristics), but which are not correlated with the treatment dummy because they are measured *before* the beginning of the program. That is, one can estimate the following regression, in which, as above, T indicates the treatment status and E the eligibility. The subscripts i and v refer to the household and the local economy, respectively.

$$Y_{iv} = \alpha_0 + \alpha_1 T_v + \alpha_2 E_i + \alpha_3 T_v E_i + \alpha_4 X_i + \varepsilon_{iv}$$

The parameters $\alpha_1 + \alpha_3$ and α_1 identify the ATE and the ITE. The variables (X) are: household poverty index and land size; locality poverty index and number of households; and the gender, age, and literacy of the head of household and whether he or she speaks an indigenous language. All the X variables are at 1997 values.

Before estimating average treatment effects, however, it is necessary to check that the randomization "worked," that is, that the characteristics of subjects in treatment and control villages did not differ systematically. Berhman and Todd (1999) perform this exercise, confirming that the individual and household characteristics are balanced between treatment and control villages.

Table 1 shows the actual data. As hypothesized, both effects are positive and significant. The consumption ATE is between 24 and 30 *pesos*, a 15 to 20% increase in monthly food consumption per adult equivalent, compared to the levels in control villages (which, by assumption, are not affected by the program's existence). The consumption ITE is also positive: PROGRESA causes ineligibles' monthly food consumption to increase by about 20 *pesos*, which is roughly a 10% increase compared to the consumption level of ineligible households in control villages.

While one has now established the existence of positive spillover effects on consumption, provided the village randomization was successful and the consumption spillover effects occurred only within villages, it is not yet clear whether the proposed mechanism (informal gifts and loans between villagers, especially family members) is causing these indirect effects. Besides working through informal gifts and transfers between relatives and friends, the program may affect the labor and goods markets, as well as savings.

For example, it is possible that the parents of children whose education is now subsidized may decide to work less, or may choose to take their children out of work and into school. Both these effects may end up increasing local wages, which would increase ineligibles' income, and therefore consumption. In addition, if the eligibles started purchasing more goods and services, the ineligibles may earn more from the sales of such commodities, and again this would enable them to increase their consumption.

If one is interested in understanding to what extent each of these channels is causing the observed increased consumption for the ineligibles, it is necessary to collect data on labor income or wages, goods prices or quantities sold, for instance. Angelucci and De Giorgi (2009) conduct this analysis and show that the program's indirect effect on consumption is not caused by changes in the labor and goods markets. The increase in consumption, on the other hand, is paralleled by an increase in loans and gifts received by ineligible households in treatment villages (and by a small decrease in the value of their savings).

The implications of these findings are that failing to account for the program's spillover effects result in a 12% underestimation of the effect of PROGRESA on consumption in treatment villages. More generally, this exercise teaches one how households cope with high income variability and the absence of formal credit and insurance.

Learning about the mechanism is useful for at least two additional reasons. First, it helps one understand what to expect when implementing the program at the national level. In this case, the findings suggest the program will continue to have sizeable spillover effects on consumption, as long as the informal sharing institutions—such as the extended family—are as important in the rest of Mexico as in the sampled villages. Second, the findings might help policy makers design more cost-effective policies.

Suppose one were to find that all village members benefit from the program because they redistribute the PROGRESA grant in the village regardless of the actual identity of the households that receive the transfer. This would mean that costly and time-consuming household-level targeting may not be required. Geographic targeting, using poverty maps to identify indigent areas, could be a cheaper and faster alternative.

7. Conclusions

Subjects that are not eligible for a particular treatment will often benefit from it indirectly in a wide variety of interventions. Measuring these "spillover effects" (or Indirect Treatment Effects) and understanding the mechanisms that generate them can be crucial to the design of effective policies.

Failing to have an evaluation design that accounts for spillover effects can result in a double underestimation of the program's impact. First, the average effect on eligible subjects is underestimated; second, the effect of the treatment on ineligible subjects goes completely unnoticed. This double underestimation may lead to the wrong policy recommendations (for example dropping the program because it does not seem cost effective).

This guideline makes some key recommendations to aid the design of evaluations that account for the presence of spillover effects. Here is a summary of these recommendations:

1. Have a theory about what may cause the spillover effects, who may be affected by them, and how. Knowledge of the local economy and of the types of interactions between its members is crucial for the design of an evaluation that at least accounts for the existence of spillover effects. Ideally it should also enable one to measure the effects and understand why they arise. Having a hypothesis about spillover effects often means having better assumptions on how the program works which should help to design better policies.
2. To measure spillover effects effectively, the evaluation design must take them into account from the start. This often means collecting more data, for instance by surveying geographical areas which are not affected at all by the intervention. Evaluation designs which do not consider these issues from the very beginning will typically not be able to measure spillover effects.
3. To identify and estimate the average treatment effect on eligible subjects (ATE) in the presence of spillover effects, select a control group that is not indirectly affected by the program. Theory will guide in this choice. Consider, for example, the case in which the indirect effects operate at village level. With an experimental design, randomize at village level. With a selection-on-observables design (matching), select treatment and control groups from different villages.

4. To identify and estimate the indirect treatment effect (ITE), that is the effect of the treatment on individuals who are not directly affected by the treatment, have two groups of subjects with similar characteristics (the same distribution of characteristics, if there is random assignment) but only one group that is indirectly affected by the treatment. Consider, for example, the case in which the indirect effects operate at village level. With an experimental design, do a double randomization across and within villages. With a selection-on-observables design (matching), select subjects that live in treated villages and that are indirectly affected by the treatment and compare them with similar subjects from different villages.
5. To understand the *mechanisms* that cause the spillover effects, think about potential competing explanations and collect data on relevant outcomes. In many cases, unveiling the mechanisms behind the spillover effects will lead to a better understanding of how the program works in general.

References

Angelucci, M., and G. De Giorgi. 2009. “Indirect Effects of an Aid Program: How do Cash Injections Affect Ineligibles’ Consumption?” *American Economic Review*, 99(1), 486-508, March 2009.

Angrist, J.D. and J. Pischke. 2009. *Mostly Harmless Econometrics.* Princeton, NJ, United States: Princeton University Press.

Avitabile, C., and V. Di Maro. 2007. “Spillover Effects in Healthcare Programs: Evidence of Social Norms and Information Sharing”. Paper presented at the 2009 Latin America and Caribbean Economic Association. Buenos Aires, Argentina. A copy of the paper can be provided upon request.

Berhman, J., and P. Todd. 1999. “Randomness in the Experimental Sample of Progresa (Education, Health, and Nutrition Program)”. International Food Policy Research Institute Working Paper. Washington DC, United States: IFPRI.

Caliendo, M., and S. Kopeining. 2008. "Some Practical Guidance for the Implementation of Propensity Score Matching". *Journal of Economic Surveys*. Blackwell Publishing, vol. 22(1), pages 31-72, 02.

Chay, K., P. Ibarraran, and J. M. Villa. 2010. "Regression Discontinuity and Impact Evaluation". Forthcoming in the Inter-American Development Bank-SPD Impact Evaluation Guidelines series. Washington DC, United States: IDB.

Hendrick, C., A. Maffioli, and G. Vazquez. 2010. "Guidelines on Matching". Forthcoming in the Inter-American Development Bank-SPD Impact-Evaluation Guidelines series. Washington DC, United States: IDB.

Lee, D., and T. Lemieux. 2009. "Regression Discontinuity Designs in Economics". National Bureau of Economic Research .Working Paper No. 14723, Feb 2009. Cambridge, MA, United States: NBER.

Miguel, E., and M. Kremer. 2004. “Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities”. *Econometrica* 72(1): 159-217.

Other useful references (not cited in the paper)

Angelucci, M., G. De Giorgi, M.A. Rangel et al. 2007. *Extended Family Networks in Rural Mexico: A Descriptive Analysis*. In; T. Besley and R. Jayaraman, editors. Forthcoming in CESifo Conference Volume on Institutions and Development. Cambridge, MA, United States: MIT Press.

Bobonis, G., and F. Finan. 2008. "Endogenous Social Interaction Effects in School Participation in Rural Mexico". *The Review of Economics and Statistics* 91 (4) (2009), 695–716.

Duflo, E., and E. Saez. 2003. "The Role of Information and Social Interactions in Retirement Plan Decisions: Evidence from a Randomized Experiment". *Quarterly Journal of Economics* 118, 815-842.

Hahn, J., and K. Hirano. 2009. "Design of Randomized Experiments to Measure Social Interaction Effects," forthcoming in *Economics Letters*.

Lalive, R., and M. A. Cattaneo. 2009. "Social Interactions and Schooling Decisions". *The Review of Economics and Statistics* 91(3), pages 457-477, 01. Cambridge MA, United States: MIT Press.

Moffitt, R. 2001. "Policy Interventions, Low-Level Equilibria, and Social Interactions". In: S.N. Durlauf and P. Young, editors. *Social Dynamics* 45-82. Cambridge, MA, United States: MIT Press.

Manski, C. 1993. "Identification of Endogenous Social Effects: The Reflection Problem". *The Review of Economic Studies* 60(3): 531-42.

Appendix A. Technical issues

This appendix briefly discusses some more technical issues which were not addressed, or were only introduced, in the main text.

1. Total effect of the program

One might be interested in calculating the total treatment effect of the program. Intuitively, this should be defined as the sum of all the effects the program is having and could be named as *TATE* (Total Average Treatment Effect):

$$TATE = ATE + ITE$$

While the intuition is correct, some caveats apply when one wants to calculate this effect. Consider again the simple example given in the main text, in which we had a school with 100 pupils and treatment and control groups of equal size (50 students). Before the treatment 80 students were infected in total and, assuming randomization worked well, we expected that these were shared equally among treatment and control groups (40 pupils infected in each group). After treating 50 students, only 5 treated pupils and 30 control pupils were infected, which meant that we now had 35 pupils in total infected in the school.

If we compare this with the 80 pupils infected before the intervention, then we conclude that the total effect of the program is that $35-80=45$ pupils stopped being infected, which can be divided in an $ATE=40-5=35$ pupils stopped being infected and an $ITE=40-30=10$ pupils. In this example there are no particular complications in defining the TATE as we are considering a well-defined population (the treated school) and we are not taking averages of some outcome but are just counting pupils that are infected before and after the intervention.

In practical applications things are not so straightforward. For example, consider the general case we have been using in this paper in which we have four different groups: treatment (T=1) and control localities (T=0) and eligible (E=1) and ineligible households (E=0). We have estimated the ATE as the difference in outcomes of E=1 households between T=1 and T=0 localities, and the ITE with the same difference for E=0 households. In this case if one wants to define the TATE as the sum of ATE and ITE, at least two complications must be considered. First, and most importantly, the ATE and ITE are based on populations that are different as eligibility for the program was not assigned randomly. In particular, the ATE is estimated for a

group of people (E=1) who are "poorer" than the population on which the ITE is based (E=0). This is an issue which cannot be easily solved (in practical applications it is usually not possible to assign eligibility for the program randomly) but that has to be taken into account when interpreting the TATE.

In other words, in the case of random assignment of eligibility one could make the point that the ITE is a parameter which applies to the same group of subjects on which we estimated the ATE. On the contrary, with nonrandom assignment of eligibility ITE applies to a group with different characteristics. In general, we are interested in estimating effects of interventions in the "real world", which will typically have a rule for assigning eligibility, and so the difference in the interpretation of the ATE and ITE will be exactly consistent with the nature of the program. However, in those cases in which one wants to extend the results of the ITE to the group on which the ATE was estimated and eligibility was not assigned randomly, one should take into account that these are parameters which apply to different groups[17].

Another complication is that the ATE and ITE are calculated based on samples which do not necessarily have the same characteristics. In particular, the sample sizes of the E=1 and E=0 groups are not necessarily the same. However, this is a less important point which can be easily solved by calculating the TATE as a weighted average of the ATE and ITE, where the weights are the sample sizes of the E=1 and E=0 groups.

2. Identification assumptions of ATE and ITE

Consider the parameters ATE and ITE, which, for ease of exposition, we define below again:

$$ATE = E(Y_1 - Y_0 \mid T = 1, E = 1) = E(Y_1 \mid T = 1, E = 1) - E(Y_0 \mid T = 1, E = 1)$$
$$ITE = E(Y_1 - Y_0 \mid T = 1, E = 0) = E(Y_1 \mid T = 1, E = 0) - E(Y_0 \mid T = 1, E = 0)$$

Of the objects in the definition of ATE and ITE we cannot observe $E(Y_0 \mid T = 1, E = 1)$ and $E(Y_0 \mid T = 1, E = 0)$, that is we cannot observe the potential outcome in absence of the treatment Y_0 for those subjects which are in program areas (T=1).

[17] For example, analysis of the mechanisms behind the ITE might unveil a very important channel through which the program works (say the program is weakening the social norm of husbands' opposition to their spouses being screened for cervical cancer by male doctors). If eligibility was assigned nonrandomly, and E=1 and E=0 groups had different characteristics, we should take into account that this mechanism might not work for the eligible group (or work in other ways).

Under some assumptions, which vary according to the method used, one can recover the unobservable objects above using data from untreated areas (T=0). That is, the following equalities will hold true:

$$E(Y_0 \mid T=1, E=1) = E(Y_0 \mid T=0, E=1)$$
$$E(Y_0 \mid T=1, E=0) = E(Y_0 \mid T=0, E=0)$$

In order to identify ATE and ITE one needs to make assumptions under which these equalities are true. There are two main assumptions:

- The treatment has no effect in T=0 areas (this is shared by experimental and nonexperimental methods)
- There are no systematic differences between T=1 and T=0 groups.
    - *Experimental design:* Assumption holds because of random assignment of treatment and control group.
    - *Propensity score matching:* Assumption holds after matching T=1 and T=0 on the basis of pre-program observable characteristics.
    - *Regression discontinuity:* Assumption holds only in an interval around the cutoff point for eligibility for the program.
    - *Instrumental variable:* Assumption holds after controlling for observable characteristics X and using an instrumental variable Z which is correlated to the program assignment variable T but does not affect the outcome variable in addition to the effect it has through the correlation with the T variable.

3. Estimation issues

Consider the following regression model in which, as above, T indicates the treatment status and E the eligibility. The subscripts i, v and t refer to the household, the local economy and time (we assume that there are only two periods, before the program $t=0$ and after the program $t=1$) respectively. X_{ivt} is a vector of observable characteristics, ε_{iv} represents unobservable characteristics which vary across household and local economies but are constant over time and ν_t stands for unobservable characteristics which vary over time but are constant across

households and local economies.

$$Y_{ivt} = \alpha_0 + \alpha_1 T_v + \alpha_2 E_i + \alpha_3 T_v E_i + \alpha_4 X_{ivt} + \varepsilon_{iv} + \nu_t$$

Under the assumptions in section two, the parameters $\alpha_1 + \alpha_3$ and α_1 identify the ATE and the ITE respectively.

While this regression model will stay unchanged, estimation issues can vary depending on the type of data available and the method used. With an experimental design and only ex post (after the intervention's inception) the ATE and ITE can be estimated with a simple OLS. If one believes that randomization worked perfectly then the conditioning variables (X) are not really needed, as in principle there should not be systematic differences between treatment and control groups. However, in practical applications one may want to include controls even in the case of random assignment as these would increase the precision of the estimates. In the case of nonexperimental methods and only *ex-post* data, controls are needed to control for the differences in treatment and comparison groups.

The availability of *ex-ante* (or baseline) data makes things substantially better in terms of estimation. First, with baseline data one can assess the differences between treatment and control groups at time $t = 0$ in the case of experimental design (and so check whether randomization worked well) as well as assess the differences between treatment and comparison groups in the case of nonexperimental designs. In addition to this, with baseline data a difference-in-difference (DID) strategy can be used. The basic idea of DID is to use the difference between outcome values in the counterfactual group (control or comparison group) before and after the program ($Y_{T=0,t=1} - Y_{T=0,t=0}$) as an indication of the trend the outcome would have had in the treatment areas had the program not been in operation. After controlling for this trend the remaining difference between treatment and counterfactual group can then be attributed to the program.

This strategy is especially useful when using nonexperimental methods in which pre-program differences between treatment and comparison group are present by definition and so outcome values are likely to be different at the baseline. In the case of experimental designs DID might be less useful as randomization should assure that the outcome values of treatment and control group are the same at the baseline, which means that value comparisons after the program should automatically take into account the trend effect. However, it can be shown that DID can increase the precision of the estimates in all the cases we are considering, including the

experimental one. In a regression framework the DID can be written as:

$$Y_{ivt} = \alpha_0 + \alpha_1 t + \alpha_2 T_v + \alpha_3 E_i + \alpha_4 T_v E_i + \alpha_4 T_v t + \alpha_5 E_i t + \alpha_6 T_v E_i t + \alpha_7 X_{ivt} + \varepsilon_{iv} + \nu_t$$

In this specification $\alpha_4 + \alpha_6$ estimates the ATE and α_6 estimates the ITE. Notice that the main assumption behind DID is that it can control for unobserved characteristics which are time invariant (ε_{iv}) but not for those that vary over time (ν_t).

One final consideration refers to standard errors: if the primary-sampling units are local economies (which we argue they should be in an ideal experimental design) then standard errors should be clustered at the local-economy level.

Figure 1. PROGRESA: Data structure

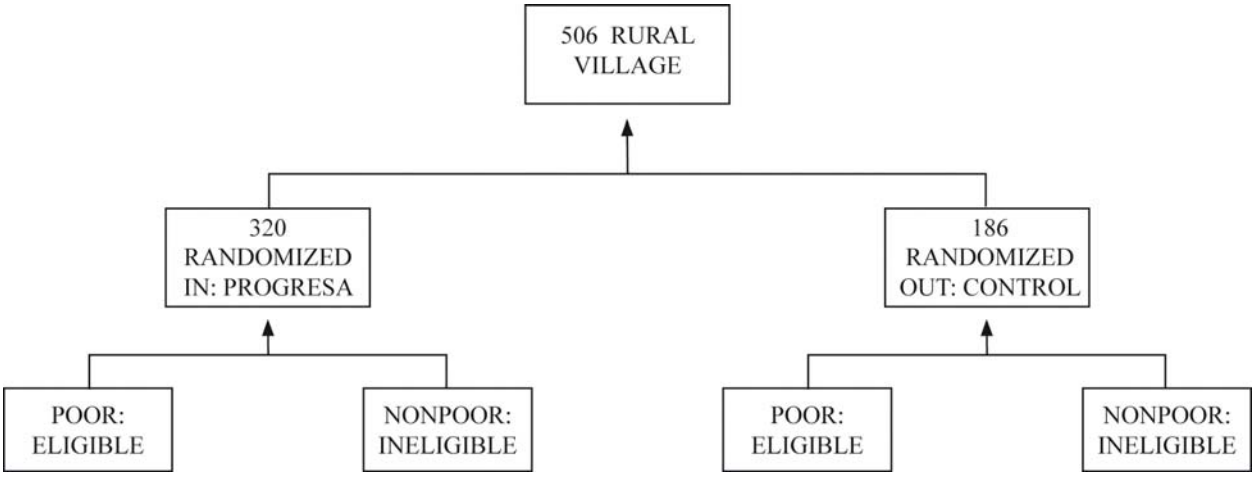

Table 1: Average peso monthly food consumption per adult equivalent: levels and differences

	Ineligibles			Eligibles	
	may-99	nov-99		may-99	nov-99
Control	213,69	206,71		159,92	153,7
	(212.19)	(232.56)		(158.33)	(126.72)
Treatment	233,06	224,09		185,66	184,31
	(303.79)	(285.61)		(193.81)	(172.25)
ITE	20,72	18,84	**ATE**	24,42	29,86
	(10.19)**	(9.42)**		(5.64)***	(4.79)***

Source: Angelucci and De Giorgi (2009).

Note: Monthly pesos per adult equivalent at Nov. 1998 prices; the exchange rate is roughly 10 pesos per USD. We report the standard deviations of the means and the standard errors, in brackets, of the treatment effects. The latter are clustered at village level with *** and ** indicating significance at the 1% and 5% levels respectively. The set of conditioning variables we add to the regressions in the left panel are: household poverty index, land size, head of household gender, age, whether he/she speaks indigenous language, literacy; at the locality level poverty index and number of households. All variables are at 1997 values.