Using Randomization in Development Economics Research: A Toolkit

Esther Duflo, Rachel Glennerster and Michael Kremer
BREAD Working Paper No. 136
December 2006

B R E A D

Working Paper

Bureau for Research and Economic Analysis of Development

Using Randomization in Development Economics Research: A Toolkit*

Esther Duflo†, Rachel Glennerster‡and Michael Kremer§

December 12, 2006

Abstract

This paper is a practical guide (a toolkit) for researchers, students and practitioners wishing to introduce randomization as part of a research design in the field. It first covers the rationale for the use of randomization, as a solution to selection bias and a partial solution to publication biases. Second, it discusses various ways in which randomization can be practically introduced in a field settings. Third, it discusses designs issues such as sample size requirements, stratification, level of randomization and data collection methods. Fourth, it discusses how to analyze data from randomized evaluations when there are departures from the basic framework. It reviews in particular how to handle imperfect compliance and externalities. Finally, it discusses some of the issues involved in drawing general conclusions from randomized evaluations, including the necessary use of theory as a guide when designing evaluations and interpreting results. *JEL Classification: I0; J0; O0; C93.* Keywords: Randomized evaluations; Experiments; Development; Program evaluation.

*We thank the editor T.Paul Schultz, as well Abhijit Banerjee, Guido Imbens and Jeffrey Kling for extensive discussions, David Clingingsmith, Greg Fischer, Trang Nguyen and Heidi Williams for outstanding research assistance, and Paul Glewwe and Emmanuel Saez, whose previous collaboration with us inspired parts of this chapter.

†Department of Economics, Massachusetts Institute of Technology and Abdul Latif Jameel Poverty Action Lab

‡Abdul Latif Jameel Poverty Action Lab

§Department of Economics, Harvard University and Abdul Latif Jameel Poverty Action Lab

Contents

1 Introduction

Randomization is now an integral part of a development economist's toolbox. Over the last ten years, a growing number of randomized evaluations have been conducted by economists or with their input. These evaluations, on topics as diverse as the effect of school inputs on learning (Glewwe and Kremer 2005), the adoption of new technologies in agriculture (Duflo, Kremer, and Robinson 2006), corruption in driving licenses administration (Bertrand, Djankov, Hanna, and Mullainathan 2006), or moral hazard and adverse selection in consumer credit markets (Karlan and Zinman 2005b), have attempted to answer important policy questions and have also been used by economists as a testing ground for their theories.

Unlike the early "social experiments" conducted in the United States—with their large budgets, large teams, and complex implementations—many of the randomized evaluations that have been conducted in recent years in developing countries have had fairly small budgets, making them affordable for development economists. Working with local partners on a smaller scale has also given more flexibility to researchers, who can often influence program design. As a result, randomized evaluation has become a powerful research tool.

While research involving randomization still represents a small proportion of work in development economics, there is now a considerable body of theoretical knowledge and practical experience on how to run these projects. In this chapter, we attempt to draw together in one place the main lessons of this experience and provide a reference for researchers planning to conduct such projects. The chapter thus provides practical guidance on how to conduct, analyze, and interpret randomized evaluations in developing countries and on how to use such evaluations to answer questions about economic behavior.

This chapter is not a review of research using randomization in development economics.[1] Nor is its main purpose to justify the use of randomization as a complement or substitute to other research methods, although we touch upon these issues along the way.[2] Rather, it is a practical guide, a "toolkit," which we hope will be useful to those interested in including

[1] Kremer (2003) and Glewwe and Kremer (2005) provide a review of randomized evaluations in education; Banerjee and Duflo (2005) review the results from randomized evaluations on ways to improve teacher's and nurse's attendance in developing countries; Duflo (2006) reviews the lessons on incentives, social learning, and hyperbolic discounting.

[2] We have provided such arguments elsewhere, see Duflo (2004) and Duflo and Kremer (2005).

randomization as part of their research design.

The outline to the chapter is as follows. In Section 2, we use the now standard "potential outcome" framework to discuss how randomized evaluations overcome a number of the problems endemic to retrospective evaluation. We focus on the issue of selection bias, which arises when individuals or groups are selected for treatment based on characteristics that may also affect their outcomes and makes it difficult to disentangle the impact of the treatment from the factors that drove selection. This problem is compounded by a natural publication bias towards retrospective studies that support prior beliefs and present statistically significant results. We discuss how carefully constructed randomized evaluations address these issues.

In Section 3, we discuss how can randomization be introduced in the field. Which partners to work with? How can pilot projects be used? What are the various ways in which randomization can be introduced in an ethically and politically acceptable manner?

In section 4, we discuss how researchers can affect the *power* of the design, or the chance to arrive at statistically significant conclusions. How should sample sizes be chosen? How does the level of randomization, the availability of control variables, and the possibility to stratify, affect power?

In section 5, we discuss practical design choices researchers will face when conducting randomized evaluation: At what level to randomize? What are the pros and cons of factorial designs? When and what data to collect?

In section 6 we discuss how to analyze data from randomized evaluations when there are departures from the simplest basic framework. We review how to handle different probability of selection in different groups, imperfect compliance and externalities.

In section 7 we discuss how to accurately estimate the precision of estimated treatment effects when the data is grouped and when multiple outcomes or subgroups are being considered. Finally in section 8 we conclude by discussing some of the issues involved in drawing general conclusions from randomized evaluations, including the necessary use of theory as a guide when designing evaluations and interpreting results.

2 Why Randomize?

2.1 The Problem of Causal Inference

Any attempt at drawing a causal inference question such as "What is the causal effect of education on fertility?" or "What is the causal effect of class size on learning?" requires answering essentially counterfactual questions: How would individuals who participated in a program have fared in the absence of the program? How would those who were not exposed to the program have fared in the presence of the program? The difficulty with these questions is immediate. At a given point in time, an individual is either exposed to the program or not. Comparing the same individual over time will not, in most cases, give a reliable estimate of the program's impact since other factors that affect outcomes may have changed since the program was introduced. We cannot, therefore, obtain an estimate of the impact of the program on a given individual. We can, however, obtain the average impact of a program, policy, or variable (we will refer to this as a treatment, below) on a group of individuals by comparing them to a similar group of individuals who were not exposed to the program.

To do this, we need a comparison group. This is a group of people who, in the absence of the treatment, would have had outcomes similar to those who received the treatment. In reality, however, those individuals who are exposed to a treatment generally differ from those who are not. Programs are placed in specific areas (for example, poorer or richer areas), individuals are screened for participation (for example, on the basis of poverty or motivation), and the decision to participate in a program is often voluntary, creating self-selection. Families chose whether to send girls to school. Different regions chose to have women teachers, and different countries chose to have the rule of law. For all of these reasons, those who were not exposed to a treatment are often a poor comparison group for those who were. Any difference between the groups can be attributed to both the impact of the program or pre-existing differences (the "selection bias"). Without a reliable way to estimate the size of this selection bias, one cannot decompose the overall difference into a treatment effect and a bias term.

To fix ideas it is useful to introduce the notion of a *potential outcome*, introduced by Rubin (1974). Suppose we are interested in measuring the impact of textbooks on learning. Let us call Y_i^T the average test score of children in a given school i if the school has textbooks and Y_i^C the test scores of children in the same school i if the school has no textbooks. Further, define Y_i as outcome that is actually observed for school i. We are interested in the difference $Y_i^T - Y_i^C$,

which is the effect of having textbooks for school i. As we explained above, we will not be able to observe a school i both with and without books at the same time, and we will therefore not be able to estimate individual treatment effects. While every school has two potential outcomes, only one is observed for each school.

However, we may hope to learn the expected average effect that textbooks have on the schools in a population:

$$E[Y_i^T - Y_i^C]. \tag{1}$$

Imagine we have access to data on a large number of schools in one region. Some schools have textbooks and others do not. One approach is to take the average of both groups and examine the difference between average test scores in schools with textbooks and in those without. In a large sample, this will converge to

$$D = E[Y_i^T|\text{School has textbooks}] - E[Y_i^C|\text{School has no textbooks}] = E[Y_i^T|T] - E[Y_i^C|C].$$

Subtracting and adding $E[Y_i^C|T]$, i.e., the expected outcome for a subject in the treatment group had she not been treated (a quantity that cannot be observed but is logically well defined) we obtain,

$$D = E[Y_i^T|T] - E[Y_i^C|T] - E[Y_i^C|C] + E[Y_i^C|T] = E[Y_i^T - Y_i^C|T] + E[Y_i^C|T] - E[Y_i^C|C]$$

The first term, $E[Y_i^T - Y_i^C|T]$, is the *treatment effect* that we are trying to isolate (i.e., the effect of treatment on the treated). In our textbook example, it is the answer to the question: on average, in the treatment schools, what difference did the books make?

The second term, $E[Y_i^C|T] - E[Y_i^C|C]$, is the *selection bias*. It captures the difference in potential untreated outcomes between the treatment and the comparison schools; treatment schools may have had different test scores on average even if they had not been treated. This would be true if schools that received textbooks were schools where parents consider education a particularly high priority and, for example, are more likely to encourage their children to do homework and prepare for tests. In this case, $E[Y_i^C|T]$ would be larger than $E[Y_i^C|C]$. The bias could also work in the other direction. If, for example, textbooks had been provided by a non-governmental organization to schools in particularly disadvantaged communities, $E[Y_i^C|T]$ would likely be smaller than $E[Y_i^C|C]$. It could also be the case that textbooks were part of

a more general policy intervention (for example, all schools that receive textbooks also receive blackboards); the effect of the other interventions would be embedded in our measure D. The more general point is that in addition to any effect of the textbooks there may be systematic differences between schools with textbooks and those without.

Since $E[Y_i^C|T]$ is not observed, it is in general impossible to assess the magnitude (or even the sign) of the selection bias and, therefore, the extent to which selection bias explains the difference in outcomes between the treatment and the comparison groups. An essential objective of much empirical work is to identify situations where we can assume that the selection bias does not exist or find ways to correct for it.

2.2 Randomization Solves the Selection Bias

One setting in which the selection bias can be entirely removed is when individuals or groups of individuals are randomly assigned to the treatment and comparison groups. In a randomized evaluation, a sample of N individuals is selected from the population of interest. Note that the "population" may not be a random sample of the entire population and may be selected according to observables; therefore, we will learn the effect of the treatment on the particular sub-population from which the sample is drawn. We will return to this issue. This experimental sample is then divided *randomly* into two groups: the *treatment* group (N_T individuals) and the *comparison* (or control) group (N_C individuals).

The treatment group then is exposed to the "treatment" (their treatment status is T) while the comparison group (treatment status C) is not. Then the outcome Y is observed and compared for both treatment and comparison groups. For example, out of 100 schools, 50 are randomly chosen to receive textbooks, and 50 do not receive textbooks. The average treatment effect can then be estimated as the difference in empirical means of Y between the two groups,

$$\hat{D} = \hat{E}[Y_i|T] - \hat{E}[Y_i|C],$$

where $\hat{E}$ denotes the sample average. As the sample size increases, this difference converges to

$$D = E[Y_i^T|T] - E[Y_i^C|C].$$

Since the treatment has been randomly assigned, individuals assigned to the treatment and

control groups differ in expectation only through their exposure to the treatment. Had neither received the treatment, their outcomes would have been in expectation the same. This implies that the selection bias, $E[Y_i^C|T] - E[Y_i^C|C]$, is equal to zero. If, in addition, the potential outcomes of an individual are unrelated to the treatment status of any other individual (this is the "Stable Unit Treatment Value Assumption" (SUTVA) described in Angrist, Imbens, and Rubin (1996)),[3] we have

$$E[Y_i|T] - E[Y_i|C] = E[Y_i^T - Y_i^C|T] = E[Y_i^T - Y_i^C],$$

the causal parameter of interest for treatment T.

The regression counterpart to obtain $\hat{D}$ is

$$Y_i = \alpha + \beta T + \epsilon_i, \tag{2}$$

where T is a dummy for assignment to the treatment group. Equation (2) can be estimated with ordinary least squares, and it can easily be shown that $\hat{\beta}_{OLS} = \hat{E}(Y_i|T) - \hat{E}(Y_i|C)$.[4]

This result tells us that when a randomized evaluation is correctly designed and implemented, it provides an unbiased estimate of the impact of the program in the sample under study—this estimate is internally valid. There are of course many ways in which the assumptions in this simple set up may fail when randomized evaluations are implemented in the field in developing countries. This chapter describes how to correctly implement randomized evaluations so as to minimize such failures and how to correctly analyze and interpret the results of such evaluations, including in cases that depart from this basic set up.

Before proceeding further, it is important to keep in mind what expression (1) means. What is being estimated is the *overall impact* of a particular program on an outcome, such as test scores, allowing other inputs to change in response to the program. It may be different from the impact of textbooks on test scores *keeping everything else constant.*

To see this, assume that the production function for the outcome of interest Y is of the form $Y = f(I)$, where I is a vector of inputs, some of which can be directly varied using policy

[3]This rules out externalities–the possibility that treatment of one individual affects the outcomes of another. We address this issue in Section 6.3.

[4]Note that estimating equation 2 with OLS does not require us to assume a constant treatment effect. The estimated coefficient is simply the average treatment effect.

tools, others of which depend on household or firm responses. This relationship is structural; it holds regardless of the actions of individuals or institutions affected by the policy changes. The impact of any given input in the vector I on academic achievement that is embedded in this relationship is a structural parameter.

Consider a change in one element of the vector I, call it t. One estimate of interest is how changes in t affect Y when all other explanatory variables are held constant, i.e., the partial derivative of Y with respect to t. A second estimate of interest is the total derivative of Y with respect to t, which includes changes in other inputs in response to the change in t. In general, if other inputs are complements to or substitutes for t, then exogenous changes in I will lead to changes in other inputs j. For example, parents may respond to an educational program by increasing their provision of home-supplied educational inputs. Alternatively, parents may consider the program a substitute for home-supplied inputs and decrease their supply. For example, Das, Krishnan, Habyarimana, and Dercon (2004) and others suggest that household educational expenditures and governmental non-salary cash grants to schools are substitutes, and that households cut back on expenditures when the government provides grants to schools.

In general, the partial and total derivatives could be quite different, and both may be of interest to policymakers. The total derivative is of interest because it shows what will happen to outcome measures after an input is exogenously provided and agents re-optimize. In effect it tells us the "real" impact of the policy on the outcomes of interest. But the total derivative may not provide a measure of overall welfare effects. Again consider a policy of providing textbooks to students where parents may respond to the policy by reducing home purchases of textbooks in favor of some consumer good that is not in the educational production function. The total derivative of test scores or other educational outcome variables will not capture the benefits of this re-optimization. Under some assumptions, however, the partial derivative will provide an appropriate guide to the welfare impact of the input.

Results from randomized evaluations (and from other internally valid program evaluations) provide reduced form estimates of the impacts of the treatment, and these reduced form parameters are total derivatives. Partial derivatives can only be obtained if researchers specify the model that links various inputs to the outcomes of interest and collect data on these intermediate inputs. This underscores that to estimate welfare impact of a policy, randomization needs to be combined with theory, a topic to which we return in section 8.

2.3 Other Methods to Control for Selection Bias

Aside from randomization, other methods can be used to address the issue of selection bias. The objective of any of these methods is to create comparison groups that are valid under a set of identifying assumptions. The identifying assumptions are not directly testable, and the validity of any particular study depends instead on how convincing the assumptions appear. While it is not the objective of this chapter to review these methods in detail,[5] in this section we discuss them briefly in relation to randomized evaluations.[6]

2.3.1 Controlling for Selection Bias by Controlling for Observables

The first possibility is that, conditional on a set of observable variables X, the treatment can be considered to be as good as randomly assigned. That is, there exists a vector X such that

$$E[Y_i^C|X,T] - E[Y_i^C|X,C] = 0. \tag{3}$$

A case where this is obviously true is when the treatment status is randomly assigned conditional on a set of observable variables X. In other words, the allocation of observations to treatment or comparison is not unconditionally randomized, but within each strata defined by the interactions of the variables in the set X, the allocation was done randomly. In this case, after conditioning on X, the selection bias disappears. We will discuss in section 6.1 how to analyze the data arising from such a set-up. In most observational settings, however, there is no explicit randomization at any point, and one must *assume* that appropriately controlling for the observable variables is sufficient to eliminate selection bias.

There are different approaches to control for the set of variables X. A first approach, when the dimension of X is not too large, is to compute the difference between the outcomes of the treatment and comparison groups within each cell formed by the various possible values of X. The treatment effect is then the weighted average of these within-cell effects (see Angrist (1998) for an application of this method to the impact of military service). This approach (fully non-

[5] Much fuller treatments of these subjects can be found, notably in this and other handbooks (Angrist and Imbens 1994, Card 1999, Imbens 2004, Todd 2006, Ravallion 2006).

[6] We do not discuss instrumental variables estimation in this section, since its uses in the context of randomized evaluation will be discussed in section 6.2, and the general principle discussed there will apply to the use of instruments that are not randomly assigned.

parametric matching) is not practical if X has many variables or includes continuous variables. In this case, methods have been designed to implement matching based on the "propensity score," or the probability of being assigned to the treatment conditional on the variables X.[7] A third approach is to control for X, parametrically or non-parametrically, in a regression framework. As described in the references cited, matching and regression techniques make different assumptions and estimate somewhat different parameters. Both, however, are only valid on the underlying assumption that, conditional on the observable variables that are controlled for, there is no difference in potential outcomes between treated and untreated individuals. For this to be true, the set of variables X must contain all the relevant differences between the treatment and control groups. This assumption is not testable and its plausibility must be evaluated on a case-by-case basis. In many situations, the variables that are controlled for are just those that happen to be available in the data set, and selection (or "omitted variable") bias remains an issue, regardless of how flexibly the control variables are introduced.

2.3.2 Regression Discontinuity Design Estimates

A very interesting special case of controlling for an observable variable occurs in circumstances where the probability of assignment to the treatment group is a discontinuous function of one or more observable variables. For example, a microcredit organization may limit eligibility for loans to women living in household with less than one acre of land; students may pass an exam if their grade is at least 50%; or class size may not be allowed to exceed 25 students. If the impact of any unobservable variable correlated with the variable used to assign treatment is smooth, the following assumption is reasonable for a small ϵ:

$$E[Y_i^C|T, X < \overline{X} + \epsilon, X > \overline{X} - \epsilon] = E[Y_i^C|C, X < \overline{X} + \epsilon, X > \overline{X} - \epsilon], \tag{4}$$

where X is the underlying variable and $\overline{X}$ is the threshold for assignment. This assumption implies that within some ε-range of $\bar{X}$, the selection bias is zero and is the basis of "regression discontinuity design estimates" (Campbell 1969); see Todd (2006) chapter in this volume for

[7] The results that controlling for the propensity score leads to unbiased estimate of the treatment effect under assumption 3 is due to Rosenbaum and Rubin (1983) see chapters by Todd (2006) and Ravallion (2006) in this volume for a discussion of matching).

further details and references). The idea is to estimate the treatment effect using individuals just below the threshold as a control for those just above.

This design has become very popular with researchers working on program evaluation in developed countries, and many argue that it removes selection bias when assignment rules are indeed implemented. It has been less frequently applied by development economists, perhaps because it faces two obstacles that are prevalent in developing countries. First, assignment rules are not always implemented very strictly. For example, Morduch (1998) criticizes the approach of Pitt and Khandker (1998), who make implicit use of a regression discontinuity design argument for the evaluation of Grameen Bank clients. Morduch shows that despite the official rule of not lending to household owning more than one acre of land, credit officers exercise their discretion. There is no discontinuity in the probability of borrowing at the one acre threshold. The second problem is the officials implementing a program may be able to manipulate the level of the underlying variable that determines eligibility, which makes an individual's position above or below the threshold endogenous. In this case, it cannot be argued that individuals on either side of the cutoff have similar potential outcomes and equation (4) fails to hold.

2.3.3 Difference-in-Differences and Fixed Effects

Difference-in-difference estimates use pre-period differences in outcomes between treatment and control group for control for pre-existing differences between the groups, when data exists both before and after the treatment. Denote by Y_1^T (Y_1^C) the potential outcome "if treated" ("if untreated") in period 1, after the treatment occurs, and Y_0^T (Y_0^C) the potential outcome "if treated" ("if untreated") in period 0, before the treatment occurs. Individuals belong to group T or group C. Group T is treated in period 1 and untreated in period 0. Group C is never treated.

The difference-in-differences estimator is

$$\widehat{DD} = [\hat{E}[Y_1^T|T] - \hat{E}[Y_0^C|T]] - [\hat{E}[Y_1^C|C] - \hat{E}[Y_0^C|C]]$$

and provides an unbiased estimate of the treatment effect under the assumption that $[\hat{E}(Y_1^C|T) - \hat{E}(Y_0^C|T)] = [\hat{E}(Y_1^C|C) - \hat{E}(Y_0^C|C)]$, i.e., that absent the treatment the outcomes in the two groups would have followed parallel trends.

Fixed effects generalizes difference-in-differences estimates when there is more than one time period or more than one treatment group. The fixed effects estimates are obtained by regressing the outcome on the control variable, after controlling for year and group dummies. Both difference-in-differences and fixed effect estimates are very common in applied work. Whether or not they are convincing depends on whether the assumption of parallel evolution of the outcomes in the absence of the treatment is convincing. Note in particular that if the two groups have very different outcomes before the treatment, the functional form chosen for how outcomes evolved over time will have an important influence on the results.

2.4 Comparing Experimental and Non-Experimental Estimates

A growing literature is taking advantage of randomized evaluation to estimate a program's impact using both experimental and non-experimental methods and then test whether the non-experimental estimates are biased in this particular case. LaLonde's seminal study found that many of the econometric procedures and comparison groups used in program evaluations did not yield accurate or precise estimates and that such econometric estimates often differ significantly from experimental results (Lalonde 1986). A number of subsequent studies have conducted such analysis focusing on the performance of propensity score matching (Heckman, Ichimura, and Todd 1997, Heckman, Ichimura, Smith, and Todd 1998, Heckman, Ichimura, and Todd 1998, Dehejia and Wahba 1999, Smith and Todd 2005). Results are mixed, with some studies finding that non-experimental methods can replicate experimental results quite well and others being more negative. A more comprehensive review by Glazerman, Levy, and Myers (2003) compared experimental non-experimental methods in studies of welfare, job training, and employment service programs in the United States. Synthesizing the results of twelve design replication studies, they found that retrospective estimators often produce results dramatically different from randomized evaluations and that the bias is often large. They were unable to identify any strategy that could consistently remove bias and still answer a well-defined question.

Cook, Shadish, and Wong (2006) conducts a comparison of randomized and non-randomized studies, most of which were implemented in educational settings, and arrives at a more nuanced conclusion. He finds that experimental and non-experimental results are similar when the non-experiment technique is a regression discontinuity or "interrupted time series" design (difference-in-differences with long series of pre-data), but that matching or other ways to control for

observables does not produce similar results. He concludes that well designed quasi-experiments (regression discontinuity designs in particular) may produce results that are as convincing as those of a well-implemented randomized evaluation but that "You cannot put right by statistics what you have done wrong by design." While Cook's findings are extremely interesting, the level of control achieved by the quasi-experiments he reviews (in terms, for example, of strictly following threshold rules) is such that for developing countries these designs may actually be less practical than randomized evaluations.

We are not aware of any systematic review of similar studies in developing countries, but a number of comparative studies have been conducted. Some suggest omitted variables bias is a significant problem; others find that non-experimental estimators may perform well in certain contexts. Buddlemeyer and Skofias (2003) and Diaz and Handa (2006) both focus on PROGRESA, a poverty alleviation program implemented in Mexico in the late 1990s with a randomized design. Buddlemeyer and Skofias (2003) use randomized evaluation results as the benchmark to examine the performance of regression discontinuity design. They find the performance of such a design to be good, suggesting that if policy discontinuities are rigorously enforced, regression discontinuity design frameworks can be useful. Diaz and Handa (2006) compare experimental estimates to propensity score matching estimates, again using the PROGRESA data. Their results suggest that propensity score matching does well when a large number of control variables is available.

In contrast, several studies in Kenya find that estimates from prospective randomized evaluations can often be quite different from those obtained using a retrospective evaluation in the same sample, suggesting that omitted variable bias is a serious concern. Glewwe, Kremer, Moulin, and Zitzewitz (2004) study an NGO program that randomly provided educational flip charts to primary schools in Western Kenya. Their analysis suggests that retrospective estimates seriously overestimate the charts' impact on student test scores. They found that a difference-in-differences approach reduced but did not eliminate this problem.

Miguel and Kremer (2003) and Duflo, Kremer, and Robinson (2006) compare experimental and non-experimental estimate of peer effects in the case of the take up of deworming drug and fertilizer adoption, respectively. Both studies found that the individual's decision is correlated with the decisions of their contacts. However, as Manski (1993) has argued, this could be due to many factors other than peer effects, in particular, to the fact that these individuals share the

same environment. In both cases, randomization provided exogenous variation in the chance that some members of a particular network adopted the innovation (deworming or fertilizer, respectively). The presence of peer effect can then be tested by comparing whether others in the network were then more likely to adopt as well (we will come back to this specific method of evaluating peer effects in section 6.3). Both studies find markedly different results from the non-experimental results: Duflo, Kremer, and Robinson (2006) find no learning effect, while Miguel and Kremer (2003) find *negative* peer effects. Furthermore, Miguel and Kremer (2003) run on the non-experimental data a number of specifications checks that have been suggested in the peer effects literature, and all these checks support the conclusion that peer effects are in fact positive, suggestive that such checks may not be sufficient to erase the specification bias.

Future research along these lines would be valuable, since comparative studies can be used to assess the size and prevalence of biases in retrospective estimates and provide more guidance which methodologies are the most robust. However, as discussed below, these types of studies have to be done with care in order to provide an accurate comparison between different methods. If the retrospective portions of these comparative studies are done with knowledge of the experimental results, there is a natural tendency to select from plausible comparison groups and methodologies in order to match experimental estimates. To address these concerns, future researchers should conduct retrospective evaluations before the results of randomized evaluations are released or conduct blind retrospective evaluations without knowledge of the results of randomized evaluations or other retrospective studies.

2.5 Publication Bias

2.5.1 Publication bias in non-experimental studies

Uncertainty over bias in the reported results from non-experimental studies is compounded by publication bias, which occurs when editors, reviewers, researchers, or policymakers have a preference for results that are statistically significant or support a certain view. In many cases, as we just reviewed, researchers will have many possible choices of how to specify empirical models, and many of these might still be subject to remaining omitted variable bias.

Consider an example in which the true treatment effect is zero, but each non-experimental technique yields an estimated treatment effect equal to the true effect plus an omitted variable

bias term that is itself a normally distributed random variable with mean zero. Unfortunately, what appears in the published literature may not reflect typical results from plausible specifications, which would be centered on zero, but may instead be systematically biased.

In any study, a number of choices will need to be made about how the analysis would most appropriately be conducted, which method should be used, which control variable should be introduced, which instrumental variable to use. There will often be legitimate arguments for a variety of different alternatives. Researchers focus their time and effort on completing studies that seem to produce statistically significant results, those that do not end up in the "file drawer."

Some researchers may inappropriately mine various regression specifications for one that produces statistically significant results. Even researchers who do not deliberately search among specifications for those that yield significant results may do so inadvertently. Consider a researcher undertaking a retrospective study for which there are several potentially appropriate specifications, not all of which are known to the researcher before beginning the analysis. The researcher thinks of one specification and runs a series of regressions. If the results are statistically significant and confirm what the researcher expected to find, it is perhaps likely that he or she will assume the specification is appropriate and not spend much time considering possible alternatives. However, if the regression results are not statistically significant or go against what the researcher expected to find, he or she is perhaps more likely to spend substantial amounts of time considering other possible specifications. In addition to generating too many false positives in published papers, this type of specification searching probably leads to under-rejection of commonly held views.

Even if a researcher introduces no bias of this kind, the selection by journals of papers with significant results introduces another level of publication bias. Moreover, citation of papers with extreme results by advocates on one side or another of policy debates is likely to compound publication bias with citation bias. The cumulative result of this process is that even in cases when a program has no effect, strongly positive and/or strongly negative estimates are likely to be published and widely cited.

A growing body of available evidence suggests publication bias is a serious problem within the economics literature. DeLong and Lang (1992) devise a test to determine the fraction of un-rejected null hypotheses that are false. They note that under the null, the distribution of test

statistics is known; their cumulative distributions are given by the marginal significance levels associated with them. For example, any test statistic has a 5% chance of falling below the value of the .05-significance level. They use this observation to examine whether the distribution of these test statistics conform to what would be expected if a pre-specified fraction of the null hypotheses were in fact true. Using data from articles published in major economics journals, the authors find they can reject at the .05 level the null hypothesis that more than about one-third of un-rejected null hypotheses were true. Although the authors acknowledge several potential explanations for their findings, they argue publication bias provides the most important explanation.

Statisticians have developed a meta-analysis framework to examine whether inferences are sensitive to publication bias. Hedges (1992) proposed a formal model of publication bias where tests that yield lower p-values are more likely to be observed. Ashenfelter, Harmon, and Oosterbeek (1999) apply Hedges's analytic framework to the literature on rates of return to education and find strong evidence that publication bias exists for instrumental variables (IV) estimates on the rate of return to education, which suggests that the often cited results that IV estimates of returns to education are larger than OLS estimates may just be an artifact of publication bias. Likewise, Card and Krueger (1995) find evidence of significant publication bias in the time-series minimum wage literature, leading to an over-reporting of significant results.

2.5.2 Randomization and publication bias

Some, though not all, of the problems of publication bias can be addressed by randomized evaluations. First, if a randomized evaluation is correctly implemented, there can be no question that the results, whatever they are, give us the impact of the particular intervention that was tested (subject to a known degree of sampling error). This implies that if results are unexpected, they have less chance of being considered the result of specification error and discarded. Miguel and Kremer (2003) evaluation of the impact of network and peer effects on the take up of deworming medicine, which we discussed above, provides an interesting example. Ex ante, the researchers probably expected children with more links to students in treatment schools to increase their uptake of deworming treatment in subsequent rounds of the program as they learned about the medicine's benefits. Instead, their experimental findings showed a significant effect in the opposite direction. Had they obtained these results in a retrospective study, most

researchers would have assumed there was a problem with their data or specification and explored alternative specifications. The experimental design leaves little doubt that peer effects are in fact negative.

Second, in randomized evaluation the treatment and comparison groups are determined before a researcher knows how these choices will affect the results, limiting room for ex post discretion. There is usually still some flexibility ex post—including what variables to control for, how to handle subgroups, and how to deal with large numbers of possible outcome variables—which may lead to "cherry-picking" amongst the results. However, unlike omitted variable bias which can become arbitrarily large, this ex post discretion is bounded by the ex ante design choices.

Nevertheless, "cherry-picking" particular outcomes, sites, or subgroups where the evaluation concluded that the program is effective provides a window through which publication biases can appear in randomized evaluations. This is why the FDA does not consider results from sub-group analysis in medical trials valid evidence for the effectiveness of a drug. Below, we discuss how to handle these issues.

Third, randomized evaluations can also partially overcome the file drawer and journal publication biases as their results are usually documented even if they suggest insignificant effects. Even when unpublished, they are typically circulated and often discussed in systematic reviews. This is because researchers discover whether the results are significant at a much later stage and are much less likely to simply abandon the results of an evaluation that has taken several years to conduct than they are to abandon the results of a quick dip into existing data to check out an idea. In addition, funding agencies typically require a report of how their money was spent regardless of outcomes.

Despite this, it would still be extremely useful to put institutions in place to ensure that negative as well as positive results of randomized evaluations are disseminated. Such a system is in place for medical trial results, and creating a similar system for documenting evaluations of social programs would help to alleviate the problem of publication bias. One way to help maintain such a system would be for grant-awarding institutions to require researchers to submit results from all evaluations to a centralized database. To avoid problems due to specification searching ex post, such a database should also include the salient features of the ex ante design (outcome variables to be examined, sub-groups to be considered, etc.) and researchers should

have to report results arising from this main design. While researchers could still report results other than those that they initially had included in their proposal, they should be clearly signposted such that users should be able to distinguish between them. This information would be useful from a decision-theoretic perspective—several positive, unbiased estimates that are individually statistically insignificant can produce significant meta-estimates—and be available to policymakers and those seeking to understand the body of experimental knowledge regarding the effectiveness of certain policies.

3 Incorporating Randomized Evaluation in a Research Design

In the rest of this chapter, we discuss how randomized evaluations can be carried out in practice. In this section, we focus on how researchers can introduce randomization in field research in developing countries. Perhaps the most widely used model of randomized research is that of clinical trials conducted by researchers working in laboratory conditions or with close supervision. While there are examples of research following similar templates in developing countries,[8] most of the projects involving randomization differ from this model in several important ways. They are in general conducted with implementation partners (governments, NGOs, or private companies) who are implementing real-world programs and are interested in finding out whether they work or how to improve them. In some cases, randomization can then be included during pilot projects, in which one or several version of a program are tried out. There are also cases where randomization can be included outside pilot projects, with minimal disruption to how the program is run, allowing the evaluation of on-going projects.

Section 3.1 discusses the possible partners for evaluation. Section 3.2 discusses how randomization can be introduced in the context of pilot projects and how these pilot projects have the potential to go beyond estimating only program effects to test specific economic hypotheses. Section 3.3 discuses how randomization can be introduced outside pilot projects.

[8]The best example is probably the "Work and Iron Supplementation Experiment" (see Thomas, Frankenberg, Friedman, Habicht, and Al (2003)), where households were randomly assigned to receive either iron supplementation or a placebo for a year, and where compliance with the treatment was strictly enforced.

3.1 Partners

Unlike conducting laboratory experiments, which economists can do on their own, introducing randomization in real-world programs almost always requires working with a partner who is in charge of actually implementing the program.

Governments are possible partners. While government programs are meant to serve the entire eligible population, pilots programs are some times run before the programs are scaled-up. These programs are limited in scope and can sometimes be evaluated using a randomized design. Some of the most well-known early social experiments in the US —for example the Job Partnership Training Act and the Negative Income Tax—were conducted following this model.

There are also a few examples from developing countries. PROGRESA (now called Oportunidades) (also discussed in Todd (2006) and Parker, Rubalcava, and Teruel (2005) chapters in this volume) is probably the best known example of a randomized evaluation conducted by a government. The program offers grants, distributed to women, conditional on children's school attendance and preventative health measures (nutrition supplementation, health care visits, and participation in health education programs). In 1998, when the program was launched, officials in the Mexican government made a conscious decision to take advantage of the fact that budgetary constraints made it impossible to reach the 50,000 potential beneficiary communities of PROGRESA all at once, and instead started with a randomized pilot program in 506 communities. Half of those were randomly selected to receive the program, and baseline and subsequent data were collected in the remaining communities.

The task of evaluating the program was given to academic researchers through the International Food Policy Research Institute. The data was made accessible to many different people, and a number of papers have been written on its impact (most of them are accessible on the IFPRI Web site). The evaluations showed that it was effective in improving health and education. (see in particular Gertler and Boyce (2001) and Schultz (2004)).

The PROGRESA pilot had an impressive demonstration effect. The program was continued and expanded in Mexico despite the subsequent change in government, and expanded in many other Latin American countries, often including a randomized pilot component. Some examples are the Family Allowance Program (PRAF) in Honduras (International Food Policy Research 2000), a conditional cash transfer program in Nicaragua (Maluccio and Flores 2005), a condi-

tional cash transfer program in Ecuador (Schady and Araujo 2006), and the Bolsa Alimentação program in Brazil.

These types of government-sponsored (or organized) pilots are becoming more frequent in developing countries than they once were. For example, such government pilot programs have been conducted in Cambodia (Bloom, Bhushan, Clingingsmith, Hung, King, Kremer, Loevinsohn, and Schwartz 2006), where the impact of public-private partnership on the quality of health care was evaluated in a randomized evaluation performed at the district levels. In some cases, governments and researchers have worked closely on the designs of such pilot. In Indonesia, Olken (2005) collaborated with the World Bank and the government to design an experiment where different ways to fight corruption in locally administered development projects were tried in different villages. In Rajasthan, India, the police department is working with researchers to pilot a number of reforms to improve police performance and limit corruption across randomly selected police stations in nine districts.

Randomized evaluations conducted in collaboration with governments are still relatively rare. They require cooperation at high political levels, and it is often difficult to generate the consensus required for successful implementation. The recent spread of randomized evaluations in development owes much to a move towards working with non-governmental organizations (NGOs). Unlike governments, NGOs are not expected to serve entire populations, and even small organizations can substantially affect budgets for households, schools, or health clinics in developing countries. Many development-focused NGOs frequently seek out new, innovative projects and are eager to work with researchers to test new programs or to assess the effectiveness of existing operations. In recent years, numerous randomized evaluations have been conducted with such NGOs, often with evaluation-specific sponsorship from research organizations or foundations. A number of examples are covered throughout this chapter.

Finally, for-profit firms have also started getting interested in randomized evaluations, often with the goal of understanding better how their businesses work and thus to serve their clients better and increase their profits. For example, Karlan and Zinman (2006b), Karlan and Zinman (2005a), Karlan and Zinman (2006a) and Bertrand, Karlan, Mullainathan, Shafir, and Zinman (2005) worked with a private consumer lender in South Africa. Many micro-finance organizations are now working with researchers to understand the impact of the salient features of their

products or to design products that serve their clients better.[9]

What are the advantages of different partners? As mentioned, NGOs are more willing to want to partner on an evaluation, so they are often the only partner of choice. When possible, working with governments offers several advantages. First, it can allow for much wider geographic scope. Second, the results may be more likely to feed into the policy process. Third, there will be less concern about whether the results are dependent on a particular (and impossible to replicate) organizational culture. However, NGOs and firms offer a much more flexible environment, where it can be easier for the researchers to monitor the implementation of the research design. One of the main benefits of the move to working with NGOs has been an increased scope for testing a wider range of questions, implementing innovative programs, and allowing for greater input from researchers into the design of programs, especially in the pilot stage.

3.2 Pilot projects: From program evaluations to field experiments

A natural window to introduce randomization is before the program is scaled up, during the pilot phase. This is an occasion for the implementation partner to rigorously assess test the effectiveness of the program and can also be a chance to improve its design.

Many early field randomized studies (both in the US and in developing countries) simply sought to test the effectiveness of particular programs. For example, the PROGRESA pilot program implemented the program in the treatment villages and did not introduce it in the comparison villages. The evaluation is only able to say whether, taken together, all the components of PROGRESA are effective in increasing health and education outcomes. They cannot disentangle the various mechanisms at play without further assumptions.

Such evaluations are still very useful in measuring the impact of policies or programs before they are scaled up, but increasingly, researchers have been using such pilot programs to go beyond the simple question of whether a particular program works or not. They have helped their partners design programs or interventions with specific theories in mind. The programs are designed to help solve the partner's practical problems but they also serve as the test for the theories. A parallel movement has also happened in developed countries, fuelled by the concerns on the external validity of laboratory experiments (see Harrison and List (2004) for a review of this literature).

[9] See the research coordinated by the Centre for Micro Finance in India, for many examples.

These pilot programs allow standard "program evaluation" to transform itself into "field experiments", in the sense that both the implementing agency and the researchers are experimenting together to find the best solution to a problem (Duflo 2006). There is an explosion of such work in development economics.[10] In practice, the distinction between "simple" program evaluation and field experiment is of course too stark: there is a wide spectrum of work ranging from the most straightforward comparison between one treatment and one comparison group for a program, to evaluation involving a large number of groups allowing researchers to test very subtle hypotheses in the field.

Here we mention only two examples which illustrate well the power of creative designs. Ashraf, Karlan, and Yin (2006) set out to test the importance of time-inconsistent ("hyperbolic") preferences. To this end they designed a commitment savings product for a small rural bank in the Philippines. The rural bank was interested in participating in a program that had the potential to increase savings. Individuals could restrict the access to the funds they deposited in the accounts until either a given maturity or a given amount of money was saved. Relative to standard accounts, the accounts carried no advantage other than this feature. The product was offered to a randomly selected half of 1,700 former clients of the bank. The other half of the individuals were assigned either to a pure comparison group or to a group who was visited and given a speech reminding them of the importance of savings. This group allowed them to test whether the simple fact of discussing savings is what encourages clients to save, rather than the availability of a time-commitment device. Having these two separate groups was possible only in the context of a pilot with a relatively flexible organization.

Duflo, Kremer, and Robinson (2006) evaluated a series of different interventions to understand the determinants of the adoption of fertilizer in Western Kenya. The design of the interventions made it possible to test some of the standard hypotheses of the hindrance to the adoption of new technology. Field demonstrations with treatment and control plots were conducted to evaluate the profitability of fertilizer in the local conditions. Because the farmers were randomly selected, those field trials also allowed them to study the impact of information provision and the channels of information transmission; other ways to provide information (starter kits, school based demonstrations) were also examined; finally, financing constraints and difficulties in saving were also explored, with interventions helping farmers to buy fertilizer at the

[10] Many of these very recent or on going studies are reviewed in the review articles mentioned in the introduction.

time when they have most money in the field.

3.3 Alternative Methods of Randomization

The examples we have looked at so far are in one respect similar to classic clinical trials: in all cases the randomized study was introduced concurrent with a new program and where the sample was randomly allocated into one or more treatment groups and a comparison group that never received the treatment. However, one of the innovations of recent work is to realize that there are many different ways to introduce an element of randomization into programs. It is often possible to introduce randomization into existing programs with minimal disruption. This has spurred rapid growth in the use of randomized studies by development economists over the last ten years. In this section we run through the four key methods—oversubscription, phase-in, within-group randomization, and encouragement design—for introducing randomization into new and existing programs.

3.3.1 Oversubscription Method

A natural opportunity for introducing randomization occurs when there are limited resources or implementation capacity and demand for a program or service exceeds supply. In this case, a natural and fair way to ration resources is to select those who will receive the program by lottery among eligible candidates.

Such a method was used to ration the allocations of school vouchers in Colombia and the resulting randomization of treatment and comparison groups allowed for a study to accurately assess the impact of the voucher program (Angrist, Bettinger, Bloom, King, and Kremer) evaluated the impact of expanded consumer credit in South Africa by working with a lender who randomly approved some marginal loan applications that would normally have been rejected. All applicants who would normally have been approved received loans, and those who were well below the cutoff were rejected. Such an evaluation was only possible because the experimental design caused minimal disruption to the bank's normal business activities. When interpreting these results, one must be careful to keep in mind that they only apply to those "marginal" borrowers or, more generally, to the population over which assignment to the treatment group was truly random.

3.3.2 Randomized Order of Phase-In

Financial and administrative constraints often lead NGOs to phase-in programs over time, and randomization will often be the fairest way of determining the order of phase-in. Randomizing the order of phase-in can allow evaluation of program effects in contexts where it is not acceptable for some groups or individuals to receive no support. In practical terms, it can facilitate continued cooperation by groups or individuals that have randomly been selected as the comparison group. As such, where logistics permit, randomization of phase-in may be preferable to a pure lottery because the expectation of future benefits provides subjects an incentive to maintain contact with researchers and thus alleviates issues associated with attrition (see section 6.4).

The Primary School Deworming Project provides an example of this type of randomized phase-in trial (Miguel and Kremer 2004). This program provided medical treatment for intestinal worms (helminths) and schistosomiasis as well as worm-prevention health education lessons to children in 75 primary schools in rural Busia district, Kenya during 1998-2002. The program randomly divided the schools into three groups, each consisting of 25 primary schools. Treatment in the schools was done as follows: 25 Group 1 schools began receiving treatment in 1998; 25 Group 2 schools began receiving treatment in 1999; and 25 Group 3 schools began receiving treatment in 2000. The impact of the program on the health, nutrition, and education of the children was evaluated by comparing the results from group 1 schools in 1998 with group 2 and 3 acting as comparisons and the results of group 1 and 2 schools in 1999 with group 3 schools acting as a comparison. The researchers found that deworming led to improved health and increased school participation.

One drawback of randomized phase-in designs is that they often prevent researchers from estimating a program's long-run effects; however, when a program targets well-identified cohorts, this is still possible. In the deworming program, for example, children were not eligible for treatment after they left school. Taking advantage of this, Miguel and Kremer are following the cohorts who "missed out" on the program because they were too old to receive treatment when it was phased-in to their school. These cohorts provide a valid control group to study the long-run effects of the program on the equivalent cohorts from treated schools.

In contrast, if a randomized phase-in is too rapid relative to the time it takes for program

effects to materialize, it will be impossible to detect treatment effects at all. For example, one would be unlikely to detect the effect of a microcredit program that was phased-in to control villages only six months after it was introduced to the treatment group. When planning a phase-in design, the time between phases should be sufficient to encompass any treatment lag.

Randomized phase-in becomes problematic when the comparison group is affected by the expectation of future treatment. For example, in the case of a phased in microcredit program, individuals in the comparison groups may delay investing in anticipation of cheaper credit once they have access to the program. In this case, the comparison group is also affected by its participation in the experiment and does not provide a valid counterfactual. Some have argued that this may have been at play in the case of the PROGRESA experiment.

3.3.3 Within-Group Randomization

Even a randomized phase-in may not spread the benefits sufficiently smoothly across the whole group to ensure good cooperation with the study. For example, schools may refuse to let researchers collect test scores on their students while the schools do not benefit from participating in the study. In this case, it is still possible to introduce an element of randomization by providing the program to some subgroups in each area.

The evaluation of the balsakhi program, a remedial education assistance in poor urban schools in India provided by Pratham, an Indian education NGO (Banerjee, Duflo, Cole, and Linden 2007) provides an example. The program was designed to provide those children falling behind in school the basic skills they need to learn effectively. Pratham hires and trains tutors, referred to as *balsakhi* or "child's friend," to give remedial math and reading comprehension instruction to children. To ensure cooperation from school authorities, every school in the study received a balsakhi in every year. However, based on random assignment, some schools were asked to use the balsakhi in grade 3 and others in grade 4.[11]

This design was deemed fair by school teachers, since all schools received the same assistance. Further more, since the NGO could make a credible case that they could not provide more than one balsakhi per school, there was no expectation that all children in a school should benefit from the program.

[11] Glewwe, Kremer, and Moulin (2004) used a similar approach in their study of a program to give textbooks to schools in Kenya.

The drawback of such designs is that they increase the likelihood that the comparison group is contaminated. For example, in the balsakhi program, one may have been worried that headmasters reallocated resources from grade 3 to grade 4 if grade 3 got a balsakhi but grade 4 did not. In this particular application, such contamination was unlikely because schools have a fixed number of teachers per grade and few other resources to reallocate. But this risk needs to be considered when deciding whether or not to adopt such a design.

3.3.4 Encouragement Designs

Encouragement designs allow researchers to evaluate the impact of a program that is available in the entire study area but whose take up is not universal. They are particularly useful for evaluating programs over which randomization of access is not feasible for ethical or practical reasons. Rather than randomize over the treatment itself, researchers randomly assign subjects an encouragement to receive the treatment. One of the early encouragement designs was a study of whether studying for the GRE could lead to an increase in test scores (Holland 1988). While studying is available to everyone, researchers increased the number of students who studied for it by mailing out free materials to a randomly selected set of GRE candidates. More recently, Duflo and Saez (2003) studied the impact of receiving information about tax deferred accounts (TDA) by providing financial incentives to university employees to attend the session organized by their university (to which everybody is invited). The incentive increased the fraction of people who chose to attend the session in the group where it was sent, and the TDA adoption of those individuals can then be followed over time and compared to that of groups that did not receive the incentive.

In Kenya, Duflo, Kremer, and Robinson (2006) evaluated the impact of witnessing fertilizer demonstration on another farmer's plot on future adoption of fertilizer by farmers. To do so, they set up fertilizer demonstrations on randomly selected farmers' plots and then explicitly invited a randomly selected subset of the farmers' friends to view the demonstration. While a farmer's other friends were also welcome to come, the fraction who attended was much larger among those "invited" than "not invited." Since the invitation was randomly assigned, it provides a natural instrumental variable with which to evaluate the impact of the treatment.

Because they only *increase* the probability that a treatment is received without changing it from zero to one, encouragement designs pose specific analytical challenges. We discuss the

analytical requirements of this approach in section 6.2.

4 Sample size, design, and the power of experiments

The power of the design is the probability that, for a given effect size and a given statistical significance level, we will be able to reject the hypothesis of zero effect. Sample sizes, as well as other design choices, will affect the power of an experiment.

This section does not intend to provide a full treatment of the question of statistical power or the theory of the design of experiment.[12] Rather, its objective is to draw attention on the key factors that influence the statistical power of randomized evaluations. It presumes a basic knowledge of statistics and ignores some of the more complicated or subtle issues. We first review basic principles of power calculations. We then discuss the influence of design factors such as multiple treatment groups, randomization at the group level, partial compliance, control variables, and stratification. Finally, we discuss the practical steps involved in making power calculations, and the roles they should be given when planning evaluations.

4.1 Basic Principles

The basic principles of power calculation can be illustrated in a simple regression framework. As we discussed above, the difference in sample means for two groups (our estimate of the average treatment effect) is the OLS coefficient of β in the regression

$$Y_i = \alpha + \beta T + \epsilon_i \,. \tag{5}$$

Assume that there is only one possible treatment, and that a proportion P of the sample is treated. Assume for now that each individual was randomly sampled from an identical population, so that observations can be assumed to be i.i.d., with variance σ^2.

The variance of $\hat{\beta}$, the OLS estimator of β, is given by

$$\frac{1}{P(1-P)}\frac{\sigma^2}{N} \tag{6}$$

[12] A good reference for power calculations is Bloom (1995). An good reference on the theory of design of experiments is Cox and Reid (2000).

We are generally interested in testing the hypothesis, H_0, that the effect of the program is equal to zero against the alternative that it is not.[13] The *significance level*, or size, of a test represents the probability of a type I error, i.e., the probability we reject the hypothesis when it is in fact true.

[Insert fig. 1 about here]

The bell shape picture on the left in figure 1 is the distribution of $\hat{\beta}$ under the null hypothesis of no effect.[14] For a given significance level, H_0 will be rejected if $\hat{\beta}$ falls to the right of the critical level, that is if $|\hat{\beta}| > t_\alpha * SE_{\hat{\beta}}$, where t_α depends on the significance level ($t_{\alpha/2}$ for a two sided test) and is obtained from a standard t-distribution.

In figure 1, the curve to the right shows the distribution of $\hat{\beta}$ if the true impact is β. The *power* of the test for a true effect size β is the fraction of the area under this curve that falls to the right of the critical value t_α, i.e., the probability that we reject H_0 when it is in fact false.

To achieve a power κ, it must therefore be that

$$\beta > (t_{1-\kappa} + t_\alpha) SE(\hat{\beta})$$

where $t_{1-\kappa}$ is again given by a t-table. For example, for a power of 80%, $t_{1-\kappa} = 0.84$.

The *minimum detectable effect size* for a given power (κ), significance level (α), sample size (N), and portion of subjects allocated to treatment group (P) is therefore given by

$$MDE = \left(t_{(1-\kappa)} + t_\alpha\right) * \sqrt{\frac{1}{P(1-P)}} \sqrt{\frac{\sigma^2}{N}} \tag{7}$$

for a single sided test (t_α is replaced by $t_{\alpha/2}$ for a two-sided test). Alternatively, equation (7) implicitly defines the sample size N required to achieve a given power, given the effect size that is posited and the level of significance chosen.

Equation (7) shows that there is a trade-off between power and size. When the size decreases, t_α increases, so that the minimum effect size increases for a given level of power. Thus there is trade-off between the probability of falsely concluding that the program has an effect when it

[13] Note that in some cases, it may be interesting and important to evaluate programs that researchers *do not expect* will have a large effect (for example to counteract a policy fad). The power calculation should then be run not to test $H_0 = 0$ (since failing to reject that a program has zero effect does not mean "accepting" that the program has zero effect), but to test that the effect of the program is no larger than some number.

[14] The exposition here follows Bloom (1995).

does not and the probability of falsely concluding that it has no effect when it does. The other parameters that are relevant for this basic power calculation are the minimum effect size the researcher wants to be able to detect, the standard deviation of ϵ, the proportion of the sample that is allocated to the treatment and comparison groups, and the sample size.

Equation (7) also provides some guidance on how to divide the sample between the treatment and the comparison group. With one treatment, if the main cost of the evaluation is data collection, it shows that an equal division between treatment and comparison group is optimal, since the equation is minimized at $P = 0.5$. However, when the treatment is expensive and data collection is cheap (for example, when administrative data on the outcome is available for both the treatment and the comparison group), the optimal sample size will have a larger comparison group. More generally the optimal proportion of treated observations can be obtained by minimizing equation 7 under the budget constraint $Nc_d + NPc_t \leq B$, where N is the total sample size, c_c is the unit cost per comparison subject, and c_t is the unit cost per treatment subject (including both data collection and the treatment cost). This gives the following optimal allocation rule:

$$\frac{P}{1-P} = \sqrt{\frac{c_c}{c_t}},$$

that is, the ratio of subjects in the treatment group to those in the comparison should be proportional to the inverse of the square root of their costs.

The logic of equation (7) can be extended to apply to sample size calculations when more than one treatment is evaluated. Suppose an experiment involves two treatments (for example, the evaluation of the SEED commitment savings product program described above had a comparison group, a social marketing group, and a "commitment savings" group). A first possibility is that the researchers are only interested in the contrast between the comparison group and the marketing group on the one hand and the comparison and the commitment savings group on the other hand. In this case, if the researcher puts equal weight on these estimates, he wants to minimize the sum of the minimum detectable effect (MDE) for the two treatments. The optimal allocation thus requires twice as many observations in the comparison than in each treatment group. Of course, the researcher may want to be able to detect a smaller effect in one intervention than in the other, which can be translated by a higher weight put on one MDE than the other. The main point that remains is that the sample size of the comparison group

should be larger.[15]

If, on the other hand, a researcher interested in the contrast between the two treatments (which was the case in the SEED evaluation) they will need a sample size sufficient to be able to detect a difference between the two groups. If the difference between the two treatments is not very large, this may require a larger sample size than the evaluation of either treatment separately would.

4.2 Grouped Errors

Many of the designs we discussed above involve randomizing over groups rather than individuals. In such cases, researchers nevertheless often have access to individual data. For example, in the PROGRESA program, the village was the unit of randomization, but individual data were available.

When analyzing individual data from programs randomized at a group level, it is important to take into account that the error term may not be independent across individuals. People in the same group may be subject to common shocks, which means their outcomes may be correlated. Because treatment status is also uniform within groups, this correlation in the outcomes of interest may be mistakenly be interpreted as an effect of the program. For example, consider a case where among two districts with large populations, all individuals in one district are given a nutritional supplement program, and those in the other district are assigned to comparison group. Now assume that the comparison district suffers a drought. It will not be possible to distinguish the effect of the drought from the effect of the program.

Formally, consider a modified version of equation 4.5 (this treatment follows (Bloom 2005)).

$$Y_{ij} = \alpha + \beta T + \upsilon_j + \omega_{ij} \tag{8}$$

where j indexes the group and ij the individual. For simplicity of exposition, assume there are J clusters of identical size n, υ_j is i.i.d. with variance τ^2, and ω_{ij} is i.i.d. with variance σ^2. The OLS estimator of $\hat{\beta}$ is still unbiased, and its standard error is

[15]The optimal allocation is given by

$$\frac{N_i}{N_j} = \frac{\sum_{H_I} \omega_h}{\sum_{H_J} \omega_h} \sqrt{\frac{c_j}{c_i}}$$

where ω_h is the weight placed on testing hypothesis h, and H_I is the set of all hypotheses including group I.

$$\sqrt{\frac{1}{P(1-P)}}\sqrt{\frac{n\tau^2+\sigma^2}{nJ}}. \tag{9}$$

If the randomization had been conducted at the level of the individual, the standard error of $\hat{\beta}$ would have been

$$\sqrt{\frac{1}{P(1-P)}}\sqrt{\frac{\tau^2+\sigma^2}{nJ}}. \tag{10}$$

Equations (9) and (10) imply that the ratio between the standard errors for group level randomization and for individual level randomization given a fixed number of members per group, the *design effect*, is equal to

$$D = \sqrt{1+(n-1)\rho} \tag{11}$$

where n is the number of individuals in each group and $\rho = \tau^2/(\tau^2+\sigma^2)$ is the intracluster correlation, i.e., the proportion of the overall variance explained by within group variance. As equation (11) shows, the design effect increases with both the intracluster correlation and the number of individuals per group. This effect can be quite large, even for modest values of the intraclass correlation. The standard error will more than double, for example, with group size of 50 and intraclass correlation of 0.06.

This increase in variance has obvious implication for sample sizes calculations. Specifically, Bloom (2005) shows that the MDE with J groups of size n each is given by

$$MDE = \frac{M_{J-2}}{\sqrt{P(1-P)J}}\ \sqrt{\rho+\frac{1-\rho}{n}}\ \sigma \tag{12}$$

where $M_{J-2} = t_{\alpha/2} + t_{1-\kappa}$, for a two sided test.

Equation 12 shows that, ignoring the effect of J on the critical values of the t distribution, the MDE varies roughly proportionally as a function of the number of groups J. On the other hand, the number of observations per group affects precision much less, especially when ρ is relatively large. This implies that, for a given sample size, an increase in the number of individuals sampled per cluster increases the precision much less than increasing the number of clusters being randomized. Intuitively, when group outcomes are correlated, data from another individual in an existing group provides less information than data from the first individual in a new cluster.

Finally, the equation shows that both the total number of clusters to be sampled and the number of people to sample per cluster are very dependent on ρ.

Note that all these calculations were carried out under the assumption of common variances, which may not be the case, though the assumption simplifies the power calculation. In section 7 below we discuss how to compute standard errors with grouped data without making this assumption.

4.3 Imperfect Compliance

We saw a number of cases where the randomized design only influences the *probability* that someone receives a treatment. In section 6.2 below we will discuss in detail how to analyze and interpret data arising from such experiments.

What is important to note here is that the possibility that compliance may not be perfect should be taken into account when determining the optimal sample size. In section 6.2 we will show that the ratio of the difference between the initial treatment and control groups to the difference in the probability of being treated in the two groups is an estimate of the causal effect of the treatment among the compliers (those induced by the randomization to receive the treatment).

The power of the design is therefore going to arise from the difference between the outcomes in those who were *initially* assigned to the treatment and those who were not, irrespective of whether they were treated or not (this is the reduced form effect of the initial assignment). Denote by c the share of subjects initially assigned to the treatment group who actually receive the treatment and by s the share of subjects initially assigned to the comparison group who receive the treatment. The reduced form effect is going to be the actual treatment effect multiplied by $c - s$.

Hence, the minimum detectable treatment effect size accounting for partial compliance is now given by

$$MDE = \left(t_{(1-\kappa)} + t_{\alpha}\right) * \sqrt{\frac{1}{P(1-P)}}\sqrt{\frac{\sigma^2}{N}}\frac{1}{c-s}. \tag{13}$$

Partial compliance thus strongly affects the power of a design, since the MDE increases *linearly* with the compliance rate, while it increases proportionally to the square root of the number of observations. Thus, if there is only an 80% difference in take up between the treatment

and control group, the sample size would have to be 56% larger to achieve the same minimum detectable effect. Alternatively, at the same sample size, the minimum detectable effect would be 25% larger.

This (in addition to the interpretation issues that arise when compliance is imperfect which we discuss in details below) underscores the importance of compliance in experiments. This implies that if a researcher has the choice between two designs with different level of compliance, choosing one which will have the highest compliance can have important implication on sample size. This is useful to think about when to introduce randomization. Suppose for example that one wants to evaluate a voluntary business training program for microcredit clients. A first approach would be an encouragement design, where randomly selected clients would be asked whether they want to participate in the program, and they could choose whether or not to do it. The evaluation would then compare those invited to those who were not invited. A second approach would be an oversubscription design, where the clients would be asked to apply, and the program would be randomized among applicants. The take-up of the program in the second design would presumably be much larger than that in first design. The MDE for the effect of the training on those who choose to participate when offered the option will decrease in proportion.[16]

4.4 Control Variables

In a simple randomized experiment, controlling for baseline values of covariates likely to influence or predict the outcome does not affect the expected value of an estimator of β, but it can reduce its variance. Note that controlling for covariates affected by the treatment would bias the estimate of the treatment effect by capturing part of its impact. Information on covariates should therefore be collected in the baseline surveys. A special case of a covariate of interest is the pre-treatment value of the outcome.

[16] One may worry that the evaluation would be less "representative" in the second sample since the evaluation is carried out in a sample of applicants. This is, however, not quite correct: as we will see below, the right interpretation of the treatment effect in the first design is that it is the effect of the intervention among *compliers*, that is those who get the training if they are selected for it, and not otherwise. Therefore, even in the first design, we will evaluate the effect of the training on those who are interested in getting it. The population that will be affected is not necessarily exactly the same in the two designs: it is of course possible that some people would not think of applying by themselves but would accept to be trained if offered an option. But the difference may not be that large.

Consider the equation:

$$Y_{ij} = \alpha + \beta T + X_{ij}\gamma + \tilde{\upsilon_j} + \tilde{\omega_{ij}}, \tag{14}$$

where X_{ij} is a set of control variables, which can be the group- or individual-levels. $\tilde{\upsilon_j}$ and $\tilde{\omega_{ij}}$ now represent the unexplained variance after controlling for X_{ij}. Ignoring the effect of adding covariates on degrees of freedom, controlling for covariates has three effects on variance estimates. First, it reduces the (true) residual variance and thereby tends to reduce the variance of parameter estimates. Second, in a completely randomized experiment, it *may* increase $(W'W)^{-1}$, where W is the matrix of all covariates including the treatment indicator, and thereby increase the variance of $\hat{\beta}$. Note that this effect is not present with stratification (see below) because stratification ensures that the treatment indicator is orthogonal to the other covariates in practice, whereas in a completely randomized experiment this is only true in expectation. Finally, the estimated variance of $\hat{\beta}$ is noisier than without controlling for covariates. It can be larger or smaller, but it is unbiased.

In general, controlling for variables that have a large effect on the outcome can help reduce standard errors of the estimates and thus the sample size needed. This is a reason why baseline surveys can greatly reduce sample size requirement when the outcome variables are persistent. For example, controlling for baseline test scores in evaluations of education interventions greatly improves the precision of estimates, which reduces the cost of these evaluations when a baseline test can be conducted. Note, however, that controlling for variables that explain little or none of the variation in the outcome will increase standard errors by reducing degrees of freedom.[17] Choosing which variables to control for is therefore a difficult exercise. Note that this choice must in principle be specified in advance to avoid the risk of specification searching.

4.5 Stratification

Since the covariates to be used must be chosen in advance in order to avoid specification searching and data mining, they can be used to stratify (or *block*) the sample in order to improve the precision of estimates. This technique (first proposed by Fisher (1926)) involves dividing

[17] Even controlling for the pre-treatment value of the outcome may reduce precision if the outcome is not highly persistent and it is measured with error.

the sample into groups sharing the same or similar values of certain observable characteristics. The randomization ensures that treatment and control groups will be similar *in expectation*. But stratification is used to ensure that along important observable dimensions this is also true *in practice* in the sample. For example, in the balsakhi program, described in section 3.3.3, researchers stratified according to class size, language of instruction, and school gender (boys, girls, or coed) as well as according to pre-test scores for schools in the Mumbai area. A block is constituted of all the schools that share the same language of instruction, the same school gender, and fall in the same "bin" of pre-test scores. By doing so, they ensured that the treatment and comparison groups would be balanced by gender and language of instruction and that pre-test score would be similar. An extreme version of blocked design is the pairwise matched design where pairs of units are constituted, and in each pair, one unit is randomly assigned to the treatment and one unit is randomly assigned to the control.

When the same proportion of observations are assigned to the treatment and the comparison groups in each block, the average treatment effect is equal to the difference between the outcomes of all treated and all untreated units or, equivalently, to the weighted average of the difference between treated and untreated units in each group (with the number of observations in each group as weight).

Very much like controlling for baseline variables ex post, blocking according to variables will improve precision to the extent the variables used for blocking explain the variation in the treatment of interest (Cox and Reid 2000). However, blocking is more efficient than controlling ex post for these variables, since it ensures an equal proportion of treated and untreated unit within each block and therefore minimizes variance. An easy way to see this is to observe that, at the extreme, a completely randomized design could lead to a situation where, in some blocks, there are only treatment or only control units. These blocks would not contribute anything to the analysis of difference between treatment and comparison groups when we control for the variables ex post, thus reducing the effective sample size and the precision of the treatment effect estimate.

More generally, Imbens, King, and Ridder (2006) show that when the proportion of treated and control units is the same in all strata (and equal to that in a completely randomized experiment) the variance of the treatment effect estimate is always weakly lower in the stratified design, with or without ex post controls. The same logic also implies that if several binary

variables are available for stratification, it is a good idea to use all of them, even if some of them may not end up having large explanatory power for the final outcome. The fact that the blocks within which the randomization will be performed may end up being very small is not a concern, since the estimate will be computed as an average over all these blocks.

When one or several of the possible stratification variables are continuous, so that one could form pairs on the basis of just one continuous variable, it will be necessary to make choices about which variables to use for stratification. For example, if one stratifies first according to gender and then by income, the treatment and comparison group's average incomes will be less similar than if one stratified only according to income. This choice is made taking into consideration the extent to which the candidate stratification variables are likely to explain the outcome variable and the treatment effect.

An estimate of the treatment effect and its variance that takes into account stratification can be obtained by estimating

$$Y_{ij} = \alpha + \beta T + M_{ij} + \tilde{\upsilon_j} + \tilde{\omega_{ij}}, \tag{15}$$

by OLS, where M is a set of dummy variables indicating the observation's block, by least squares using either the standard or robust estimates of least squares variance and taking into account the reduction in degrees of freedom due to stratification. Alternatively, one could ignore the stratification and estimate (), which is simply (15) without the block dummies.

Both methods are acceptable: With equal proportion of treatment and comparison units within each strata, ignoring stratification and estimating (15) without the block dummies leads to the exact same point estimates for β but a higher residual variance. The standard OLS variance based on that regression is a conservative estimator for the variance of $\hat{\beta}$. Although, *in expectation* the variance estimator based on (15) is less than or equal to that of the regression ignoring the block dummies , it is also noisier and, in a given sample, could be higher (Imbens, King, and Ridder 2006).

Apart from reducing variance, an important reason to adopt a stratified design is when the researchers are interested in the effect of the program on specific subgroups. If one is interested in the effect of the program on a sub-group, the experiment must have enough power for this subgroup (each sub-group constitutes in some sense a distinct experiment). Stratification

according to those subgroups then ensure that the ratio between treatment and control units is determined by the experimenter in each sub-group, and can therefore be chosen optimally. It is also an assurance for the reader that the sub-group analysis was planned in advance.

4.6 Power calculations in practice

This section reviewed the basic theoretical principles behind the calculation of power in an experiment. But how should power calculations be carried out by researchers in practice when planning an experiment, and for what purpose should they be used?

A first comment is that, despite all the precision of these formulas, power calculations involve substantial guess work in practice. To carry out power calculations, one must first have an idea of the mean and the variance of the outcome in the absence of the experiment, after controlling for possible covariates and/or stratification. For grouped designs, one must also have a sense of what the correlation in the outcomes of interest between different group members is likely to be. The best way to obtain guesses about these parameters is in general to use previously collected data, ideally from the same country or region. Sometimes such data are not available, and it is necessary to conduct a baseline survey to get a sense of the magnitude of these variables. But it may be difficult and time consuming, particularly when one of the variables one plans to control for is the baseline value of the outcome variable (this would require two surveys separated by some length of time). For clustered designs, finding reliable estimates for ρ can prove to be a challenge in practical applications. Table 1 displays a range of intra-class correlation for test scores (where the "class" corresponds to a grade level within a school. It shows that the intraclass correlation is high for test scores, ranging from 0.2 to 0.6. It is often smaller in other applications. It is worth performing power calculations with a variety of levels for ρ to get a range of required sample sizes.

One must then chose a level for the test. This is conventionally set at 5% or 10%, since this is the probability of a type-I error generally accepted as significant in published papers. Finally, one must specify the effect size that one wishes to be able to detect. As a rule of thumb for a policy intervention evaluation, this should be the smallest effect size that is large enough such that the intervention would be cost effective if it were to be scaled up. Cheap interventions should therefore be evaluated using larger samples. This, however, ignores the cost of the experiment itself. Moreover, for economists interested in getting insights about structural parameters, this

rule of thumb may not apply. It may be of intrinsic interest from an economics point of view to answer the question of whether a given intervention can have even a small effect on an outcome, irrespective of the immediate policy implications.

A shortcut when data on mean and standard deviation of the outcomes are not available is to directly specify the effect size one wishes to detect in multiples of the standard deviation of the outcome. Cohen (1988) proposes that an effect of 0.2 standard deviation is "small", 0.5 is "medium" and 0.8 is "large." Unfortunately, without a sense of what the standard deviation is, it is not clear that the distinction between large, medium and small has much practical meaning. But it can at least provide some idea to the researcher on whether the design will have power to perform a given design. This information can then be plugged into software which computes power under different scenarios.[18]

The final question is the level of power for which a researcher should aim and, more generally, how to use power calculations. A first view of power calculations is that they should be conducted ex-ante to determine the necessary sample to obtain a given power—many funding agency consider 80% to 90% an appropriate target. However, sample size is often determined in large part by budget or implementation constraints. In this case, a second view of power calculations is that they can help evaluate the power of a specific design, and thus help the researcher decide whether to embark on the project at all. Here, a question that naturally arises is therefore whether or not a researcher should accept to conduct a low-powered study. The answer is not obvious. Some argue that since the study has little chance to deliver conclusive results, it is not worthwhile to conduct it. From a social point of view (in particular if one adopt a Bayesian or decision theoretic point of view), however, this is forgetting that any particular study is only one of many that may be conducted on the topic. The greatest incremental precision on a given question comes from the first few observations. Moreover, results from several low power studies can be combined in a meta-analysis which will have more power. It remains that, from a purely private point of view, and taking into account the fact that each experiment involves fixed costs (in designing the experiments, the questionnaires, etc.), low-powered designs are probably best avoided in most cases by individual researchers.

A third use of power calculations is to help make decisions about how to design the experi-

[18] "Optimal Design" is a free tool that performs such power calculations (see Raudenbush, Spybrook, Liu, and Congdon (2005)).

ment to achieve the maximum power within a given budget. For example, is it worth conducting a baseline? In clustered designs, how many clusters should be sampled and how many units per cluster, given the fixed cost of surveying each cluster and the intra-class correlation within each cluster? How should many should individuals be allocated to different treatment groups? How many treatments can be reliably evaluated given the available sample size? Of course, these designs issues choices are not determined only by the need for precision. In the next section, we discuss them in more details.

5 Practical Design and Implementation Issues

This section discusses various design and implementation issues faced by those conducting randomized evaluations. We begin with the choice of randomization level. Should one randomize over individuals or some larger group? We then discuss cross-cutting designs that test multiple treatments simultaneously within the same sample. Finally, we address some data collection issues.

5.1 Level of Randomization

An important practical design choice is whether to randomize the intervention at the level of the individual, the family, the village, the district, etc. While early social experiments in the US were all randomized at the individual level, many evaluations in developing countries are randomized across groups.[19] For some interventions, such as those seeking to influence an entire community, the choice does not arise. For example, Chattopadhyay and Duflo (2004) study the reservation for women of leadership positions in village councils. The randomization necessarily takes place at the level of the *gram panchayat*, a local council encompassing several villages. All villages in a given *gram panchayat* are therefore either treatment or comparison; there is no room for randomization at the village level.

But for many interventions, it is possible to choose whether to randomize at the individual or the group level, and this choice is not always evident. For example, early randomization of deworming medicines were carried out at the individual level within schools (Dickson and Garner 2000), while Miguel and Kremer (2004) look at similar programs by randomly phasing-

[19] See Bloom (2005) for discussion of clustered randomized trials in the US context.

in the program at the school level. Interventions such as input provisions in classrooms could be carried out at the school level (for example, schools get selected to receive textbooks or flip charts) or at the level of individuals students (in the Tennessee Star experiment, students within schools were randomly assigned to either a large class, a small class, or a class with a teacher aid (Krueger and Whitmore 2002)

When there is flexibility in the level at which to randomize, several factors need to be taken into account. First, as discussed in section 4.2, the larger the groups that are randomized, the larger the total sample size needed to achieve a given power. The level of randomization thus has a potentially large effect on the budget and administrative burden of the evaluation, making individual-level randomization attractive when possible.

Second, however, spillovers from treatment to comparison groups can bias the estimation of treatment effects. In such case, the randomization should occur at a level that captures these effects. For example, Miguel and Kremer (2004) found much larger effects of deworming drugs than did earlier evaluations that randomized across individuals. They argue that because worm infections spread easily among children, the comparison group in individual-level randomizations also benefited from treatment, reducing the difference between the outcomes of treated and control children. While such spillovers are not necessarily absent when randomizing at larger levels (for example, Miguel and Kremer do show spillover across schools in their samples), they are typically much smaller. This can be an argument for randomizing at a level that captures any large spillover effects. Another form of externality that can occur is that individuals in the comparison group may change their behavior in anticipation of being treated in the future. It may be easier in a village-level randomization to leave the comparison group unaware of the existence of a treated group.

Third, randomization at the group level may some times be much easier from the implementation point of view, even if it requires larger sample sizes. There are various reasons for this. In interventions that have a strong fixed cost element in each location, it is cost-efficient to allow as many people as possible to take advantage of the interventions. For example, Banerjee, Duflo, and Glennerster are currently evaluating a community based iron fortification method. A local NGO trains the village miller to fortify flour with iron and supplies the fortification compound to him. Not allowing everyone in the community to take advantage of the trained miller would imply that the initial set up costs are not fully leveraged.

Another reason why group-level randomization may be preferred is that individual-level randomization of a program perceived as desirable in a village or a neighborhood may create resentment towards the implementation organization. Organizations may simply refuse to participate in such evaluations and, even if they agree, may be less likely to implement the experiment as designed. There may be slippage from the comparison to the treatment group (for example, some students assigned to large classes in the Tennessee Star Experiment found their way into small classes), either because the comparison individuals manage to get treated after a while, or because the field staff, intentionally or not, do not treat the right individuals based on the initial random assignment. It is much easier for a research team to ensure that villages are being treated according to an initial random assignment than to monitor individuals.

The choice of the level at which to randomize is therefore very context specific. It depends on the nature of the intervention as well as the nature of the interactions between the individuals to be treated.

5.2 Cross-Cutting Designs

One of the institutional innovations that led to a large increase in the number of randomized evaluations is the increased use of cross-cutting (or factorial) designs. In cross-cutting designs several different treatments are tested simultaneously with randomizations being conducted so that treatments are orthogonal to each other. Kremer (2003) describes many of those experiments conducted in education in Western Kenya.

There are two ways to think about cross-cutting designs. First, they can be used to test various interventions and combinations of interventions relative to a comparison groups and relative to each other. They can also establish whether treatments have important interaction effects. Policymakers are often interested in using a variety of strategies to change an outcome. For example, the PROGRESA program we discussed above is a combination of several programs: a cash transfer, a redistribution of resources towards women, and an incentive component. From a policy perspective, the evaluation of the "full PROGRESA" package may be sufficient for the Mexican government when deciding whether or not to continue with PROGRESA. But in order to learn about behavior and, for policy purposes, to understand which components of PROGRESA should be scaled up, one might want to know whether the incentive part of the scheme is necessary, whether distributing the money to women rather than to men matters, etc.

In principle, a cross cutting design could have been used in order to disentangle the various component of PROGRESA.

If a researcher is cross-cutting interventions A and B, each of which has a comparison group, she obtains four groups: no interventions (*pure control)*; A only; B only; and A and B together (*full intervention).* If a researcher wants to test whether B has a different effect when combined with A than alone, the sample sizes must be sufficient to allow her to statistically distinguish A versus A and B, as well as B versus A and B. As we discussed in section 4, one may consider making the *full intervention* and *pure control* groups larger than the *A only* and *B only* groups.

When such cross-cutting designs are too costly or require too large a sample size, a practical question that often arises is whether to evaluate a combined program (A and B) or to separately evaluate the two components. Policymakers may have a preference to evaluate the A and B combination as long it has the potential to be scaled up, since the A and B combination is more likely to have an effect than either A or B separately.

From an economist's perspective, the drawback of evaluating packages of interventions is that it makes it difficult to understand what drove the response and thus to extract lessons more general than just "this particular package worked." The advantage is that a more intensive intervention is more likely to have an impact and thus to show that outcomes can indeed be affected. If there is substantial uncertainty about the fact that either component may make a big difference to the outcomes of interest, it may make sense to first evaluate the combined package and then follow up with later studies designed to disentangle the various potential mechanisms at work. In the initial study, intermediate variables likely to be affected by one intervention but not the other can be used to shed light on which part of the intervention was effective. For example, in the deworming pilot mentioned above (Miguel and Kremer 2004), two programs were combined: deworming pills were distributed, and children were given advice about preventive behavior (wearing shoes, washing hands, etc.). Researchers collected variables on behavior which suggested that no behavior changed in the treatment schools. This strongly suggests that the component of the intervention that made the difference was the provision of the deworming pill.

Even when there is no interest in potential interactions between programs, cross-cutting designs can also be useful for testing multiple hypotheses rather than one, with little increase in cost, since the main cost of randomized evaluations typically consists of conducting the surveys to

establish baseline conditions and to measure outcome variables. In this case, the overall sample size need only be large enough to have sufficient power for the intervention that is expected to have the smaller effect. For example, Banerjee, Duflo, Cole, and Linden (2007) tested in the same sample (the municipal schools in Vadodara, India) the effect of remedial education and the effects of Computer Assisted Learning. As we saw above, half the schools received the remedial education program for grade 4. Half the schools received the computer assisted learning program, also for grade 4 students. The randomization for the computer assisted learning was stratified according to treatment status for the remedial education program. The same test scores were used to look at the effect of both programs.

In this case, the effect of remedial education we obtain is that of remedial education conditional on half the school getting computer assisted learning as well. This may have been problematic if computer assisted learning had little chance to be scaled up and if the effect of remedial education turned out to be very different in schools with and without computers. In this case, the two programs did not seem to interact with each other at all, so that the existence of two treatments did not diminish the external validity of the evaluation of each of them.

Because they significantly reduce costs, cross-cutting different treatments has proved very important in allowing for the recent wave of randomized evaluations in development economics.[20] They may also provide a window for graduate students or others who have limited access to resources to implement randomized research projects as additional treatments as part of larger projects. For example, using a cross-cutting design, Duflo, Dupas, Kremer, and Sinei (2006) evaluate the effect on risky sexual behavior of Kenya's teacher training for HIV/AIDS education and that of helping children stay in school longer by reducing the cost of education. As part of her dissertation, Dupas (2006) evaluated an additional intervention which she designed and implemented with the help of the NGO that was facilitating the initial project: the program was an information intervention where teenagers were informed of the relative prevalence of HIV/AIDS in different age groups. The intervention itself is very cheap and could be added to the program at minimal costs. Collecting the data would have been extremely expensive, but the necessary data was collected as part of the initial project. It turns out that this intervention proved to be much more effective in reducing pregnancy rates (the marker of risky sexual behavior) than the regular teacher training program. This suggests that adding this component to the regular

[20] Many of the experiments on education in Kenya described in Kremer (2003) were cross-cutting designs.

program has the potential to make it much more effective in reducing risky sexual behavior.

A final advantage of cross-cutting designs, evident in this example, is that, while a full welfare analysis is, as we discussed earlier, difficult, it is possible to compare the effectiveness of several different techniques for achieving a specific outcome. That is at least a second best.

5.3 Data Collection

We do not discuss specific survey design issues here as they are already covered by a substantial literature (see for example Deaton (1997)). Our main focus here is on the choice of what type of data to collect, the value of a baseline survey, and the use of administrative data.

5.3.1 Conducting Baseline Surveys

One of the first data collection questions that a researcher must address is whether or not to conduct a baseline survey. In principle, randomization renders baseline surveys unnecessary, since it ensures the treatment and comparison groups are similar in expectation. However, there are several reasons why researchers may want to conduct a baseline survey.

First, as we have discussed, a baseline survey generates control variables that will reduce the variability in final outcomes and therefore reduces sample size requirements. In terms of the cost of the evaluation, the trade-off between conducting a baseline survey and not conducting one boils down to comparing the cost of the intervention, the cost of data collection, and the impact that variables for which data can be collected in a baseline survey may have on the final outcome. When the intervention is expensive and data collection is relatively cheap, conducting a baseline will save money. When the intervention is cheap but data collection is expensive, it may be more cost effective to run a larger experiment without conducting a baseline.

Cost is not the only consideration, however. There are several other advantages of conducting baseline surveys. First, they make it possible to examine interactions between initial conditions and the impact of the program. In many cases this will be of considerable importance for assessing external validity. Second, a baseline survey provides an opportunity to check that the randomization was conducted appropriately. Third, collecting baseline data offers an opportunity to test and refine data collection procedures.

The alternative strategy of collecting "pre-intervention data" retrospectively in the post-survey will usually be unacceptable, because even if the program does not affect those variables

it may well affect recall of those variables. Sometimes sufficient administrative data is already available and can substitute for a baseline to gauge the validity of randomization and provide control variable for looking at the significance of interventions.

5.3.2 Using Administrative Data

Using administrative data (data collected by the implementing organization as part of their normal functioning) linked to information on treatment can greatly reduce the cost of data collection and reduce attrition. Use of administrative data is more common in developed countries, but even in developing countries researchers can have access to such data. For example, Angrist, Bettinger, and Kremer (2006) examine the medium-run impact of the Colombia voucher program by linking data on the voucher lottery with data on registration for Colombia's school completion/college entrance exam.

It is, however, important in such cases to ensure that data is comparable between treatment and comparison groups. For example, it may be that outcome variables of interest are collected as part of a program but only in program areas. It might be tempting to reduce data collection costs by only putting in place a new survey in comparison areas and relying on program data to get outcome variables in treatment areas. However, this could introduce biases as a difference in measured outcomes between treatment and comparison areas could reflect different data collection methodologies. For example, Duflo and Hanna (2006) study the impact of providing incentives based on teacher attendance in informal schools. In treatment schools, attendance is measured every day using date and time-stamped photographs. In comparison schools, attendance needs to be measured by through unannounced visits to the schools. In order to ensure uniformity of data collection, the program is evaluated by comparing random visits in both types of schools. And indeed, the average absence rate measured through the daily camera data is different than that measured through the random visit.

Another issue to be aware of is that a program may impact the *measurement* of an underlying variable of interest more than the variable itself. Consider an evaluation for which the outcome of interest is some underlying latent variable (such as learning) that is imperfectly measured by some proxy (such as test scores). In many cases the relationship between the latent variable and the proxy is plausibly unaffected by the program. However, if the program itself creates incentives which are tied to the proxy, then it will be desirable to measure the effect of the

intervention using another proxy variable which is also highly correlated with the latent variable but which is not linked to the incentives of the program. For example, in their evaluation of a teacher incentives program based on district test scores, Glewwe and Kremer (2003) collected data not only on the district test scores (on which the incentives was based) but also on a "low stakes" NGO-administered test, which provides an independent measure of learning.

6 Analysis with Departures from Perfect Randomization

This section discusses potential threats to the internal validity of randomized evaluation designs, and ways to either eliminate them ex-ante, or handle them in the analysis ex post. Specifically, we discuss how to analyze data when the probability of selection depends on the strata; analysis of randomized evaluations with imperfect compliance; externalities; and attrition.

6.1 The Probability of Selection Depends on the Strata

A first departure from perfect randomization is when randomization is conditional on observable variables, with different probability of being selected depending on the value of the observable variables. In section 4.5 we discussed designs where blocked designs (or stratification) were used to reduce the variance of the estimated treatment effect. The allocation of observations to treatment and comparison groups was the same in all blocks. It may also happen, however, that the probability of selection differs in different strata. Consider for example the Colombia voucher program already discussed. The randomization was done within each city, with a pre-fixed number of winners in each city. The ratio of lottery winners to total applicants was therefore different in each city. This implies that the lottery status is not random in the overall sample (for example, there may be more losers in Bogota than in Cali if Bogota had more applicants for a given number of places). However, it is still random within each city. In other words, the treatment status is random conditional of a set of observable variables (a stratum: in this case, a city).

Denote as T the treatment status, and X a set of dummy variables indicating the strata. Randomization conditional on observables implies that

$$E[Y_i^C|X,T] - E[Y_i^C|X,C] = 0,$$

so

$$E[Y_i|X,T] - E[Y_i|X,C] = E[Y_i^T|X,T] - E[Y_i^C|X,T].$$

Therefore

$$E_X\{E[Y_i^T|X,T] - E[Y_i^C|X,C]\} = E[Y_i^T - Y_i^C|T],$$

our parameter of interest. Finally,

$$E_X\{E[Y_i^T|X,T] - E[Y_i^C|X,T]\} = \int \{E[Y_i^T|x,T] - E[Y_i^C|x,C]\}P(X=x|T)dx.$$

This means that, if X takes discrete values, we can compare treatment and comparison observations in each strata and then take a weighted average over these strata, using as weights the proportion of treated units in the cells (this is the sample analog of the above expression). This gives the average effect of the treatment on the treated. The cells where everybody is treated or nobody is are dropped. This method can be applied whenever the randomization is conditional on a set of pre-determined characteristics. An alternative is simply to control for X in an OLS regression of the outcome Y on T. One must, however, be sure to include all the relevant dummies in the regression. Suppose for example that the probability to receive a program depends both on city and income (with two income categories: rich and poor). Then X must include dummy variables for each city, for each income categories, and all their interactions.

6.2 Partial Compliance

In some cases, an evaluation is designed to reach all individuals assigned to the treatment group and great care is taken to ensure that compliance is near perfect. This was the case in the Indonesian iron supplementation experiment discussed in the introduction of section 3, where compliance rates exceeded 92% (Thomas, Frankenberg, Friedman, Habicht, and Al 2003). There are many other cases, however, where compliance is not expected to be perfect. Sometimes only a fraction of the individuals who are offered the treatment take it up. Conversely, some members of the comparison group may receive the treatment. This is referred to as "partial (or imperfect) compliance."

A common reason for partial compliance is that researchers rarely have perfect control over what the comparison group chooses to do. To go back to the example of the iron supplementation study, some individuals in both experimental groups may have already been taking iron supplements, and may have continued to do so even after the evaluation started, since they knew they had one chance in two to be part of the placebo group. Even though nearly all individuals in the treatment groups were treated, since some individuals in the comparison group may have been treated as well, the difference in treatment probability between the treatment and comparison groups was not one. In some instances, members of the comparison group are treated directly by the program. For example, when randomization is at the school level for example, some students may decide to transfer from the comparison group to treatment group in order to benefit from the program offered to the treatment group. In the well-known Tennessee STAR class size evaluation, some children initially assigned to a large class also moved to a small class (Krueger and Whitmore 2002).

There are also cases where it is not possible to enforce compliance in the treatment group. For example, in the deworming program, only children present on the day of the deworming received the deworming pills. Tracking the children at home to administer the pills would have been prohibitively expensive. Thus not every child in the treated schools was treated.

In many cases, experiments do not intend to treat everybody in the treatment group. This is always the case in encouragement designs. For example, in the evaluation of the effect of information sessions on tax deferred account take up discussed previously (Duflo and Saez 2003), treatment individuals were offered financial incentives to attend an information session. Individuals in the treatment and the comparison groups were, however, both free to attend. The probability of attending the session among those who received the letter was 19% and only 6% among those who did not. While the difference in the probability to attend was fairly large (13 percentage points) and very statistically significant, it was far from being one.

In this case, the manipulation that the experimenters performed was to send a letter informing the employee of the benefit. However, the benefits office is more concerned about the impact of the information session itself. More generally, we are often interested in the effect of a given treatment, but the randomization only affects the *probability* that the individual is exposed to the treatment, rather than the treatment itself.

To be valid and to prevent the reintroduction of selection bias, an analysis needs to focus

on groups created by the initial randomization. One must compare *all* those initially allocated to the treatment group to *all* those initially randomized to the comparison group, whatever their actual behavior and their actual treatment status. The analysis cannot exclude subjects or cut the sample according to behavior that may have been affected by the random assignment. Doing so can lead to erroneous results. This was the case in several early studies examining a program in Cambodia that contracted out government health services to NGOs (Keller and Schwartz 2001, Bhushan, Keller, and Schwartz 2002, Schwartz and Bhushan 2004). Using a 1997 baseline survey and 2001 midterm survey, these studies found that outcomes improved more in the districts with contracting than in comparison districts; however, the 2001 midterm survey did not collect data in three of the eight districts initially assigned to treatment, but where acceptable bids were not received. Thus any estimates would be biased if districts that received acceptable bids differed from those that did not in unobserved variables that influence outcomes. For example, if potential contractors were more likely to bid on districts in which it appeared easiest to reach the contract targets, program effects could be overestimated. Bloom, Bhushan, Clingingsmith, Hung, King, Kremer, Loevinsohn, and Schwartz (2006) corrected the problem by collecting data on all districts that were randomly assigned to either the treatment or comparison groups and by comparing all districts initially assigned to the treatment group with all those assigned to the comparison group, regardless of their final assignment.

In cases where the actual treatment is distinct from the variable that is randomly manipulated, call Z the variable that is randomly assigned (for example, the letter inviting the university employees to the fair and offering them \$20 to attend), while T remains the treatment of interest (for example, attending the fair). Denote $Y_i(0)$ the potential outcome for an individual if $Z = 0$, and $Y_i(1)$ the potential outcome for an individual if $Z = 1$.

Because of random assignment, we know that $E[Y_i(0)|Z = 1] - E[Y_i(0)|Z = 0]$ is equal to zero, and that the difference $E[Y_i|Z = 1] - E[Y_i|Z = 0]$ is equal to the causal effect of Z. However, this is not equal to the effect of the treatment, T, since Z is not equal to T. Because Z has been chosen to at least influence the treatment, this difference is called the *Intention to Treat estimate* (ITT).

In many contexts, the intention-to-treat estimate is actually a parameter of interest. For example, in the case of the deworming program, if policymakers are interested in the cost effectiveness of a universal school based deworming treatment, and tracking children at home

is not practical, any estimate of the effectiveness of the program needs to take into account the fact that not all children will be present at school on the day of the treatment. In this case, the parameter of interest is the intention-to-treat.

There are, however, many circumstances where researchers are interested in the effect of the intervention (T) itself, rather than that of the instrument. This is particularly true when the evaluation is not designed to be scaled up as a policy but rather to understand the impact of a treatment that could potentially be delivered in many other ways. This was the case in the iron supplementation experiment. A policy that delivers iron pills and carefully monitors compliance is not a practical policy option. There are much cheaper ways to deliver iron, for example by investing in supplementation of food. Policymakers and researchers are therefore interested in the impact of a diet rich in iron, which only individuals who complied with the treatment are getting.

We now investigate what can be learned about the causal effect of the treatment T when compliance is imperfect, so that the randomization generates an instrument Z for the treatment of interest T. This is discussed in Angrist and Imbens (1994, 1995) and related work, and the analysis here follows their treatment.

6.2.1 From Intention To Treat to Average Treatment Effects

Consider the Wald estimate, which is the ratio of the intention-to-treat estimate and the fraction of individuals who were treated in the treatment and the comparison group.

$$\beta_W = \frac{E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0]}{E[T_i|Z_i = 1] - E[T_i|Z_i = 0]}. \tag{16}$$

Note that the Wald estimate is the IV estimate of β in equation 2, using the dummy Z as an instrument. Imbens and Angrist show that, under two assumptions below, this ratio can be interpreted as the average treatment effect for a well-defined group of individuals, namely those who are induced by the instrument Z to take advantage of the treatment.

The two identification assumptions are the following:

1. Independence: $(Y_i^C, Y_i^T, T_i(1), T_i(0))$ is independent of Z; and

2. Monotonicity: Either $T_i(1) \geq T_i(0)$ for all i or $T_i(1) \leq T_i(0)$ for all i.

The independence assumption subsumes two requirements. First, the fact that the comparison between outcomes for individuals exposed to different values of the instrument identify the causal impact of the instrument. This will be true by construction in the case of randomized evaluation, since the instrument is randomly assigned. Second, that potential outcomes are not directly affected by the instrument. This assumption does not necessarily hold in randomized evaluations and will need to be examined carefully.

The monotonicity assumption requires that the instrument makes *every person* either weakly more or less likely to actually participate in the treatment. For example, every person in the treatment group for the iron study is no less likely to get iron than had they been in the comparison group. This assumption needs to be examined on a case by case basis, but in most cases it will be reasonable.

We can manipulate the numerator of expression 16.

$$\begin{aligned} E[Y_i|Z_i=1]-E[Y_i|Z_i=0] &= E[T_i(1)Y_i^T+(1-T_i(1))Y_i^C|Z_i=1]-E[T_i(0)Y_i^T+(1-T_i(0))Y_i^C|Z_i=0] \\ &= E[(T_i(1)-T_i(0))(Y_i^T-Y_i^C)+E[Y_i^C|Z_i=1]-E[Y_i^C|Z_i=0], \end{aligned}$$

which, by the independence assumption, is equal to $E[(T_i(1)-T_i(0))(Y_i^T-Y_i^C)]$. This can be expanded to

$$\begin{aligned} &E[-(Y_i^T-Y_i^C)|T_i(1)-T_i(0)=-1]P[T_i(1)-T_i(0)=-1]+ \\ &E[(Y_i^T-Y_i^C)*0|T_i(1)-T_i(0)=0]P[T_i(1)-T_i(0)=0]+E[Y_i^T-Y_i^C|T_i(1)-T_i(0)=1]P[T_i(1)-T_i(0)=1]. \end{aligned}$$

The first term cancels out due to the monotonicity assumption. The second term cancels out since the difference is pre-multiplied by zero. This expressions therefore simplifies to

$$E[Y_i^T-Y_i^C|(T_i(1)-T_i(0)=1]P[T_i(1)-T_i(0)=1].$$

Meanwhile,

$$\begin{aligned} P[T_i(1)-T_i(0)=1] &= P[T_i(1)=1,T_i(0)=0]\mathbf{1}[T_i(1)=1,T_i(0)=0] \\ &= \mathbf{1}[T_i(1)=1]-\mathbf{1}[T_i(0)=1]. \end{aligned}$$

Taking expectations,

$$\begin{aligned} P[T_i(1) - T_i(0) = 1] &= E[T_i(1)] - E[T_i(0)] \\ &= E[T_i|Z = 1] - E[T_i|Z = 0]. \end{aligned}$$

Hence

$$\begin{aligned} \hat{\beta}_W &= \frac{E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0]}{E[T_i|Z_i = 1] - E[T_i|Z_i = 0]} \\ &= E[Y_i^T - Y_i^C|(T_i(1) - T_i(0) = 1]. \end{aligned}$$

Under the monotonicity and the independence assumptions, the Wald estimator gives us the effect of the treatment *on those whose treatment status was affected by the instrument*, which is known as the local average treatment effect (LATE) (Angrist and Imbens 1994). These are those who, in the absence of the randomly assigned instrument, would not have been treated but are induced to receive treatment by the assigned instrument. They are often referred to as the *compliers*.

A special case is when nobody in the comparison group is treated, $T_i(0) = 0$. In this case, the Wald estimate is the *effect of the treatment on the treated*. For example, in the second year of the balsakhi study, some schools that had been assigned a balsakhi did not get one. The difference between the average test score of all children in the initial, randomly-assigned treatment group and all children in the initial, randomly-assigned comparison group, divided by the probability that a school receives a balsakhi conditional on having been assigned to the treatment group is an estimate of the average effect of the balsakhi program on children in schools that were actually treated. To the extent that the treatment schools that did not get a balsakhi are different from those that did, the estimated effect may not be representative of the programs impact on the average school.

Another special case is when everybody in the treatment group is treated, $T_i(1) = 1$. This will often be true in "oversubscription" designs, where successful applicants are selected from a list, and offered the treatment. In this case (subject to the caveats we discuss below), the Wald estimate identifies the effect of the treatment on those who would not be treated without inducement.

When the randomization only induces imperfect assignment to the treatment and the comparison groups, it is therefore still possible to make meaningful causal statements. However, the average causal effect that is estimated is not necessarily representative of the average causal effect for the entire population. Depending on circumstances, it may or may not be representative of a sub-population of interest. Those who are induced by the evaluation to take up a particular treatment may be different than those who were already taking it up or who would be likely to be induced to take it up by another policy intervention. In other words, another selection effect appears in this case, although it does not bias the estimation of the causal effect for compliers.[21]

In some cases, the group of compliers is exactly the group of interest, precisely because they are the ones that are likely to be affected by a policy. In some cases, policymakers are really interested in the impact of average impact of the policy in as representative a group as possible.

In this case, there may be a tradeoff between the benefits of a less tightly controlled evaluation, where the initial random assignment is used as an instrument for the policy and that of an evaluation where the initial randomization is very closely respected. The first evaluation may be much easier to implement, and can be conducted in a larger area and on a larger sample, but the group of compliers may be small and the treatment effects may not be representative of what would be obtained in the population at large. The second evaluation is more difficult to carry out, and requires much larger budgets or a smaller sample size.

This analysis also helps us compare prospective randomized evaluations and natural experiments. Some non-experimental studies take advantage of "natural experiments," where assignment to a treatment or a policy is in part due to a randomly assigned variable that can be used as instrument for the treatment. For example, Angrist (1990) studies the impact veteran status on civilian earnings by taking advantage of the Vietnam draft lottery, where draft assignment to service was in part based on a random number. Because in many cases the random factor only explains part of the variation in the actual treatment, such studies often have a first-stage that is small in magnitude (even if it is very statistically significant), which implies that compliers represent only small share of the population. Even when a natural experiment utilizes very large samples that may be representative of a country's entire population, the resulting estimate may not have more external validity (that is, be less applicable in other contexts) than those of a

[21] See Heckman and Vytlacil (2005) for a more extensive discussion of marginal treatment effects.

randomized evaluation conducted in a smaller sample, but carefully controlled such that the first stage is much larger, because identification comes from a very narrow and non-random group of the population.[22]

6.2.2 When is IV Not Appropriate

In order for the IV estimate to be interpreted as the causal effect of a treatment on the compliers, both the monotonicity and the independence assumptions must hold. Randomized evaluations make it likely that both assumptions are satisfied, but they do not necessarily ensure it. Recall that the independence assumption discussed above requires that potential outcomes of any treatment state, (Y_i^T, Y_i^C), are independent of Z, the instrument. In some case, this assumption fails to hold.

First, the instrument may affect non-compliers in the treatment group. Return to the balsakhi example, and consider only the first year of the evaluation where compliance was perfect at the school level (Banerjee, Duflo, Cole, and Linden 2007). Because schools were randomly assigned to the treatment and comparison groups, the comparison between the test scores of *all* children in the treatment groups and *all* children in the comparison groups provides an unbiased estimate of the average effect of the program on all children, or the intention-to-treat estimate. Noting that only 20% of the children in treatment schools were actually assigned to work with the balsakhi—recall that this is a remedial education program—it is tempting to divide the ITT estimate by the probability of being sent to the balsakhi in the treatment group in order to obtain the effect of working with the balsakhi on the children who actually received the remedial education. This, however, is inappropriate, because children in treatment schools who were not sent to the balsakhi may have benefited from the fact that their class is now smaller for most of the day, and their weakest peers have left the classrooms. At the extreme, it could have been the case that the entire effect on the average class was due to an improvement in the learning level of top scoring children who were not sent to the balsakhi and now enjoy better learning conditions. By using the fact that the school was assigned to the program as an instrument for the fact that the child actually received the remedial education, one would, possibly erroneously,

[22] As an example, Card (1999) discusses the interpretation of natural experiment estimates of returns to education and shows that many of them can be interpreted as the returns to education for individuals with high discount rates or credit constrained individuals. This may explain why they tend to be larger than OLS estimates.

force all the impact of the treatment to work through the remedial education classes. If these effects are positive (as we may expect in this case), the IV will be an overestimate of the impact of the treatment on compliers.

The same situation occurred in the deworming program (Miguel and Kremer 2004). Since not all children in treatment schools were treated, it may again be tempting to divide the intention-to-treat estimate by the fraction of children treated to estimate an average treatment effect. However, as Miguel and Kremer show (and we discuss in more detail below), non-treated children in treatment schools actually benefited from the treatment, since they were exposed to fewer worm-carrying children. Once again, the Wald estimator would overestimate the effect of treatment on the treated, while the intention-to-treat estimate is a valid estimate of the effect of the program on the entire school.

It is important to note that, even if these externalities are small, the bias will be magnified because the ITT estimate is divided by a number smaller than one. While this is less of a concern when the first stage is very powerful, when it is weak the bias can become extremely large.

6.3 Externalities

Experimental interventions can create spillover effects such that untreated individuals are affected by the treatment. Spillovers may be physical—substantial disease reduction externalities were found in the evaluation of a Kenyan primary school deworming program for example (Miguel and Kremer 2004). They may also result from price changes—Vermeersch and Kremer (2004) found that the provision of school meals to preschoolers at some schools in Kenya led nearby schools to reduce school fees. Spillovers can also occur in the form of learning and imitation effects (see Duflo and Saez (2003), Miguel and Kremer (2004)).

To see how spillovers can lead to biased estimates of treatment effects, consider the simple situation in which a treatment is randomly allocated across a population of individuals and compliance is perfect. Using the potential outcome framework, the intention-to-treat estimate is $ITT = E[Y_i^T|T = 1] - E[Y_i^C|T = 0]$. In order to interpret this difference as the effect of the treatment, the standard unit treatment value assumption (SUTVA) must hold. It says that the potential outcomes for each individual are independent of his treatment status, as well as the treatment group status of any other individual (Angrist, Imbens, and Rubin 1996). If this

is violated, $\hat{E}[Y_i^C|T=0]$ in the sample is not equal to $E[Y_i^C|T=0]$ in the population, since the sample contains both treated and untreated individuals. The potential outcome for each individual (and therefore the ITT) now depends on the entire vector of allocations to treatment and comparison groups. If the spillover effects on untreated individuals are generally positive, then the intention-to-treat estimate ITT will generally be smaller than it would have been without spillovers.

It is easy to see that this assumption will be violated when externalities are present. Consider once again the deworming program in Kenya. Those children who received the treatment were directly protected against worms; however, the untreated children with whom they have contact and against whom they are compared will experience fewer infections as well since treated individuals no longer transmit worms. Miguel and Kremer (2004) note that previous work on deworming programs may have underestimated treatment effects because it randomized treatment among individuals in the presence of positive spillovers. If spillovers are negative, the estimate would be upwards biased.

If spillovers are global (e.g., changes in world prices), identification of program effects will be problematic with any methodology. If they are local, randomization at the group level can allow for estimation of the total program effect on the group. If externalities do not operate across groups, group-level randomization is sufficient to identify overall treatment effects. It cannot, however, decompose the direct and spillover effects.

Where spillovers are likely to be important, experiments can be specifically designed to estimate their extent and magnitude. A first technique is to purposefully vary the level of exposure to a treatment within a group. For example, in their study of information and 401(k) participation, Duflo and Saez (2003) randomized the offer of getting an incentive to attend information session at two levels. First a set of university departments were randomly chosen for treatment, and then a random set of individuals within treatment departments were offered the prize. This allowed the authors to explore both the direct effect on attendance and plan enrollment of being offered an incentive and the spillover effect of being in a department in which others had been offered incentives.

A second technique is to exploit the variation in exposure across groups that naturally arises from randomization. For example, Duflo, Kremer, and Robinson (2006) performed on-site agricultural trials in a randomly selected sample of farmers. They then asked all farmers

the names of the three farmers they discuss agriculture with the most often (referred to below as friends). They then compare adoption of fertilizer among the "friends" of the treatment farmers to that of the "friends" of the comparison farmers. This allows them to estimate the extent of information externalities. Likewise, Miguel and Kremer (2004) compare adoption of deworming pills among the friends of children who were in early treatment schools to that of the friends of children in the late treatment schools. Miguel and Kremer (2004) estimate cross-group externalities by exploiting the fact that randomization created local variation in the density of treatment schools and hence random variation in the likelihood that a student in a non-treated school would be exposed to spillovers. Specifically, they estimate regressions of the form $y_{ij} = \beta_0 + \beta_1 T_j + \sum_d \gamma_d N_{dj}^T + \sum_d \phi_d N_{dj} + \varepsilon_{ij}$, where y_{ij} is the outcome of interest for individual i in school j, T_j indicates whether school j is a treatment school, and N_{dj}^T and N_{dj} measure the number of pupils within distance d of school j in treatment schools and all schools, respectively. The independent effect of school density on the outcome is captured by ϕ_d. The average effect of treatment on the outcome in treatment schools is given by $\beta_1 + \sum_d \gamma_d \bar{N}_d^T$, where $\bar{N}_d^T$ is the average number of treatment pupils a student in a treatment school is exposed to within distance d. The first term represents the direct effect (including within-school externalities on any non-treated pupils) and the second term represents the cross-school externality.

A third technique to estimate spillover effects is to randomly assign individuals to different peer groups. For example, the *Moving to Opportunity* experiment in the US (Liebman, Katz, and Kling 2004) offered randomly selected individuals vouchers to move to lower poverty neighborhoods. The comparison between those who received vouchers and those who didn't provides an estimate of the importance of neighborhood effects.

6.4 Attrition

Attrition refers to the failure to collect outcome data from some individuals who were part of the original sample. Random attrition will only reduce a study's statistical power; however, attrition that is correlated with the treatment being evaluated may bias estimates. For example, if those who are benefiting least from a program tend to drop out of the sample, ignoring this fact will lead us to overestimate a program's effect. While randomization ensures independence of potential outcomes in the initial treatment and comparison groups, it does not hold after non-random attrition. This problem occurred in the first large-scale randomized evaluation in

the US, the Negative Income Tax experiment, and produced a rich econometric literature of ways to address the issue (Hausman and Wise 1979, Heckman 1979).

Even if attrition rates are similar in treatment and comparison groups, it remains possible that the attritors were selected differently in the treatment and comparison groups. For example, in the evaluation of a medication, attrition due to death may be reduced in the treatment group, but attrition due to the fact that the subject feel healthier and stop complying with the experimental protocol may be increased in the treatment group.

This makes attrition a very difficult problem to solve ex post and that implies managing attrition during the data collection process is essential. Attrition can be limited by implementing systems to carefully track participants even after they leave the program. For example, in the balsakhi program (Banerjee, Duflo, Cole, and Linden 2007) children were tested at home if they were not found in school after a number of visits, which resulted in low attrition rates. This requires good information on where to find participants even if they drop out of the program. If the goal is to follow participants for a long time after the end of the program, it is important to collect good information in the baseline on how to find them later on (for example the names of neighbors and relatives that can be interviewed if the respondent cannot be found). When following up with *all* attritors is too expensive, a random sample of the attritors can be selected for intensive follow-up. In the analysis, these observations need to be given a higher weight, reflecting their sampling probability.

A first step in the analysis of an evaluation must always be to report attrition levels in the treatment and comparison groups and to compare attritors with non-attritors using baseline data (when available) to see if they differ systematically, at least along observable dimensions. If attrition remains a problem, statistical techniques are available to identify and adjust for the bias. These techniques can be parametric (see Hausman and Wise (1979), Wooldridge (2002) or Grasdal (2001)) or nonparametric. We will focus on non-parametric techniques here because parametric methods are more well known. Moreover, non-parametric sample correction methods are interesting for randomized evaluation, because they do not require the functional form and distribution assumptions characteristic of parametric approaches. Important studies discussing non-parametric bounds include Manski (1989) and Lee (2002)

The idea of non-parametric Manski-Lee bounds is to use plausible assumptions about the monotonicity of potential outcomes and attrition along with relative rank restrictions on the

distribution of potential outcomes to derive a bound on the treatment effect that can be estimated from available data. Ordinary treatment effect estimates will provide either upper or lower bounds on the true effect depending on the direction of attrition bias. When attrition bias is negative and the treatment effect is positive, the ordinary estimates provide a lower bound for the true effect, and the upper bound is estimated using the Manski-Lee approach.

Below we summarize the approach to attrition taken by Angrist, Bettinger, and Kremer (2006) to evaluate the long term impact of a Colombian voucher program on latent learning. As we discussed above, secondary school vouchers were allocated by lottery among a set of applicants. The authors matched lottery winners and losers to records from Colombia's high school graduation/college entrance exam, finding that winners were more likely to take the exam. The differential high school completion rates, while interesting in their own right, make estimating the impact of the program on learning tricky. Let y_{1i} be the outcome for individual i if offered treatment and y_{0i} the outcome they would otherwise obtain. D_i is an indicator variable for random assignment to treatment. Let T_{1i} be an indicator variable for whether the individual would remain in the sample conditional being assigned to the treatment group, and T_{0i} similarly indicate whether they would remain in the sample if assigned to the comparison group. Assume that $y_{1i} \geq y_{0i}$ and $T_{1i} \geq T_{0i}$ for all i. These assumptions mean that treatment is never harmful and that those offered treatment at least as likely to remain in the sample as those who are not. Now define an outcome variable that is zero for attritors: $Y_{Xi} = T_{Xi} y_{Xi}$ for $X = \{0, 1\}$. Then we can write the following equation linking the actually observed outcome Y_i to potential outcomes, attrition status, and treatment group:

$$Y_i = Y_{0i} + (Y_{1i} - Y_{0i}) D_i = T_{0i} y_{0i} + (T_{1i} y_{1i} - T_{0i} y_{0i}) D_i.$$

Let $q_{i0}(\theta)$ be the θ-quintile of the distribution of Y_{i0} and let $q_{i1}(\theta)$ be the θ-quintile of the distribution of Y_{i1}. Now define a rank-preservation restriction: Y_{1i} is said to be a θ-quintile preserving transformation of the random variable Y_{0i} if $P(Y_{i1} \geq q_{i1}(\theta) | Y_{i0} \geq q_{i0}(\theta)) = 1$. In other words, the rank preservation restriction says that when the potential outcome in the comparison state is above a certain quantile in its own distribution, then the potential outcome in the treatment state is also above that quantile in its own distribution. Given the assumptions already outlined and a choice of θ such that $\theta \geq \theta_0$ where $q_{0i}(\theta_0) = 0$, Angrist, Bettinger, and Kremer (2006) prove that

$$
\begin{aligned}
E\left[Y_i|D_i=1, Y_i > q_{i1}\left(\theta\right)\right] - E\left[Y_i|D_i=0, Y_i > q_{i0}\left(\theta\right)\right] &\geq E\left[y_{i1} - y_{i0}|y_{i0} > q_{i0}\left(\theta\right), T_{i0}=1\right] \\
&\geq E\left[Y_i|D_i=1, Y_i > q_{i0}\left(\theta\right)\right] - E\left[Y_i|D_i=0, Y_i > q_{i0}\left(\theta\right)\right].
\end{aligned}
$$

One can then choose a quantile, θ_0, such that $q_{0i}\left(\theta_0\right)=0$, and then drop the lower θ_0 percent of the Y_{i1} distribution to obtain an upper bound on $E\left[y_{i1}-y_{i0}|T_{i0}=1\right]$ while the unadjusted treatment effect provides a lower bound. Note that the bound will be the tighter, the lower is the attrition. This underscores the need for limiting attrition bias as much as possible.

7 Inference Issues

This section discusses a number of the key issues related to conducting valid inference from randomized evaluations. We begin by returning to the issue of group data addressing how to compute standard errors that account for the grouped structure. We then consider the situation when researchers are interested in assessing a program's impact on several (possibly related) outcome variables. We next turn to evaluating heterogeneous treatment effect across population subgroups, and finally discuss controlling for covariates in estimation.

7.1 Grouped Data

As was introduced in section 4.2, when the randomization takes place at the group level, standard errors need to take into account possible correlation in the outcome variables between members of the same group. Equation (11) gives us the inflation factor for the standard errors under the assumption of no heteroskedasticity and a common covariance structure across group (Moulton 1990). In this case, equation (2) can also be estimated more efficiently by Generalized Least Squares, assuming a group random effect.

If one wants to avoid the assumption of a common covariance structure, one approach to computing standard errors with grouped data is to use the cluster-correlated Huber-White covariance matrix estimator. This approach is recommended when the number of groups randomized is large enough. However, Donald and Lang (2001) and Wooldridge (2004) have pointed out that asymptotic justification of this estimator assumes a large number of aggregate units. Simulations in Duflo, Mullainathan, and Bertrand (2004) show the cluster-correlated Huber-White estimator

performs poorly when the number of clusters is small (less than 50), leading to over-rejection of the null hypothesis of no effect.

When the number of clusters is small, hypothesis tests can also be generated using randomization inference (Rosenbaum 2002). This approach involves generating placebo random assignments P_j and associated regression coefficients, denoted B_p. Denote the set of all possible assignments from the randomization process $\{P_j\}$. Now consider β_p in the regression equation

$$Y_{ij} = \delta + \beta_p P_j + \upsilon_{ij}.$$

Since P_j is a randomly generated placebo, $E(\beta_p) = 0$. Let $\hat{F}\left(\hat{\beta}_p\right)$ be the empirical c.d.f. of $\hat{\beta}_p$ for all elements of $\{P_j\}$ .We can now perform a hypothesis test by checking if our measured treatment effect is in the tails of the distribution of placebo treatments. We can reject H_0: $\hat{\beta}_T = 0$ with a confidence level of $1 - \alpha$ if $\hat{\beta}_T \leq \hat{F}^{-1}\left(1 - \frac{\alpha}{2}\right)$ or $\hat{\beta}_T \geq \hat{F}^{-1}\left(1 - \frac{\alpha}{2}\right)$. Since the placebo assignments, P_j, only vary across clusters, this method takes intracluster correlations into account.

The advantage of randomization inference is that it is valid for any sample size, and can thus be used even when the number of sample is very small. Bloom, Bhushan, Clingingsmith, Hung, King, Kremer, Loevinsohn, and Schwartz (2006) use this method to compute standard errors in their study of the impact of subcontracting the management of public health care center in Cambodia. Randomization was carried out at the district level, and only 12 districts participated in the study. Clustered standard errors could therefore be affected by fairly strong bias. Note, however, that while unbiased, randomization inference has low power relative to more parametric approaches when the true effect is large because it does not put even minimal structure on the error term (see the discussion in Bloom, Bhushan, Clingingsmith, Hung, King, Kremer, Loevinsohn, and Schwartz (2006)).

7.2 Multiple Outcomes

Evaluations often affect many different outcomes that the experimenter subsequently measures. Testing hypotheses regarding multiple outcomes calls for special techniques. Standard hypothesis testing supposes that the experimenter is interested in each outcome separately. But when testing multiple outcomes, the probability of rejecting a true null hypothesis for at least one outcome is greater than the significance level used for each test (Kling and Liebman 2004); a

researcher testing ten independent hypotheses at the 5% level will reject at least one of them with a probability of approximately 40%.

Suppose that in our example, individual hypotheses test showed the estimated effect of the intervention was statistically significant for scores in math but for no other subjects. While a policymaker concerned only with math scores might focus on the point estimate of the effect on math, an experimenter reporting the results of the program would be wrong to draw the inference that the program worked in math but not in other subjects. In order to make a correct inference, the standard errors must be adjusted to account for the fact that the outcome is a member of a family of hypotheses. This is often referred to as the "familywise" error.

Adjusted p-values for each outcome are to be constructed such that the probability is less than .05 that at least one of the tests in a family would exceed the critical value under the joint null hypothesis of no effects. The simplest and most conservative approach is Bonferroni adjustment, in which each p-value is multiplied by the number of tests in the family (Savin 1984). This approach is too conservative, however, since it it treats all the hypotheses as independent.[23]

An alternative approach with multiple related hypothesis (a family of hypotheses) is thus to test whether the *overall* effect of treatment on a family of outcomes is significantly different from zero. Following our example, a policy maker may be interested in the effect of their intervention on test scores in general, rather than on each subject separately. Measurement of such overall effects has its roots in the literature on clinical trials and on meta-analysis (See O'Brien (1984); Logan and Tamhane (2003); and Hedges and Olkin (1985)). A summary measure that captures this idea is the mean standardized treatment effect. Suppose that there are K different outcomes in the family of interest. Let the point estimate and standard error for each effect be given by and $\hat{\pi}_k$ and $\hat{\sigma}_k$. Following O'Brien (1984), Logan and Tamhane (2003), and Kling, Liebman, Katz, and Sanbonmatsu (2004), the mean standardized treatment effect is given by $\tilde{\pi} = \frac{1}{K}\sum_{k=1}^{K}\frac{\hat{\pi}_k}{\hat{\sigma}_k}$. The standard error of this mean effect needs to take into account the fact that all the outcomes are correlated: this can be done by running seemingly unrelated regressions (SUR) for all the standardized outcomes falling into a family. The mean standardized treatment effect is preferable to alternatives, such as a joint F-test across all the outcomes, since it is unidirectional and thus has more power to detect whether all effects go in the same direction.

[23] Alternative approaches are the Bonferroni-Holm step-down adjustment (Holm 1979) and the bootstrap method of Westfall and Young (1993) (See Kling and Liebman (2004) for an example).

These corrections help avoid the publication bias problems discussed above. Unlike a retrospective analysis which may draw data from a large data set with many potentially unrelated variables (like a census), all the variables collected in a prospective evaluation were, presumably, collected because they were considered potential outcome variables. All should therefore be reported whether they are significant or not. Where there are a large number of variables, it can also be useful for the researcher to set out ahead of time which variables fall into which families for family testing.

7.3 Subgroups

Interventions often have heterogeneous effects on the population they affect. For example, we might expect a remedial education program to have a greater effect on students who have low test scores than on those who have high test scores. If we randomly allocate an intervention across groups (such as classrooms) containing students of both types, then the effect of treatment will be an average of the effect on low and high scoring children. Researchers and policymakers may, however, be interested in testing the effect separately for low- and high-scoring children.

Ideally, researchers should know when designing the evaluation protocol, either through a priori reasoning or through knowledge gained from other studies, which of the possible subgroups should be investigated separately and make this decision clear ex ante. Theoretical restrictions may in fact give rise to additional testable hypotheses—for example, that the program has a significant impact on low-scoring children, but not on high-scoring children.

Note that in clustered designs evaluations may have almost as much power for such sub-group as it is for the entire sample. This is because the number of observations per cluster matter more for power than the number of clusters. Moreover, if one of the reason for the correlation within cluster was that the fraction of individuals belonging to each subgroups vary from a cluster to another, examining sub-groups separately may reduce the within-group correlation enough to compensate for the loss in sample size (Bloom 2005).

In some cases (though not when the randomization is carried out at the group level and each group contain members of different sub-groups of interest) it is possible to stratify our randomization of individuals into treatment and comparison groups by the subgroups, and to estimate treatment effects in each subgroup. This makes the intention to looking at the effects for different subgroup explicit. Even when subgroups are determined in advance, standard errors

need to be adjusted for the fact that there are several subgroups. The possible adjustments are very much like the adjustments for multiple outcomes just reviewed.

What if a researcher discovers after the randomization has been conducted and the evaluation begun that a particular subgroup seems to have a very different treatment effects? For example, Glewwe, Kremer, and Moulin (2004) find no evidence that providing textbooks to rural Kenyan primary schools increased scores for the typical student. However, there is evidence that the program increased test scores for students with initially higher academic achievement. While this "cut" was not intended in advance, it can be easily be rationalized: since the textbooks were in English, they were unlikely to help the weaker student. Should such evidence be reported? Likewise, evaluation results are often separately reported for different sites.[24] In Kenya, Kremer, Miguel, and Thornton (2004) separately report the result of an incentives program for two districts, one where it "worked" and the other where it was ineffective. One could argue that dividing the sample according to a variable that was not part of the original randomized design may lead to data mining. Because the universe of possible "cuts" of the data is not known, computing Bonferroni bounds or similar adjustments is not possible, and standard errors cannot be properly calculated.

The most stringent standard, which is applied by the FDA in clinical trials of new drug therapies, is that unless the subgroups are specified ex ante, results for specific subgroups will not be considered sufficient evidence for drug licensure. The FDA requires a new trial designed ex ante to measure the impact on any subgroups. The reason the FDA takes this approach is that if a researcher (or pharmaceutical company) has free reign to examine an arbitrary number of subgroups, it will typically be possible to choose subgroups such that the intervention appears effective within that group.

From a Bayesian point of view, however, a policymaker who trusts that the researcher has not data mined among many potential subgroups, and who believes there is a good a priori case to believe that effects might differ by subgroups would seem to have reason to take into account the results of subgroup analysis, even if done ex post. In this context it is also important to remember that randomized trials constrain researchers' options much more than in standard

[24]This is the case for most social experiments in the US. For example, the results of the *Moving to Opportunity* experiment were first reported for the city of Boston (Liebman, Katz, and Kling 2004), where it seemed to have bigger effects than in some other sites, at least in early years of the program.

empirical economics, where, almost by construction, there are many more opportunities to engage in data mining, subgroup selection, and specification searching.

While every effort should be made to lay out the planned subgroup analysis in advance and typically to stratify the randomization by subgroup, we do believe that it may sometimes be worthwhile to report results from ex post grouping as well as those that were initially planned, since they can shed additional light on the first set of results. However, if results according to ex post grouping are reported, a researcher's report must make it very clear which subgroups were defined ex ante and which subgroup was defined ex post. From the point of view of a researcher, an evaluation that yields suggestive results *ex post* on groups that had not been designed *ex ante* should be thought of as the first step to a further evaluation focusing in particular on these effects.

7.4 Covariates

Another element of choice in the analysis of an experiment is the choice of what to control for. As we discussed above, controlling for variables not affected by the treatment may reduce the variance of the outcome, leading to more precise estimates (controlling for variables that are *affected* by the experiment would of course lead to biased estimates). But once again, those variables should be specified ex ante, to avoid specification searching. It is common practice to report both "raw" differences and as well as regression-adjusted results.

8 External Validity and Generalizing Randomized Evaluations

Up until now we have mainly focused on issues of internal validity, i.e., whether we can conclude that the measured impact is indeed caused by the intervention in the sample. In this section we discuss external validity—whether the impact we measure would carry over to other samples or populations. In other words, whether the results are generalizable and replicable. While internal validity is necessary for external validity, it is not sufficient. This question has received a lot of attention in the discussions surrounding the use of randomized evaluation. A very interesting overview of this debate is given by the papers of Bardhan, Basu, Mookherjee and Banerjee in a symposium on "new development economics" (Banerjee 2005, Basu 2005, Mookherjee 2005, Bardhan 2005, Banerjee, Bardhan, Basu, Kanbur, and Mookherjee 2005). In this section, we

discuss reasons why one may worry about the external validity of randomized evaluations, and what are the ways to ameliorate these concerns.

8.1 Partial and General Equilibrium Effects

Because randomized evaluations compare the difference between treatment and comparison populations in a given area, they are not able to pick up general equilibrium effects (Heckman, Lochner, and Taber 1998). Such effects may be particularly important for assessing the welfare implications of scaling up a program. For example, in the evaluation of the voucher program in Colombia, the researchers compared the outcomes for students who were given vouchers to attend private school with those of students who applied for but did not receive vouchers. This evaluation was able to identify the impact of winning a voucher given that there was a system of vouchers in place—in other words, it measured the partial or localized effect of the program on recipients. It was not, however, able to say what the overall impact of introducing a voucher system was on the education system in Colombia.

Both advocates and detractors of voucher systems point to potential general equilibrium effects of vouchers. Proponents suggest that the added competition introduced by vouchers increases the pressure on public schools and that these schools improve their performance as a result (Hoxby 2003). Opponents argue that vouchers are likely to increase sorting of students by ability, preferences, and race and may undermine the role of education in creating a cohesive society with common identity and shared values (Hsieh and Urquiola 2003). To the extent that vouchers pull the most motivated children and their parents out of public schools, they may reduce the pressure generated by articulate, motivated parents on public schools to perform well.

Neither of these effects can be measured in the study on Colombian vouchers. Increased competition could improve the quality of education in public schools (where more comparison children are educated) and so reduce the difference between treatment and comparison outcomes. This would reduce the measured effect of vouchers even though there was a positive impact of vouchers on the system. Increased sorting would impact both treatment and comparison schools, and so a comparison between the two would not be able to pick it up. A decline in performance in public schools due to the loss of the most committed students and parents would show up in a study of this kind as a larger gap between the performance of treatment and comparison

students. In other words, the bigger the magnitude of this negative effect of vouchers, the better it would make the outcomes of vouchers appear.

General equilibrium effects of this kind can be thought of as another variety of externality. As with externalities, it would be possible to pick up some of these general equilibrium effects if the unit of observation were large enough—although this is not always practical. For example, the impact of vouchers on between school competition, on sorting, and on the children remaining in public schools, could be analyzed by randomizing vouchers at the community level (assuming a community was large enough to incorporate several schools, some public and some private). If most children stayed within this community to go to school, a comparison of communities that (on a random basis) introduced vouchers with those that did not would tell us something about these general equilibrium effects.

In other cases, the general equilibrium effects may work at the level of the country or even the world (if, for example, they impact wages or prices). It would be difficult to implement a randomized evaluation that picked these up as it would involve randomizing at the national or even international level.

8.2 Hawthorne and John Henry Effects

Another limitation of prospective evaluations is that the evaluation itself may cause the treatment or comparison group to change its behavior. Changes in behavior among the treatment group are called Hawthorne effects, while changes in behavior among the comparison group are called John Henry effects.[25] The treatment group may be grateful to receive a treatment and conscious of being observed, which may induce them to alter their behavior for the duration of the experiment (for example, working harder to make it a success). The comparison group may feel offended to be a comparison group and react by also altering their behavior (for example, teachers in the comparison group for an evaluation may "compete" with the treatment teachers or, on the contrary, decide to slack off).

These behavioral responses are often discussed in the context of being specific concerns

[25]The Hawthorne Effect refers to the Hawthorne works of the Western Electric Company in Chicago. During a series of studies regarding the effect of work conditions on worker productivity, researchers concluded that the knowledge they were being observed induced workers to exert additional effort. The John Henry effect refers to the rail worker of American folklore.

for randomized evaluations, although similar effects can occur in other settings. For example, provision of school inputs could temporarily increase morale among students and teachers, which could improve performance in the short run. Such effects would create problems for fixed-effects, difference-in-differences and regression discontinuity estimates as well as randomized evaluations. What makes an experiment special is that individuals may know they are part of an evaluation and may thus react to the very fact of being evaluated, not only to the inputs received.

One way to disentangle Hawthorne or John Henry effects from long run impacts of the program which would obtain outside of an evaluation is to collect longer run data. For example, Duflo and Hanna (2006) continued to monitor the impact of the camera program over a year after the official "experiment" was over (but the NGO decided to continue to implement the program as a permanent program). The fact that the results are similar when the program is not being officially evaluated any more and at the beginning of the evaluation period suggest that the initial results on presence where not due to Hawthorne effects.

Evaluations can also be designed to help disentangle the various channels and help ameliorate concerns of John Henry or Hawthorne effects. One example is the evaluation of the SEED program, mentioned above (Ashraf, Karlan, and Yin 2006). In this case, the authors were concerned that individuals may have saved more not because of the special SEED savings program but just because they were visited by a team that suggested saving. Their solution was to create an additional treatment group to which they marketed a "regular" savings program. That is, of the half of these individuals not assigned to the SEED commitment treatment group, one-fourth were assigned to the pure comparison group while one-fourth were assigned to a third group—the "marketing treatment" group. Clients in this group were given virtually the same marketing campaign as received by clients in the SEED commitment treatment group, except that the marketing was strictly limited to conventional and existing savings products of the participating microfinance institution. By comparing savings levels of clients in the SEED commitment treatment and marketing treatment groups, the authors were able to isolate the direct effect of the SEED product from the effect of the marketing campaign.

While the coefficient on the indicator for the "marketing treatment" was insignificant in all regression specifications, it was also positive in every specification. This suggests that the marketing treatment may have had an impact on savings, though this was small in magnitude and the sample size did not provide sufficient statistical power to estimate it.

8.3 Generalizing Beyond Specific Programs and Samples

More generally, an issue that often comes up with randomized evaluations is the extent to which the results are replicable or generalizable to other contexts. Sure, a specific program worked in one community in Western Kenya, but can we extrapolate that it will work elsewhere? Was its success linked to a specific NGO? Would a similar program, but with minor variations, have the same impact?

Three major factors can affect the generalizability of randomized evaluation results: the ways in which the programs are implemented (are the programs implemented with special care in a way that makes it very difficult to replicate them?), the fact that the evaluation is conducted in a specific sample, and the fact that specific programs are implemented (would a slightly different program have the same results?). We consider these factors in turn.

The first question is whether the program is implemented with a level of care that makes it impossible to replicate it. Pilot programs are often run with particular care and with particularly high-quality program officials in a way that is impossible to replicate on a wider scale. When they are implemented by NGOs that are open to having programs tested, one may also worry that only these types of NGOs could implement them so effectively. As a program is expanded, the quality may deteriorate. While completely preventing this problem is difficult, it is important and possible to avoid "gold plating" programs. It is also important to clearly document the procedures followed in the program and to collect data on how well the program was implemented (in particular, the compliance rates and whether the procedures were followed) so that what is being evaluated is clearly understood.

A more difficult issue is whether we can conclude that because one population responded to a program in one way, another population will respond in the same way to a similar program. If a program worked for poor rural women in Africa, will it work for middle-income urban men in South Asia? This problem is of course not limited to randomized evaluations. Any empirical study informs us about the sample on which the research was performed, and can be generalized only under some assumptions. But for logistical reasons, randomized evaluations are often conducted in relatively small regions, which exacerbates the problem, while retrospective research can take advantage of nationally representative data sets.[26]

[26]It is nevertheless worth pointing out that in some cases, this trade-off between internal and external validity also occurs in non-experimental studies. For example, regression discontinuity designs solve internal validity

Given that implementation constraints require working in a small number of selected sites, the external validity of randomized evaluations for a given population (say, the population of a country) would be maximized by randomly selecting sites and, within these sites, by randomly selecting treatment and comparison groups. The former is almost never done. Randomized evaluations are typically performed in "convenience" samples, with specific populations. While these choices are often necessary to make an evaluation possible, they may also limit its external validity.

There are two responses to this dilemma. We should test and see whether a program or research result holds in different contexts. But as we cannot test every single permutation and combination of contexts, we must also rely on theories of behavior that can help us decide whether if the program worked in context A and B it is likely to work in C. We discuss these two responses below.

The third question is related: given that a specific version of a program had a given impact, what can we learn about similar, but not identical, programs. For example what would have been the response to the PROGRESA program if the slope of the transfer scheduled with respect to secondary school enrollment of children of different ages had been different? Here again, the same two responses hold: one would like to try various versions of programs to understand what versions matter. As the experience accumulate on a given programs, it may be possible to infer a "schedule" of responses to different transfer size for example. But the number of possible variations on a given program is potentially infinite, and a theoretical framework is definitely needed to understand which variations are important to replicate and which are not. Here again, it is a combination of replications and theory that can help generalize the lessons from a particular program.

8.4 Evidence on the Generalizability of Randomized Evaluation Results

Evidence on program replication is unfortunately limited at this point. The relatively limited experience of programs that have been tested in different environments suggest that (at least for these randomized trials) the results have generalized quite well. However, it will clearly be

issues with very mild identifying assumption by focusing on a very small subset of the population, those who are "marginal" for getting the treatment of interest; IV strategies identify the effect for a group of compliers that may be small and not representative of the general population.

important to do more of this type of testing on different types of programs. There may be another kind of publication incentive at work here: researchers do not have strong incentives to replicate existing evaluations, and journals are also less interested in publishing such studies. Ideally, institutions should emerge that both carry out such evaluations and ensure the diffusion of the results to policymakers, even if academic publications are not the ideal forum.

Replication was built into the design of the randomized evaluations of the remedial education discussed above (Banerjee, Duflo, Cole, and Linden 2007). The evaluation was simultaneously conducted in two large cities (Mumbai and Vadodara), with completely different implementation teams. Mumbai's team was experienced as the program had already been run in that city for some time; Vadodara's team was new. Mumbai and Vadodara are very different cities. Mumbai is a much richer city where initial learning levels were higher. The results were generally very similar in Mumbai and Vadodara with one interesting exception: in language, the effects in Mumbai were much smaller than in Vadodara, to the point of being insignificant. This is likely related to the fact that over 80% of the children in Mumbai had already mastered the basic language skills the program was covering, as the baseline tests demonstrated.

The deworming program discussed above was also replicated in a different context and produced similar results. The original deworming evaluation was run in rural primary schools in western Kenya. The program was then modified to meet the needs of preschool children in urban India, with iron supplementation added given the high level of anemia in that population (Bobonis, Miguel, and Sharma 2004). The program was also implemented by different organizations—Pratham in India and International Child Support Africa in Kenya. The results, however, were surprisingly similar. In Kenya, school participation increased by 7 percentage points, height-for-age z-score increased by .09, and weight-for-age did not change. In India, participation increased by 6 percentage points, there were no average gains in child height-for-age, and weight increased by 1.1 pounds.

One result that was different between the two studies was that externalities were much larger in Kenya than in India, possibly because of differences in the transmission mechanism of worms and because iron fortification does not produce externalities.

It is possible that medical interventions of this type are more likely to replicate than programs attempting to influence behavior. However, the limited evidence on the replicability of results from incentive programs is also encouraging. PROGRESA was initially introduced and tested

in Mexico but it has since been replicated in several other Latin American countries (including Ecuador, Colombia and Brazil) as well as in other countries, such as Turkey. In several cases, these programs have also been tested using randomized evaluations. These additional trials have allowed researcher both to verify that the effect of conditional cash transfers replicate in countries other than Mexico—they generally do—and also to shed some light on the importance of particular points of program design. For example, Schady and Araujo (2006) studied the impact on school enrollment in Ecuador of a cash transfer that was *not* conditional on school enrollment. They found that the program nevertheless had large impact on school enrollment, but this was concentrated among households who believed (mistakenly) that the program was conditional on enrollment.

8.5 Field Experiments and Theoretical Models

While it is necessary to replicate studies in different contexts, it will never be feasible to rigorously test the extent to which research results hold in all possible situations. However, experiments can deliver much more general lessons when they are combined with economic theories or models.

There are two main ways to combine theory and structure with randomized evaluations. First, economic modeling can be combined with variation coming from randomized evaluations to estimate a richer set of parameters. The cost is a set of additional assumptions, but the benefit is a richer set of parameters, which can be used to make prediction about how variation of the program would affect behavior. Attanasio, Meghir, and Santiago (2005) combine structural modeling with the variation coming from the experimental design in PROGRESA to estimate a flexible model of education choice, which allows them to estimate the possible effects of different variants of PROGRESA. A related use of randomized evaluation is an "out of sample" validation of the assumption made by structural models. Todd and Wolpin (2006) estimate a structural model of schooling and work decision for children in the PROGRESA control villages. They then simulate what effect PROGRESA would have if their model was right and compare it to the actual treatment effect.

A more ambitious use of theory and randomization in development economics is to set up experiments explicitly to test particular theories about economic behavior. Karlan and Zinman (2005c, 2005a) and Bertrand, Karlan, Mullainathan, Shafir, and Zinman (2005) are three related projects offering excellent examples of using field experiments to test theories. Both projects

were conducted in collaboration with a South African lender, giving small loans to high-risk borrowers at high interest rates. In both cases, the main manipulation started by sending different direct mail solicitation to different people. Karlan and Zinman (2005c) set out to test the relative weights of ex post repayment burden and ex ante adverse selection in lending. In their set up, potential borrowers with the same observable risk are randomly offered a high or a low interest rate in an initial letter. Individuals then decide whether to borrow at the solicitation's "offer" rate. Of those that respond to the high rate, half are randomly given a new lower "contract" interest rate when they actually apply for the loan, while the remaining half continue to receive the rate at which they were offered the loan. Individuals do not know beforehand that the contract rate may differ from the offer rate. The researchers then compare repayment performance of the loans in all three groups. This design allows the researchers to separately identify adverse selection effects and ex post repayment burden effects (which could be due to moral hazard or sheer financial distress ex post). Adverse selection effects are identified by considering only the sample that eventually received the low contract rate, and comparing the repayment performance of those who responded to the high offer interest rate with those who responded to the low offer interest rate. Ex post repayment burden effects are identified by considering only those who responded to the high offer rates, and comparing those who ended up with the low offer to those who ended up with the high offer. The study found that men and women behave differently: while women exhibit adverse selection, men exhibit moral hazard. This experiment constitutes a significant methodological advance because it shows how simple predictions from theory can be rigorously tested.

Bertrand, Karlan, Mullainathan, Shafir, and Zinman (2005) apply the same principle to a broader set of hypotheses, most of them coming directly from psychology. The experiment is overlaid on the Karlan and Zinman basic experiment: the offer letters are made to vary along other dimensions, which should not matter economically, but have been hypothesized by psychologists to matter for decision-making, and have been shown to have large effects in laboratory settings. For example, the lender varied the description of the offer, either showing the monthly payment for one typical loan or for a variety of loan terms and sizes. Other randomizations include whether and how the offered interest rate is compared to a "market" benchmark, the expiration date of the offer, whether the offer is combined with a promotional giveaway, race and gender features introduced via the inclusion of a photo in the corner of the

letter, and whether the offer letter mentions suggested uses for the loan. The analysis then compares the effect of all these manipulations. While not all of them make a difference, many do, and some of the effects are large and surprising. For example, for male customers, having a photo of a woman on top of the offer letter increases take-up as much as a 1% reduction in the monthly interest rate. In some sense, the juxtaposition of the two experiments may be the most surprising. On the one hand individuals react as *homo economicus* to information — they are sensitive to interest rates and poor-risk borrowers accept the highest interest rates (at least among women). On the other hand, these effects are present in the same setting where seemingly anodyne manipulations make a large difference.

The two experiments and many others already described in this chapter illustrate how development economists have gone much beyond "simple" program evaluations to use randomization as a research tool. Compared to retrospective evaluations (even perfectly identified ones), field experiments, when the collaboration with the partner is very close, offer much more flexibility and make it possible to give primacy to the hypothesis to test, rather than to the program that happens to have been implemented. With retrospective evaluations, theory is used instrumentally, as a way to provide a structure justifying the identifying assumptions (this is more or less explicit depending on the empirical tradition the researchers belong to). With prospective evaluations, it is the experimental design that is instrumental. This gives more power both to test the theory and to challenge it. A theoretical framework is necessary to suggest which experiments should be run and help give them a more general interpretation. Some of the most recent experimental results may not fit very well within the existing theories (this is what Banerjee (2005) calls the "new challenge to theory." They should prompt a back-and-forth between theoretical modeling and field experiments, with each new wave of results challenging existing theories and providing some direction about how to formulate new ones.

References

Angrist, J., E. Bettinger, E. Bloom, E. King, and M. Kremer (2002): "Vouchers for Private Schooling in Colombia: Evidence from a Randomized Natural Experiment," *American Economic Review*, 92(5), 1535–1558.

Angrist, J., E. Bettinger, and M. Kremer (2006): "Long-Term Educational Consequences of Secondary School Vouchers: Evidence from Administrative Records in Colombia," American Economic Review, forthcoming.

Angrist, J. D. (1990): "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records," *American Economic Review*, 80(3), 313–336.

——— (1998): "Estimating the Labor Market Impact of Voluntary Military Service Using Social Security Data on Military Applicants," *Econometrica*, 66(2), 249–88.

Angrist, J. D., and G. Imbens (1994): "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62(2), 467–475.

——— (1995): "Two Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity," *Journal of the American Statistical Association*, 90(430), 431–442.

Angrist, J. D., G. W. Imbens, and D. B. Rubin (1996): "Identification of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association*, 91, 444–455.

Ashenfelter, O., C. P. Harmon, and H. Oosterbeek (1999): "A Review of Estimates of the Schooling/Earnings Relationship," *Labour Economics*, 6(4), 453–470.

Ashraf, N., D. S. Karlan, and W. Yin (2006): "Tying Odysseus to the Mast: Evidence from a Commitment Savings Product in the Philippines," *Quarterly Journal of Economics*, forthcoming.

Attanasio, O., C. Meghir, and A. Santiago (2005): "Education choices in Mexico: using a structural model and a randomised experiment to evaluate Progresa," IFS working paper EWP05/01.

Banerjee (2005): "New Development Economics and the Challenge to Theory," *Economic and Political Weekly*, Vol. 40, 4340–4344.

Banerjee, A., P. Bardhan, K. Basu, R. Kanbur, and D. Mookherjee (2005): "New Directions in Development Economics: Theory or Empirics?," BREAD Working Paper No. 106, A Symposium in Economic and Political Weekly.

Banerjee, A., and E. Duflo (2005): "Addressing Absence," *Journal of Economic Perspectives*, forthcoming.

Banerjee, A., E. Duflo, S. Cole, and L. Linden (2007): "Remedying Education: Evidence from Two Randomized Experiments in India," Quarterly Journal of Economics, forthcoming.

Bardhan, P. (2005): "Theory or Empirics in Development Economics," mimeo, University of California at Berkeley.

Basu, K. (2005): "The New Empirical Development Economics: Remarks on Its Philosophical Foundations," *Economic and Political Weekly*.

Bertrand, M., S. Djankov, R. Hanna, and S. Mullainathan (2006): "Does Corruption Produce Unsafe Drivers?," NBER Working Paper #12274.

Bertrand, M., D. S. Karlan, S. Mullainathan, E. Shafir, and J. Zinman (2005): "What's Psychology Worth? A Field Experiment in the Consumer Credit Market," Working Papers 918, Economic Growth Center, Yale University, Available at http://ideas.repec.org/p/egc/wpaper/918.html.

Bhushan, I., S. Keller, and B. Schwartz (2002): "Achieving the twin objectives of efficiency and equity: contracting health services in Cambodia," *ERD Policy Brief Series, Asian Development Bank.*, 6.

Bloom, E., I. Bhushan, D. Clingingsmith, R. Hung, E. King, M. Kremer, B. Lovinsohn, and B. Schwartz (2006): "Contracting for Health: Evidence from Cambodia," Mimeo.

Bloom, H. S. (1995): "Minimum detectable effects: A simple way to report the statistical power of experimental designs.," *Evaluation Review*, 19, 547–56.

Bloom, H. S. (2005): *Randomizing groups to evaluate place-based programs*NY: Russell Sage Foundation, chap. Learning more from social experiments, pp. 115–172.

Bobonis, G. J., E. Miguel, and C. P. Sharma (2004): "Iron Deficiency Anemia and School Participation," Poverty Action Lab Paper No. 7.

Buddlemeyer, H., and E. Skofias (2003): "An Evaluation on the Performance of Regression Discontinuity Design on PROGRESA," Institute for Study of Labor, Discussion Paper No. 827.

Campbell, D. T. (1969): "Reforms as Experiments," *American Psychologist*, 24, 407–429.

Card, D. (1999): "The Causal Effect of Education on Earnings," in *Handbook of Labor Economics, Vol. 3A*, ed. by O. Ashenfelter, and D. Card. North Holland, pp. 1801–63.

Card, D., and A. B. Krueger (1995): *Myth and Measurement: The New Economics of the Minimum Wage*. Princeton, NJ: Princeton University Press.

Chattopadhyay, R., and E. Duflo (2004): "Women as Policy Makers: Evidence from a Randomized Policy Experiment in India," *Econometrica*, 72(5), 1409–1443.

Cohen, J. (1988): *Statistical Power Analysis for the Behavioral Science*. Hillsdale, NJ: Lawrence Erlbaum, 2nd edition edn.

Cook, T. D., W. R. Shadish, and V. C. Wong (2006): "Within Study Comparisons of Experiments and Non-Experiments: Can they help decide on Evaluation Policy," mimeo, Northwestern University.

Cox, D., and N. Reid (2000): *Theory of the Design of Experiments*. London: Chapman and Hall.

Das, J., P. Krishnan, J. Habyarimana, and S. Dercon (2004): "When can school inputs improve test scores?," World Bank Policy Research working paper; no. WPS 3217.

Deaton, A. (1997): *The Analysis of Household Surveys*. World Bank, International Bank for Reconstruction and Development.

Dehejia, R. H., and S. Wahba (1999): "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs," *Journal of the American Statistical Association*, 94(448), 1053–62.

DeLong, J. B., and K. Lang (1992): "Are All Economic Hypotheses False?," *Journal of Political Economy*, 100:6, 1257–72.

Diaz, J. J., and S. Handa (2006): "An Assessment of Propensity Score Matching as a Non Experimental Impact Estimator: Evidence from Mexicos Progresa Program," forthcoming, Journal of Human Resources.

Dickson, Rumona, S. A. P. W. C. D., and P. Garner (2000): "Effect of treatment for intestinal helminth infection on growth and cognitive performance in children: systematic review of randomized trials," *British Medical Journal*, 320, 1697–1701.

Donald, S., and K. Lang (2001): "Inference with Differences-in-Differences and Other Panel Data," Discussion paper, Boston University Department of Economics.

Duflo, E. (2004): *Scaling Up and Evaluation*Oxford University Press and World Bank, chap. Accelerating Development.

——— (2006): "Field Experiments in Development Economics," Discussion paper.

Duflo, E., P. Dupas, M. Kremer, and S. Sinei (2006): "Education and HIV/AIDS Prevention: Evidence from a randomized evaluation in Western Kenya," mimeo, MIT.

Duflo, E., and R. Hanna (2006): "Monitoring Works: Getting Teachers to Come to School," NBER Working Paper No. 11880.

Duflo, E., and M. Kremer (2005): "Use of Randomization in the Evaluation of Development Effectiveness," in *Evaluating Development Effectiveness*, ed. by O. Feinstein, G. K. Ingram, and G. K. Pitman. New Brunswick, New Jersey and London, U.K.: Transaction Publishers, vol. 7, pp. 205–232.

Duflo, E., M. Kremer, and J. Robinson (2006): "Understanding Technology Adoption: Fertilizer in Western Kenya, Preliminary Results from Field Experiments," mimeo.

Duflo, E., S. Mullainathan, and M. Bertrand (2004): "How Much Should We Trust Difference in Differences Estimates?," *Quarterly Journal of Economics*, 119(1), 249–275.

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(3), 2003, pp. 815-842*, 118(3), 815–842.

Dupas, P. (2006): "Relative Risks and the Market for Sex: Teenagers, Sugar Daddies, and HIV in Kenya," mimeo, Dartmouth College.

Fisher, R. A. (1926): "The Arrangement of Field Experiments," *Journal of the Ministry of Agriculture*, 33, 503–513.

Gertler, P. J., and S. Boyce (2001): "An Experiment in Incentive-Based Welfare: The Impact of PROGRESA on Health in Mexico," Mimeo, UC-Berkeley.

Glazerman, S., D. Levy, and D. Myers (2003): *Nonexperimental Replications of Social Experiments: A Systematic Review*. Princeton, NJ: Mathematica Policy Research, Inc.

Glewwe, Paul, N. I., and M. Kremer (2003): "Teacher Incentives," Working Paper 9671, National Bureau of Economic Research.

Glewwe, P., Kremer, and S. Moulin (2004): "Textbooks and Test Scores: Evidence from a Prospective Evaluation in Kenya," Mimeo, Harvard University.

Glewwe, P., and M. Kremer (2005): "Schools, Teachers, and Education Outcomes in Developing Countries," Handbook on the Economics of Education (forthcoming).

Glewwe, P., M. Kremer, S. Moulin, and E. Zitzewitz (2004): "Retrospective Vs. Prospective Analyses of School Inputs: The Case of Flip Charts in Kenya," *Journal of Development Economics*, 74, 251 268.

Grasdal, A. (2001): "The performance of sample selection estimators to control for attrition bias," *Health Economics*, 10, 385–398.

Harrison, G., and J. A. List (2004): "Field Experiments," *Journal of Economic Literature*, XLII, 1013–1059.

Hausman, J. A., and D. A. Wise (1979): "Attrition Bias in Experimental and Panel Data: The Gary Income Maintenance Experiment," *Econometrica*, 47(2), 455–73.

Heckman, J., H. Ichimura, J. Smith, and P. Todd (1998): "Characterizing Selection Bias Using Experimental Data," *Econometrica*, 66(5), 1017–1098.

Heckman, J., H. Ichimura, and P. Todd (1997): "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme," *Review of Economic Studies*, 64(4), 605–54.

——— (1998): "Matching as an Econometric Evaluation Estimator," *Review of Economic Studies*, 65(2), 261–94.

Heckman, J., and Vytlacil (2005): "Structural Equations, Treatment Effects and Econometric Policy Evaluation," *Econometrica*, 73(3), 669–738.

Heckman, J. J. (1979): "Sample Selection Bias as a Specification Error," *Econometrica*, 47(1), 153–161.

Heckman, J. J., L. Lochner, and C. Taber (1998): "General Equilibrium Treatment Effects: A Study of Tuition Policy," Working Paper 6426, National Bureau of Economic Research.

Hedges, L. (1992): "Modeling publication selection effects in meta-analysis," *Statistical Science*, 7, 227–236.

Hedges, L. V., and I. Olkin (1985): *Statistical methods for meta-analysis*. San Diego, CA: Academic Press.

Holland, P. W. (1988): "Causal Inference, Path Analysis, and Recursive Structural Equations Models," *Sociological Methodology*, 18, 449–484.

Holm (1979): "A Simple Sequentially Rejective Multiple Test Procedure," *Scandinavian Journal of Statistics*, 6, 65–70.

Hoxby, C. M. (2003): "School Choice and School Productivity (Or, Could School Choice be a Rising Tide that Lifts All Boats," Chicago: University of Chicago Press, chap. The Economics of School Choice.

Hsieh, C.-T., and M. S. Urquiola (2003): "When Schools Compete, How Do They Compete? An Assessment of Chile's Nationwide School Voucher Program," NBER Working Paper No. W10008.

Imbens, G., G. King, and G. Ridder (2006): "On the Benefits of Stratification in Randomized Experiments," Mimeo, Harvard.

Imbens, G. W. (2004): "Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review," *The Review of Economics and Statistics*, 86(1), 4–29.

International Food Policy Research (2000): "Monitoring and Evaluation System," Discussion paper, International Food Policy Research (IFPRI).

Karlan, and J. Zinman (2005a): "Elasticities of Demand for Consumer Credit," mimeo, Yale University.

——— (2006a): "Expanding Credit Access: Using Randomized Supply Decisions To Estimate the Impacts," mimeo, Yale University.

Karlan, D., and J. Zinman (2006b): "Observing Unobservables: Identifying Information Asymmetries with a Consumer Credit Field Experiment," mimeo.

Karlan, D. S., and J. Zinman (2005b): "Observing Unobservables: Identifying Information Asymmetries with a Consumer Credit Field Experiment," Working Papers 911, Economic Growth Center, Yale University, Available at http://ideas.repec.org/p/egc/wpaper/911.html.

——— (2005c): "Observing Unobservables: Identifying Information Asymmetries with a Consumer Credit Field Experiment," Discussion paper, type: Working Papers; note: Available at http://ideas.repec.org/p/egc/wpaper/911.html.

Keller, S., and B. Schwartz (2001): "Final Evaluation Report: Contracting for Health Services Pilot Project," Unpublished Asian Development Bank Report on Loan No. 1447-CAM.

Kling, and Liebman (2004): "Experimental Analysis of Neighborhood Effects on Youth," KSG Working Paper No. RWP04-034.

Kling, J. R., J. B. Liebman, L. F. Katz, and L. Sanbonmatsu (2004): "Moving To Opportunity and Tranquility: Neighborhood Effects on Adult Economic Self-sufficiency and Health from a Randomized Housing Voucher Experiment," .

Kremer, M. (2003): "Randomized Evaluations of Educational Programs in Developing Countries: Some Lessons," *American Economic Review*, 93(2), 102–106.

Kremer, M., E. Miguel, and R. Thornton (2004): "Incentives to Learn," mimeo, Harvard University.

Krueger, A., and D. Whitmore (2002): "Would smaller classes help close the blackwhite achievement gap?," in *Bridging the achievement gap.*, ed. by J. Chubb, and T. Loveless. Washington, DC: Brookings Institute Press, chap. Bridging the achievement gap.

Lalonde, R. J. (1986): "Evaluating the Econometric Evaluations of Training Programs Using Experimental Data," *American Economic Review*, 76(4), 602–620.

Lee, D. S. (2002): "Trimming for Bounds on Treatment Effects with Missing Outcomes," Working Paper 51.

Liebman, J. B., L. F. Katz, and J. Kling (2004): "Beyond Treatment Effects: Estimating the Relationship between Neighborhood Poverty and Individual Outcomes in the MTO Experiment," Princeton IRS Working Paper 493.

Logan, and Tamhane (2003): "Accurate critical constants for the one-sided approximate likelihood ratio test of a normal mean vector when the covariance matrix is estimated," *Biometrics*, 58, 650–656.

Maluccio, J. A., and R. Flores (2005): "Impact Evaluation of a Conditional Cash Transfer Program," Discussion paper, International Food Policy Research Institute, Research Report No. 141.

Manski, C. (1993): "Identification of Exogenous Social Effects: The Reflection Problem," *Review of Economic Studies*, 60, 531–542.

Manski, C. F. (1989): "Schooling as experimentation: a reappraisal of the postsecondary dropout phenomenon," *Economics of Education Review*, 8(4), 305–312.

Miguel, E., and M. Kremer (2003): "Networks, Social Learning, and Technology Adoption: The Case of Deworming Drugs in Kenya," Working Paper 61.

——— (2004): "Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities," *Econometrica*, 72(1), 159–218.

Mookherjee, D. (2005): "Is There Too Little Theory in Development Economics?," *Economic and Political Weekly*.

Morduch, J. (1998): "Does Microfinance Really Help the Poor? New Evidence from Flagship Programs in Bangladesh," Mimeo, Princeton University.

Moulton, B. R. (1990): "An Illustration of a Pitfall in Estimating the Effects of Aggregate Variables on Micro Units," *Review of Economics and Statistics*, 72(2), 334–338.

O'Brien (1984): "Procedures for Comparing Samples with Multiple Endpoints," *Biometrics*, 40, 1079–1087.

Olken, B. (2005): "Monitoring Corruption: Evidence from a Field Experiment in Indonesia," mimeo, Harvard University.

Parker, S., L. Rubalcava, and G. Teruel (2005): "Evaluating Conditional Schooling-Health Transfer Programs (PROGRESA Program)," Forthcoming in Handbook of Development Economics, Volume 4.

Pitt, M., and S. Khandker (1998): "The Impact of Group-Based Credit Programs on Poor Households in Bangladesh: Does the Gender of Participants Matter," *Journal of Political Economy*.

Raudenbush, S. W., J. Spybrook, X. Liu, and R. Congdon (2005): "Optimal Design for Longitudinal and Multilevel Research: Documentation for the Optimal Design Software," Retrieved April 15, 2005 from http://sitemaker.umich.edu/group-based/optimal_design_software.

Ravallion, M. (2006): "Evaluating Anti-Poverty Programs," forthcoming in Handbook of Development Economics Volume 4, edited by Robert E. Evenson and T. Paul Schultz, Amsterdam: North-Holland.

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

Rosenbaum, P. R. (2002): "Covariance adjustment in randomized experiments and observational studies (with discussion)," *Statistical Science*, 17, 286327.

Rubin (1974): "Estimating Causal Effects of Treatments in Randomized and Non-randomized Studies," *Journal of Educational Psychology*, 66, 688–701.

Savin, N. E. (1984): *Multiple Hypothesis Testing*Amsterdam: Elsevier Science Publishers BV, chap. Handbook of Econometrics, Volume II, pp. 827–879.

Schady, N. R., and M. C. Araujo (2006): "Cash, Conditions, School Enrollment, and Child Work: Evidence from a Randomized Experiment in Ecuador," Unpublished manuscript.

Schultz, T. P. (2004): "School subsidies for the poor: Eevaluating the Mexican PROGRESA poverty program," *Journal of Development Economics*, 74(1), 199–250.

Schwartz, B., and I. Bhushan (2004): "Reducing Inequity In The Provision Of Primary Health Care Services: Contracting In Cambodia," mimeo.

Smith, J., and P. Todd (2005): "Does Matching Overcome Lalondes Critique of Nonexperimental Estimators?," *Journal of Econometrics*, 125(1-2), 305–353.

Thomas, D., E. Frankenberg, J. Friedman, J.-P. Habicht, and E. Al (2003): "Iron Deficiency and the Well Being of Older Adults: Early results from a randomized nutrition intervention," Mimeo, UCLA.

Todd, P. (2006): "Evaluating Social Programs with Endogenous Program Placement and Self Selection of the Treated," forthcoming in the Handbook of Development Economics, Volume 4.

Todd, P. E., and K. I. Wolpin (2006): "Ex Ante Evaluation of Social Programs," Mimeo, University of Pennsylvania.

Vermeersch, and Kremer (2004): "School Meals, Educational Achievement, and School Competition: Evidence from a Randomized Evaluation," World Bank Policy Research Working Paper No. 3523.

Westfall, P., and S. Young (1993): *Resampling-based multiple testing: Examples and methods for p-value adjustement.* New York: Wiley.

Wooldridge (2002): "Inverse Probability Weighted M-Estimators for Sample Selection, Attrition, and Stratification," *Portuguese Economic Journal*, 1, 117–139.

Wooldridge, J. M. (2004): "Cluster-Sample Methods in Applied Econometrics," *American Economic Review*, 93(2), 133–138.

Table 1: Intra-class corrrelation, primary schools

Location	Subject	Estimate	Reference
Madagascar	Math+language	0.5	AGEPA data base
Busia, Kenya	Math+language	0.22	Miguel and Kremer (2004)
Udaipur, India	Math+language	0.23	Duflo and Hanna (2005)
Mumbai, India	Math+language	0.29	Banerjee et al. (2007)
Vadodara, India	Math+language	0.28	Banerjee et al. (2007)
Busia, Kenya	Math	0.62	Glewwe et al (2004)
Busia, Kenya	Language	0.43	Glewwe et al (2004)
Busia, Kenya	Science	0.35	Glewwe et al (2004)