MDRC Working Papers on Research Methodology

The Core Analytics of Randomized Experiments for Social Research

Howard S. Bloom

mdrc
BUILDING KNOWLEDGE TO IMPROVE SOCIAL POLICY

August 2006

This working paper is part of a series of publications by MDRC on alternative methods of evaluating the implementation and impacts of social and educational programs and policies. The paper will be published as a chapter in the forthcoming *Handbook of Social Research* by Sage Publications, Inc.

Many thanks are due to Richard Dorsett, Carolyn Hill, Rob Hollister, and Charles Michalopoulos for their helpful suggestions on revising earlier drafts.

This working paper was supported by the Judith Gueron Fund for Methodological Innovation in Social Policy Research at MDRC, which was created through gifts from the Annie E. Casey, Rockefeller, Jerry Lee, Spencer, William T. Grant, and Grable Foundations.

The findings and conclusions in this paper do not necessarily represent the official positions or policies of the funders.

Dissemination of MDRC publications is supported by the following funders that help finance MDRC's public policy outreach and expanding efforts to communicate the results and implications of our work to policymakers, practitioners, and others: Alcoa Foundation, The Ambrose Monell Foundation, The Atlantic Philanthropies, Bristol-Myers Squibb Foundation, Open Society Institute, and The Starr Foundation. In addition, earnings from the MDRC Endowment help sustain our dissemination efforts. Contributors to the MDRC Endowment include Alcoa Foundation, The Ambrose Monell Foundation, Anheuser-Busch Foundation, Bristol-Myers Squibb Foundation, Charles Stewart Mott Foundation, Ford Foundation, The George Gund Foundation, The Grable Foundation, The Lizabeth and Frank Newman Charitable Foundation, The New York Times Company Foundation, Jan Nicholson, Paul H. O'Neill Charitable Foundation, John S. Reed, The Sandler Family Supporting Foundation, and The Stupski Family Fund, as well as other individual contributors.

For information about MDRC and copies of our publications, see our Web site: www.mdrc.org.

Copyright © 2006 by MDRC. All rights reserved.

Abstract

This chapter examines the core analytic elements of randomized experiments for social research. Its goal is to provide a compact discussion for faculty members, graduate students, and applied researchers of the design and analysis of randomized experiments for measuring the impacts of social or educational interventions. Design issues considered include choosing the size of a study sample and its allocation to experimental groups, using covariates or blocking to improve the precision of impact estimates, and randomizing intact groups instead of individuals. Analysis issues considered include estimating impacts when not all sample members comply with their assigned treatment and estimating impacts when groups are randomized.

Table of Contents

Acknowledgments	ii
Abstract	iii
List of Tables and Figures	vii
Introduction	1
Why Randomize?	3
A Simple Experimental Estimator of Causal Effects	3
Choosing a Sample Size and Allocation	4
Estimating Causal Effects with Noncompliance	7
Using Covariates and Blocking to Improve Precision	11
Randomizing Groups to Estimate Intervention Effects	14
Future Frontiers	18
Report Tables and Figures	20
References	24
Earlier MDRC Working Papers on Research Methodology	31
About MDRC	33

List of Tables and Figures

Table

1	Minimum Detectable Effect and Effect Size for Individual Randomization	20
2	Minimum Detectable Effect Size for Balanced Group Randomization with $\rho = 0.10$	21

Figure

1	The Minimum Detectable Effect Multiplier	22
2	A Hypothetical Experiment Including No-Shows and Crossovers	23

Introduction

This chapter introduces the central analytic principles of randomized experiments for social research. Randomized experiments are lotteries that randomly assign subjects to research groups, each of which is offered a different treatment. When the method is implemented properly, differences in future outcomes for experimental groups provide unbiased estimates of differences in the impacts of the treatments offered. The method is usually attributed to Ronald A. Fisher (1925 and 1935), who developed it during the early 1900s.[1] After World War II, randomized experiments gradually became the method of choice for testing new drugs and medical procedures, and to date over 350,000 randomized clinical trials have been conducted (Cochrane Collaboration, 2002).[2]

Numerous books have been written about randomized experiments as their application has expanded from agricultural and biological research (e.g., Fisher, 1935; Cochran and Cox, 1957; Kempthorne, 1952; and Cox, 1958) to research on industrial engineering (e.g., Box, Hunter, and Hunter, 2005), to educational and psychological research (e.g., Lindquist, 1953, and Myers, 1972) to social science and social policy research (e.g., Boruch, 1997; Orr, 1999; and Bloom, (2005a). In addition, several journals have been established to promote advancement of the method (e.g., the *Journal of Experimental Criminology, Clinical Trials* and *Controlled Clinical Trials*).

The use of randomized experiments for social research has greatly increased since the War on Poverty in the 1960s. The method has been used in laboratories and in field settings to randomize individual subjects, such as students, unemployed adults, patients, or welfare recipients, and intact groups, such as schools, firms, hospitals, or neighborhoods.[3] Applications of the method to social research have examined issues such as child nutrition (Teruel and Davis, 2000); child abuse (Olds, et al., 1997); juvenile delinquency (Lipsey, 1988); policing strategies (Sherman and Weisburd, 1995); child care (Bell et al., 2003); public education (Kemple and Snipes, 2000); housing assistance (Orr et al., 2003); health insurance (Newhouse, 1996); income maintenance (Munnell, 1987); neighborhood effects (Kling, Liebman, and Katz, forthcoming); job training

[1] References to randomizing subjects to compare treatment effects date back to the seventeenth century (Van Helmont, 1662), although the earliest documented use of the method was in the late nineteenth century for research on sensory perception (Peirce and Jastrow, 1884/1980). There is some evidence that randomized experiments were used for educational research in the early twentieth century (McCall, 1923). But it was not until Fisher (1925 and 1935) combined statistical methods with experimental design that the method we know today emerged.

[2] Marks (1997) provides an excellent history of this process.

[3] See Bloom (2005a) for an overview of group-randomized experiments; see Donner and Klar (2000) and Murray (1998) for textbooks on the method.

(Bloom et al., 1997); unemployment insurance (Robins and Spiegelman, 2001); welfare-to-work (Bloom and Michalopoulos, 2001); and electricity pricing (Aigner, 1985).[4]

A successful randomized experiment requires clear specification of five elements.

1. ***Research questions:*** What treatment or treatments are being tested? What is the counterfactual state (in the absence of treatment) with which treatments will be compared? What estimates of net impact (the impact of specific treatments versus no such treatments) are desired? What estimates of differential impact (the difference between impacts of two or more treatments) are desired?

2. ***Experimental design:*** What is the unit of randomization: individuals or groups? How many individuals or groups should be randomized? What portion of the sample should be randomized to each treatment or to a control group? How, if at all, should covariates, blocking, or matching (explained later) be used to improve the precision of impact estimates?

3. ***Measurement methods:*** What outcomes are hypothesized to be affected by the treatments being tested, and how will these outcomes be measured? What baseline characteristics, if any, will serve as covariates, blocking factors, or matching factors, and how will these characteristics be measured? How will differences in treatments be measured?

4. ***Implementation strategy:*** How will experimental sites and subjects be recruited, selected, and informed? How will they be randomized? How will treatments be delivered and how will their differences across experimental groups be maintained? What steps will be taken to ensure high-quality data?

5. ***Statistical analysis:*** The analysis of treatment effects must reflect how randomization was conducted, how treatment was provided, and what baseline data were collected. Specifically it must account for: (1) whether randomization was conducted or treatment was delivered in groups or individually; (2) whether simple randomization was conducted or randomization occurred within blocks or matched pairs; and (3) whether baseline covariates were used to improve precision.

This chapter examines the analytic core of randomized experiments — design and analysis, with a primary emphasis on design.

[4]For further examples, see Greenberg and Shroder, 1997.

Why Randomize?

There are two main reasons why randomized experiments are the most rigorous way to measure causal effects.

They eliminate bias: Randomizing subjects to experimental groups eliminates all *systematic* preexisting group differences, because only chance determines which subjects are assigned to which groups. Consequently, each experimental group has the same expected values for all characteristics, observable or not. Randomization of a given sample may produce experimental groups that differ by chance, however. These differences are random errors, not biases. Hence, the absence of bias is a property of the *process* of randomization, not a feature of its application to a specific sample. The laws of probability ensure that the larger the experimental sample, the smaller preexisting group differences are likely to be.

They enable measurement of uncertainty: Experiments randomize all sources of uncertainty about impact estimates for a given sample (their internal validity). Hence, confidence intervals or tests of statistical significance can account for all of this uncertainty. No other method for measuring causal effects has this property. One cannot, however, account for all uncertainty about generalizing an impact estimate beyond a given sample (its external validity) without both randomly *sampling* subjects from a known population and randomly *assigning* them to experimental groups (which is rarely possible in social research).[5]

A Simple Experimental Estimator of Causal Effects

Consider an experiment where half of the sample is randomized to a treatment group that is offered an intervention and half is randomized to a control group that is not offered the intervention, and everyone adheres to their assigned treatment. Follow-up data are obtained for all sample members and the treatment effect is estimated by the difference in mean outcomes for the two groups, $\bar{Y}_T - \bar{Y}_C$. This difference provides an unbiased estimate of the average treatment effect (ATE) for the study sample, because the mean outcome for control group members is an unbiased estimate of what the mean outcome would have been for treatment group members had they not been offered the treatment (their counterfactual).

However, any given sample can yield a treatment group and control group with preexisting differences that occur solely by chance and can overestimate or underestimate the ATE. The standard error of the impact estimator $(SE(\bar{Y}_T - \bar{Y}_C))$ accounts for this random error, where:

[5]Two major studies that used random sampling *and* random assignment are the national evaluations of Head Start (Puma et al., 2006) and the Job Corps (Schochet, 2006).

$$SE(\bar{Y}_T - \bar{Y}_C) = \sqrt{\frac{\sigma^2}{n_T} + \frac{\sigma^2}{n_C}} \quad (1)$$

given:

n_T and n_C = the number of treatment group members and control group members,

σ^2 = the pooled outcome variance across subjects within experimental groups.[6]

The number of treatment group members and control group members are experimental design decisions. The variance of the outcome measure is an empirical parameter that must be "guesstimated" from previous research when planning an experiment and can be estimated from follow-up data when analyzing experimental findings. For the discussion that follows it is useful to restate Equation 1 as:

$$SE(\bar{Y}_T - \bar{Y}_C) = \sqrt{\frac{\sigma^2}{nP(1-P)}} \quad (2)$$

where n equals the total number of experimental sample members ($n_T + n_C$) and P equals the proportion of this sample that is randomized to treatment.[7]

Choosing a Sample Size and Allocation

The first steps in designing a randomized experiment are to specify its treatment, target group, and setting. The next steps are to choose a sample size and allocation that maximize precision given existing constraints. For this purpose, it is useful to measure precision in terms of minimum detectable effects (Bloom, 1995 and 2005b). Intuitively, a minimum detectable effect is the smallest true treatment effect that a research design can detect with confidence. Formally, it is the smallest true treatment effect that has a specified level of statistical power for a particular level of statistical significance, given a specific statistical test.

Figure 1 illustrates that the minimum detectable effect of an impact estimator is a multiple of its standard error. The first bell-shaped curve (on the left of the figure) represents a t distribution for a null hypothesis of zero impact. For a positive impact estimate to be statistically significant at the α level with a one-tail test (or at the $\alpha/2$ level with a two-tailed test), the estimate must fall to the right of the critical t-value, t_α (or $t_{\alpha/2}$), of the first distribution. The second bell-shaped curve represents a t distribution for an alternative hypothesis that the true impact equals a specific minimum detectable effect. To have a probability (1 – B) of detecting the minimum detectable effect it must lie a distance of t_{1-B} to the right of the critical t-value for the

[6]The present discussion assumes a common outcome variance for the treatment and control groups.

[7]Note that Pn equals n_T and (1-P)n equals n_C.

null hypothesis. (The probability (1 – B) represents the level of statistical power.) Hence the minimum detectable effect must lie a total distance of $t_{\alpha} + t_{1-B}$ (or $t_{\alpha/2} + t_{1-B}$) from the null hypothesis. Because t-values are multiples of the standard error of an impact estimator, the minimum detectable effect is either $t_{\alpha} + t_{1-B}$ (for a one-tail test) or $t_{\alpha/2} + t_{1-B}$ (for a two-tail test) times the standard error. These critical t values depend on the number of degrees of freedom.

A common convention for defining minimum detectable effects is to set statistical significance (α) at 0.05 and statistical power (1 – B) at 80 percent. When the number of degrees of freedom exceeds about 20, the multiplier equals roughly 2.5 for a one-tail test and 2.8 for a two-tail test.[8] Thus, if the standard error of an estimator of the average effect of a job-training program on future annual earnings were $500, the minimum detectable effect would be roughly $1,250 for a one-tail test and $1,400 for a two-tail test.

Consider how this applies to the experiment described above. The multiplier, M_{n-2}[9], times the standard error, $SE\ (\bar{Y}_T - \bar{Y}_C)$, yields the minimum detectable effect:

$$MDE(\bar{Y}_T - \bar{Y}_C) = M_{n-2}\sqrt{\frac{\sigma^2}{nP(1-P)}} \tag{3}$$

Since the multiplier M_{n-2} is the sum of two t-values, determined by the chosen of levels of statistical significance and power, the missing value that needs to be determined for the sample design is that for σ^2. This value will necessarily be a guess, but since it is a central determinant of the minimum detectable effect, it should be based on a careful search of empirical estimates for closely related studies.[10]

Sometimes impacts are measured as a standardized mean difference or "effect size," either because the original units of the outcome measures are not meaningful or because outcomes in different metrics must be combined or compared. (There is no reason to standardize the impact estimate for the preceding job training example.) The standardized mean difference effect size (ES) equals the difference in mean outcomes for the treatment group and control group, divided by the standard deviation of outcomes across subjects within experimental groups, or:

[8]When the number of degrees of freedom becomes smaller, the multiplier becomes larger as the t distribution becomes fatter in its tails.

[9]The subscript n-2 equals the number of degrees of freedom for a treatment and control group difference of means, given a common variance for the two groups.

[10]When the outcome measure is a one/zero binary variable (e.g., employed =1 or not employed =0) the variance estimate is p(1-p)/n where p is the probability of a value equal to one. The usual conservative practice in this case is to choose p=.5, which yields the maximum possible variance.

$$ES = \frac{\bar{Y}_T - \bar{Y}_C}{\sigma} \tag{4}$$

Some researchers use the pooled within-group variance to define effect sizes while others use the control-group variance. Standardized mean effect sizes are therefore measured in units of standard deviations. For example, an effect size of 0.25 implies an impact equal to 0.25 standard deviations. When impacts are reported in effect size, precision can be reported as a minimum detectable effect size, where:

$$MDES(\bar{Y}_T - \bar{Y}_C) = M_{n-2}\sqrt{\frac{1}{nP(1-P)}} \tag{5}$$

Table 1 illustrates the implications of Equations 3 and 5 for the relationship between sample size, allocation, and precision. The top panel in the table presents minimum detectable effects for a hypothetical job training program, given a standard deviation for the outcome (annual earnings) of $1,000. The bottom panel presents corresponding minimum detectable effect sizes.

The first main observation is that increasing sample size has a diminishing return for precision. For example, the first column in the table illustrates how the minimum detectable effect (or effect size) declines with an increase in sample size for a balanced allocation (P = 0.5). Doubling the sample size from 50 individuals to 100 individuals reduces the minimum detectable effect from approximately $810 to $570 or by a factor of $1/\sqrt{2}$. Doubling the sample size again from 100 to 200 individuals reduces the minimum detectable effect by another factor of $1/\sqrt{2}$ from approximately $570 to $400. Thus, quadrupling the sample cuts the minimum detectable effect in half. The same pattern holds for minimum detectable effect sizes.

The second main observation is that for a given sample size, precision decreases slowly as the allocation between the treatment and control groups becomes more imbalanced. Equation 5 implies that the minimum detectable effect size is proportional to $1/\sqrt{P(1-P)}$, which equals 2.00, 2.04, 2.18, 2.50, or 3.33 when P (or its complement) equals 0.5, 0.6, 0.7, 0.8, or 0.9. Thus, for a given sample size, precision is best with a balanced allocation (P = 0.5). Because precision erodes slowly until the degree of imbalance becomes extreme (roughly $P \leq 0.2$ or $P \geq 0.8$), there is considerable latitude for using an unbalanced allocation. Thus, when political pressures to minimize the number of control group members are especially strong, one could use a relatively small control group. Or when the costs of treatment are particularly high, one could use a relatively large control group.[11]

[11]The preceding discussion makes the conventional assumption that σ^2 is the same for the treatment and control groups. But if the treatment affects different sample members differently, it can create a σ^2 for the treatment (continued)

One of the most difficult steps in choosing a sample design is selecting a target minimum detectable effect or effect size. From an economic perspective, this target should equal the smallest true impact that would produce benefits that exceed intervention costs. From a political perspective, it should equal the smallest true impact deemed policy-relevant. From a programmatic perspective, the target should equal the smallest true impact that exceeds known impacts from related interventions.

The most popular benchmark for gauging standardized effect sizes is Cohen's (1977, 1988) prescription (based on little empirical evidence) that values of 0.20, 0.50, and 0.80 be considered small, moderate, and large. Lipsey (1990) subsequently provided empirical support for this prescription from a synthesis of 186 meta-analyses of intervention studies. The bottom third of effect sizes in Lipsey's synthesis ranges from 0.00 to 0.32, the middle third ranges from 0.33 to 0.55, and the top third ranges from 0.56 to 1.20. Both authors suggest, however, that their general guidelines do not apply to many situations. For example, recent research suggests that much smaller effect sizes are policy-relevant for educational interventions. Findings from the Tennessee Class Size Experiment indicate that reducing elementary school class size from 22-26 students to 13-17 students increased performance on standardized reading and math tests by 0.1 to 0.2 standard deviations (Nye, Hedges, and Konstantopoulos, 1999). More recently, Kane's (2004) study of grade-to-grade improvement in math and reading on a nationally normed test suggests that one full year of elementary school *attendance* increases student achievement by roughly 0.25 standard deviations. These results highlight the importance of basing decisions about needed precision on the best existing evidence for the context being studied.

Estimating Causal Effects with Noncompliance

In most social experiments, some treatment group members ("no-shows") do not receive treatment and some control group members ("crossovers") do. This noncompliance dilutes the experimental treatment contrast, causing it to understate the average treatment effect. Consequently, it is important to distinguish between the following two impact questions:

1. What is the average effect of offering treatment?

2. What is the average effect of receiving treatment?

The first question asks about the impact of a treatment offer. This impact — which can be estimated experimentally — is often called the average effect of "intent to treat" (ITT). Since

group which differs from that for the control group (Bryk and Raudenbush, 1988). This is a particular instance of heteroscedasticity. Assuming that these standard deviations are the same can bias estimates of the standard error of the impact estimator (Gail et al., 1996). Two ways to eliminate this problem are to: (1) use a balanced sample allocation and (2) estimate separate variances for the treatment and control groups (Bloom, 2005b).

voluntary programs can only offer treatment — they cannot require it — the effect of intent to treat is a relevant consideration for making policy decisions about such programs. Furthermore, since mandatory programs often have incomplete compliance, the effect of ITT can be an important consideration for judging them.

The second question above asks about the impact of treatment receipt. It is often called the average impact of "treatment on the treated" (TOT) and is typically the question of interest for developers of interventions who want to know what they can achieve by full implementation of their ideas. However, in many instances this impact question may not be as policy-relevant as the first one, because rarely can treatment receipt be mandated.

There is no valid way to estimate the second type of effect experimentally, because there is no way to know which control group members are counterparts to treatment group members who receive treatment. To estimate such impacts, Bloom (1984) developed an extension of the experimental method, which was later expanded by Angrist, Imbens, and Rubin (1996).[12] To see how this approach works, it is useful to adopt a framework and notation that is now conventional for presenting it. This framework comprises three variables: Y, the outcome measure; Z, which equals one for subjects *randomized* to treatment and zero otherwise; and D, which equals one for subjects who *receive* the treatment and zero otherwise.

Consider an experiment in which some treatment group members do not receive treatment (they become no-shows) but no control group members receive treatment (there are no crossovers). If *no-shows experience no effect* from the intervention (because they are not exposed to it) or from randomization per se, the average effect of intent to treat equals the weighted mean of TOT for treatment recipients and zero for no-shows, with weights equal to the treatment receipt rate ([E(D|Z = 1]) and the no-show rate (1 – [E(D|Z = 1)]), such that:

$$ITT = [E(D|Z = 1)]TOT + [1 - E(D|Z = 1)]0 = [E(D|Z = 1)]TOT \quad (6)$$

Equation 6 implies that:

$$TOT = \frac{ITT}{E(D \mid Z = 1)} \quad (7)$$

The effect of treatment on the treated thus equals the effect of intent to treat divided by the expected receipt rate for treatment group members. For example, if the effect of intent to treat for a job training program were a $1,000 increase in annual earnings, and half of the treatment group members received treatment, then the effect of treatment on the treated would be $1,000/0.5 or $2,000. This adjustment allocates all of the treatment effect to only those treat-

[12] Angrist (2005) and Gennetian et al. (2005) illustrate the approach.

ment group members who receive treatment. Equation 7 represents the true effect of treatment on the treated for a given population. The corresponding sample estimator, $\hat{TOT}$, is:

$$\hat{TOT} = \frac{\bar{Y}_T - \bar{Y}_C}{(\bar{D} \mid Z = 1)} \tag{8}$$

where $(\bar{D} \mid Z = 1)$ equals the observed treatment receipt rate for the treatment group. If no-shows experience no effect, this estimator is statistically consistent and its estimated standard error is approximately:

$$se(\hat{TOT}) \approx \frac{se(\bar{Y}_T - \bar{Y}_C)}{(\bar{D} \mid Z = 1)} \tag{9}$$

Hence, both the point estimate and standard error are scaled by the treatment receipt rate.

The preceding approach does not require that no-shows be similar to treatment recipients. It requires only that no-shows experience no effect from treatment or randomization.[13] In addition, because of potential heterogeneity of treatment effects, the effect of treatment on the treated generalizes only to experimental treatment recipients and does not necessarily equal the average treatment effect.

Now add crossovers (control group members who receive treatment) to the situation, which further dilutes the experimental treatment contrast. Nonetheless, the difference in mean outcomes for the treatment group and control group provides an unbiased estimate of the effect of intent to treat. Thus, it addresses the first impact question stated above. To address the second question requires a more complex analytic framework with additional assumptions. This framework — developed by Angrist, Imbens, and Rubin (1996) — is based on four conceptual subgroups, which because of randomization comprise the same proportion of the treatment group and control group, in expectation. Figure 2 illustrates the framework and how it relates to the concepts of no-shows and crossovers. The first stacked bar in the figure represents all treatment group members (for whom Z = 1) and the second stacked bar represents all control group members (for whom Z = 0). Treatment group members who do not receive treatment (for whom D = 0) are no-shows, and control group members who do receive treatment (for whom D = 1) are crossovers.

Randomization induces treatment receipt for two of the four subgroups in the Angrist, Imbens, and Rubin (1996) framework — "compliers" and "defiers." Compliers receive treatment only if they are randomized to the treatment group, and defiers receive treatment only if

[13] This is a specific case of the exclusion principle specified by Angrist, Imbens, and Rubin (1996).

they are randomized to the control group. Thus, compliers add to the effect of ITT, and defiers subtract from it. Randomization does not influence treatment receipt for the other two groups — "always-takers," who receive treatment regardless of their randomization status, and "never-takers," who do not receive treatment regardless of their randomization status. Never-takers experience no treatment effect in the treatment group or control group, and always-takers experience the same effect in both groups, which cancels out in the overall difference between treatment and control groups. Hence, always-takers and never-takers do not contribute information about treatment effects.

If defiers do not exist,[14] which is reasonable to assume in many situations, the effect of treatment for compliers, termed by Angrist, Imbens, and Rubin (1996) the Local Average Treatment Effect (LATE) is:[15]

$$LATE = \frac{ITT}{E(D \mid Z = 1) - E(D \mid Z = 0)} \tag{10}$$

Thus to estimate the local average treatment effect from an experiment, one simply divides the difference in mean outcomes for the treatment and control groups by their difference in treatment receipt rates, or:

$$\hat{LATE} = \frac{\bar{Y}_T - \bar{Y}_C}{(\bar{D} \mid Z = 1) - (\bar{D} \mid Z = 0)} \tag{11}$$

The estimated Local Average Treatment Effect is the ratio of the estimated impact of randomization on outcomes and the estimated impact of randomization on treatment receipt.[16] Angrist, Imbens, and Rubin show that this ratio is a simple form of instrumental variables analysis called a Wald estimator (Wald, 1940).

Returning to our previous example, assume that there is a $1,000 difference in mean annual earnings for a treatment group and control group; half of the treatment group receives treatment and one-tenth of the control group receives treatment. The estimated local average

[14] Angrist, Imbens, and Rubin (1996) refer to this condition as monotonicity.

[15] This formulation assumes that the average effect of treatment on always-takers is the same whether they are randomized to treatment or control status.

[16] The expression for LATE in Equation 10 simplifies to the expression for TOT in Equation 7 when there are no-shows but no crossovers. Both expressions represent ITT divided by the probability of being a complier. When there are crossovers, the probability of being a complier equals the probability of receiving the treatment if randomized to the treatment group, minus the probability of being an always-taker. When there are no crossovers, there are no always-takers.

treatment effect equals the estimated impact on the outcome ($1,000), divided by the estimated impact on treatment receipt rates (0.5 – 0.1). This ratio equals $1,000/0.4 or $2,500.[17]

When using this approach to estimate treatment effects, it is important to clearly specify the groups to which it applies, because different groups may experience different effects from the same treatment, and not all groups and treatment effects can be observed without making further assumptions. The impact of intent to treat (ITT) applies to the full treatment group. So both the target group and its treatment effect can be observed. The Local Average Treatment Effect (LATE), which can be observed, applies to compliers, who cannot be observed. The effect of treatment on the treated (TOT), which cannot be observed, applies to all treatment group members who receive treatment (compliers plus always-takers), who can be observed.

Using Covariates and Blocking to Improve Precision

The two main approaches for improving the precision of randomized experiments — covariates and blocking — use the predictive power of past information about sample members to reduce unexplained variation in their future outcomes. This reduces the standard error of the impact estimator and its corresponding minimum detectable effect.[18] To examine these approaches, it is useful to reformulate the impact of intent to treat as the following bivariate regression:

$$Y_i = \alpha + \beta_0 T_i + \varepsilon_i \qquad (12)$$

where:

Y_i = the outcome for sample member i

T_i = one if sample member i is randomized to the treatment group and zero otherwise

ε_i = a random error that is independently and identically distributed across sample members within experimental groups, with a mean of 0 and a variance of σ^2.

α is the expected mean outcome without treatment and β_0 is the average effect of intent to treat. Thus β_0 equals the difference in expected outcomes for the treatment group and control group, and its estimator, $\hat{\beta}_0$ equals the difference in mean outcomes for the treatment and control groups in the experimental sample.

[17] In the present analysis, treatment receipt is a mediating variable in the causal path between randomization and the outcome. Gennetian et al. (2005) show how the same approach (using instrumental variables with experiments) can be used to study causal effects of other mediating variables.

[18] The remainder of this chapter assumes a common variance for treatment and control groups.

Using k* baseline characteristics, X_{ki}, as covariates to reduce the unexplained variation in Y_i produces the following multiple regression model for estimating intervention effects:

$$Y_i = \alpha + \beta_0 T_i + \sum_{k=1}^{k^*} B_k X_{ki} + \varepsilon_i^* \tag{13}$$

Defining R_A^2 as the proportion of pooled unexplained variation in the outcome within experimental groups predicted by covariates, the minimum detectable effect size is:[19]

$$MDES(\hat{\beta}_0) = M_{n-k^*-2}\sqrt{\frac{1-R_A^2}{nP(1-P)}} \tag{14}$$

There are two differences between the minimum detectable effect size in Equation 14 with covariates and Equation 5 without covariates. The first difference involves the multipliers, M_{n-2} and M_{n-k^*-2}, where the latter multiplier accounts for the loss of k* degrees of freedom from estimating coefficients for k*covariates. With roughly 40 or more sample members and 10 or fewer covariates, this difference is negligible, however.[20]

The second difference is the term $1 - R^2_A$ with covariates in Equation 14, instead of the value 1 in Equation 5 without covariates. The term $1 - R^2_A$ implies that the minimum detectable effect size decreases as the predictive power of covariates increases for a given sample size and allocation. In this way, covariates can increase effective sample size. For example, an R_A^2 of 0.25 yields an effective sample that is one-third larger than that without covariates; an R_A^2 of 0.50 yields an effective sample that is twice as large; and an R_A^2 of 0.75 yields an effective sample that is four times as large.

Several points are important to note about using covariates with experiments. First, they are not needed to eliminate bias, because randomization has done so already. Thus, values for the term B_0 in Equations 12 and 13 are identical. Second, it is good practice to specify all covariates in advance of the impact analysis — preferably when an experiment is being designed. This helps to avoid subsequent data mining. Third, the best predictors of future outcomes are typically past outcomes. For example, past student achievement is usually the best predictor of future student achievement. This is because past outcomes reflect most factors that determine future outcomes. Fourth, some outcomes are more predictable than others, and thus covariates provide greater precision gains for them. For example, the correlation between individual stan-

[19] One way to estimate R^2_A from a dataset would be to first estimate Equation 12 and compute residual outcome values for each sample member. The next step would be to regress the residuals on the covariates. The resulting r-square for the second regression is an estimate of R^2_A.

[20] See Bloom (2005b) for a discussion of this issue.

dardized test scores is typically stronger for high school students than for elementary school students (Bloom, Richburg-Hayes, and Rebeck Black, 2005).

The second approach to improving precision is to block or stratify experimental sample members by some combination of their baseline characteristics, and then randomize within each block or stratum. The extreme case of two sample members per block is an example of matching. Factors used for blocking in social research typically include geographic location, organizational units, demographic characteristics, and past outcomes. To compute an unbiased estimate of the impact of intent to treat from such designs requires computing impact estimates for each block and pooling estimates across blocks. One way to do this in a single step is to add to the impact regression a series of indicator variables that represent each of the m* blocks, and suppress the intercept, α , yielding:

$$Y_i = \beta_0 T_i + \sum_{m=1}^{m^*} \gamma_m S_{mi} + \varepsilon_i \tag{15}$$

where:

S_{mi} = one if sample member i is from block (or stratum) m and zero otherwise.

The estimated value of B_0 provides an unbiased estimator of the effect of intent to treat. The minimum detectable effect size of this estimator can be expressed as:

$$MDES(\hat{\beta}_0) = M_{n-m^*-1}\sqrt{\frac{1-R_B^{\,2}}{nP(1-P)}} \tag{16}$$

where:

R_B^2 = the proportion of unexplained variation in the outcome within experimental groups (pooled) predicted by the blocks.

There are two differences in the expressions for minimum detectable effects with and without blocking (Equations 16 and 5). The first difference involves the multipliers, M_{n-m^*-1} versus M_{n-2}, which account for the loss of one degree of freedom per block and the gain of one degree of freedom from suppressing the intercept. With samples of more than about 40 members in total and 10 or fewer blocks, there is very little difference between these two multipliers. The second difference is the addition of the term $1-R_B^2$ in Equation 16 to account for the predictive power of blocking. The more similar sample members are within blocks and the more different blocks are from each other, the higher this predictive power is. This is where precision gains come from. Note, however, that for samples with fewer than about 10 subjects, precision losses due to reducing the number of degrees of freedom by blocking can sometimes outweigh precision gains due to the predictive power of blocking. This is most likely to occur in experiments that randomize small numbers of groups (discussed later).

Another reason to block sample members is to avoid an "unhappy" randomization with embarrassing treatment and control group differences on a salient characteristic. Such differences can reduce the face validity of an experiment, thereby undermining its credibility. Blocking first on the salient characteristic eliminates such a mismatch.

Sometimes researchers wish to assure treatment and control group matches on multiple characteristics. One way to do so is to define blocks in terms of combinations of characteristics (e.g., age, race, and gender). But doing so can become complicated in practice due to uneven distributions of sample members across blocks, and the consequent need to combine blocks, often in ad hoc ways. A second approach is to specify a composite index of baseline characteristics and create blocks based on intervals of this index.[21] Using either approach, the quality of the match on any given characteristic typically declines as the number of matching variables increases. So it is important to set priorities for which variables to match on.[22]

Regardless of how blocks are defined, one's impact analysis must account for them if they are used. To not do so would bias estimates of standard errors. In addition, it is possible to use blocking in combination with covariates. If so, both features of the experimental design should be represented in the experimental analysis.

Randomizing Groups to Estimate Intervention Effects

This section introduces a type of experimental design that is growing rapidly in popularity — the randomization of intact groups or clusters.[23] Randomizing groups makes it possible to measure the effectiveness of interventions that are designed to affect entire groups or are delivered in group settings, such as communities, schools, hospitals, or firms. For example, schools have been randomized to measure the impacts of whole school reforms (Borman et al., 2005, and Cook, Hunt, and Murphy, 2000) and school-based risk-prevention campaigns (Flay, 2000); communities have been randomized to measure the impacts of community health campaigns (Murray et al., 1994); small local areas have been randomized to study the impacts of police patrol interventions (Sherman and Weisburd, 1995); villages have been randomized to study the effects of a health, nutrition, and education initiative (Teruel and Davis, 2000); and public housing developments have been randomized to study the effects of a place-based HIV prevention program (Sikkema et al., 2000) and a place-based employment program (Bloom and Riccio, 2005).

[21] Such indices include propensity scores (Rosenbaum and Rubin, 1983) and Mahalanobis distance functions (http//en.wikipedia.org/wiki/Mahalanobis_distance).

[22] One controversial issue is whether to treat blocks as "fixed effects," which represent a defined population, or "random effects," which represent a random sample from a larger population. Equations 15 and 16 treat blocks as fixed effects. Raudenbush, Martinez, and Spybrook (2005) present random-effects estimators for blocking.

[23] Bloom (2005b), Donner and Klar (2000), and Murray (1998) provide detailed discussions of this approach, and Boruch and Foley (2000) review its applications.

Group randomization provides unbiased estimates of intervention effects for the same reasons that individual randomization does. However, the statistical power or precision of group randomization is less than that for individual randomization, often by a lot. To see this, consider the basic regression model for estimating intent-to-treat effects with group randomization:

$$Y_{ij} = \alpha + \beta_0 T_j + e_j + \varepsilon_{ij} \tag{17}$$

where:

Y_{ij} = the outcome for individual i from group j

α = the mean outcome without treatment

B_0 = the average impact of intent to treat

T_j = 1 for groups randomized to treatment and 0 otherwise

e_j = an error that is independently and identically distributed between groups with a mean of 0 and a variance of τ^2

ε_{ij} = an error that is independently and identically distributed between individuals within groups with a pooled mean of zero and variance of σ^2.

Equation 17 for group randomization has an additional random error, e_j, relative to Equation 12 for individual randomization. This error reflects how mean outcomes vary across groups, which reduces the precision of group randomization.

To see this, first note that the relationship between group-level variance, τ^2, and individual-level variance, σ^2, can be expressed as an intra-class coefficient, ρ, where:

$$\rho = \frac{\tau^2}{\tau^2 + \sigma^2} \tag{18}$$

ρ equals the proportion of total variation across all individuals in the target population ($\tau^2+\sigma^2$) that is due to variation between groups (τ^2). If there is no variation between groups, ($\tau^2 = 0$) ρ equals zero. If there is no variation within groups, ($\sigma^2 = 0$) ρ equals one.

Consider a study that randomizes a total of J groups in proportion P to treatment with a harmonic mean value of n individuals per group. The ratio of the standard error of this impact estimator to that for individual randomization of the same total number of subjects (Jn) is referred to as a design effect (DE), where:

$$DE = \sqrt{1 + (n-1)\rho} \tag{19}$$

As the intra-class correlation (ρ) increases, the design effect increases, implying a larger standard error for group randomization relative to individual randomization. This is because a larger ρ implies greater random variation across groups. The value of ρ varies typically from about 0.01 to 0.20, depending on the nature of the outcome being measured and the type of group being randomized.

For a given total number of individuals, the design effect also increases as the number of individuals per group (n) increases. This is because for a given total number of individuals, larger groups imply fewer groups randomized. With fewer groups randomized, larger treatment and control group differences are likely for a given sample.[24]

The design effect has important implications for designing group-randomized studies. For example, with ρ equal to 0.10 and n equal to 100, the standard error for group randomization is 3.3 times that for individual randomization. To achieve the same precision, group randomization would need almost 11 times as many sample members. Note that the design effect is independent of J and depends only on the values of n and ρ.

The different standard errors for group randomization and individual randomization also imply a need to account for group randomization during the experimental analysis. This can be done by using a multilevel model that specifies separate variance components for groups and individuals (for example, see Raudenbush and Bryck, 2002). In the preceding example, using an individual-level model, which ignores group-level variation, would estimate standard errors that are one-third as large as they should be. Thus, as Jerome Cornfield (1978, 101) aptly observed: "Randomization by group accompanied by an analysis appropriate to randomization by individual is an exercise in self-deception."

Choosing a sample size and allocation for group-randomized studies means choosing values for J, n, and P. Equation 20 illustrates how these choices influence minimum detectable effect size (Bloom, Richburg-Hayes, and Rebeck Black, 2005).

$$MDES\ (\hat{\beta}_0) = M_{J-2}\sqrt{\frac{\rho}{P(1-P)J} + \frac{1-\rho}{P(1-P)Jn}}\ , \qquad (20)$$

This equation indicates that the group-level variance (ρ) is divided by the total number of randomized groups, J, whereas the individual-level variance, (1- ρ) is divided by the total number of individuals, Jn.[25] Hence, increasing the number of randomized groups reduces both

[24]The statistical properties of group randomization in experimental research are much like those of cluster sampling in survey research (Kish, 1965).

[25]When total student variance ($\tau^2 + \sigma^2$) is standardized to a value of one by substituting the intra-class correlation (ρ) into the preceding expressions, ρ represents τ^2 and (1- ρ) represents σ^2.

variance components, whereas increasing the number of individuals per group reduces only one component. This result illustrates one of the most important design principles for group-randomized studies: *The number of groups randomized influences precision more than the size of the groups randomized*.

The top panel of Table 2 illustrates this point by presenting minimum detectable effect sizes for an intra-class correlation of 0.10, a balanced sample allocation, and no covariates. Reading across each row illustrates that, after group size reaches about 60 individuals, increasing it affects precision very little. For very small randomized groups (with less than about 10 individuals each), changing group size can have a more pronounced effect on precision.

Reading down any column in the top panel illustrates that increasing the number of groups randomized can improve precision appreciably. Minimum detectable effects are approximately inversely proportional to the square root of the number of groups randomized once the number of groups exceeds about 20.

Equation 29 illustrates how covariates affect precision with group randomization.[26]

$$MDES\ (\hat{\beta}_0) = M_{J-g^*-2}\sqrt{\frac{\rho(1-R_2{}^2)}{P(1-P)J} + \frac{(1-\rho)(1-R_1{}^2)}{P(1-P)Jn}} \qquad (21)$$

where

$R_1{}^2$ = the proportion of individual variance (at level one) predicted by covariates,

$R_2{}^2$ = the proportion of group variance (at level two) predicted by covariates,

g* = the number of group covariates used (n.b.: The number of individual covariates does not affect the number of degrees of freedom).

With group randomization, multiple levels of predictive power are at play — $R_1{}^2$ for level one (individuals) and $R_2{}^2$ for level two (groups).[27] Group-level covariates can reduce the unexplained group-level variance (τ^2), whereas individual-level covariates can reduce both the group-level and individual-level variances (τ^2 and σ^2). However, because group-level variance is typically the binding constraint on precision, its reduction is usually most important. This is analogous to the fact that increasing the number of groups is usually more important than increasing group size. Thus, in some cases group-level covariates — which can be simple and

[26]Raudenbush (1997) and Bloom, Richburg-Hayes, and Rebeck Black (2005) discuss in detail how covariates affect precision with group randomization.

[27]The basic principles discussed here extend to situations with more than two levels of clustering.

inexpensive to obtain — provide as much gain in precision as do individual covariates (Bloom, Richburg-Hayes, and Rebeck Black, 2005, and Bloom, Bos, and Lee, 1999.)

Because of group-randomization's large sample size requirements, it is especially important to use covariates to predict group-level variances. The bottom panel of Table 2 illustrates this point. It presents the minimum detectable effect size for each sample configuration in the top panel when a covariate that predicts 60 percent of the group-level variance ($R_2^2 = 0.6$) is included. For example, adding this covariate to a design that randomizes 30 groups with 60 individuals each reduces the minimum detectable effect size from 0.36 to 0.25, which is equivalent to doubling the number of groups randomized.

Widespread application of group randomization is only beginning, and much remains to be learned about how to use the approach effectively for social research. One of the most important pieces of information required to do so is a comprehensive inventory of parameter values needed to design such studies — ρ, $R_1^{2,}$ and R_2^2. These values vary widely, depending on the type of outcome being measured, the type of group being randomized, and the type of covariate/s being used.[28]

Future Frontiers

During the past several decades, randomized experiments have been used to address a rapidly expanding range of social science questions, experimental designs have become increasingly sophisticated, and statistical methods have become more advanced. So what are the frontiers for future advances?

One frontier involves expanding the geographic scope of randomized experiments in the social sciences. To date, the vast majority of such experiments have been conducted in the United States, although important exceptions exist in both developed and developing countries.[29] Given the promise of the approach, much more could be learned by promoting its use throughout the world.

A second frontier involves unpacking the "black box" of social experiments. Experiments are uniquely qualified to addresses questions like: What did an intervention cause to hap-

[28]Existing sources of this information include, among others: Bloom, Richburg-Hayes, and Rebeck Black (2005); Bloom, Bos, and Lee (1999); Hedges and Hedberg (2005); Murray and Blitstein (2003); Murray and Short (1995); Schochet (2005); Siddiqui, Hedeker, Flay, and Hu (1996); and Ukoumunne et al. (1999).

[29]Some other countries where randomized social experiments have been conducted include: the UK (Walker, Hoggart, Hamilton, and Blank, 2006); Mexico (Shultz, forthcoming); Colombia (Angrist et al., 2002); Israel (Angrist and Lavy, 2002); India (Banerjee, Cole, Duflo, and Linden, 2005, and Duflo and Hanna, 2005); and Kenya (Miguel and Kremer, 2004). For a review of randomized experiments in developing countries, see Kremer (2003).

pen? But they are not well suited to address questions like: Why did an intervention have or not have an effect?[30] Two promising approaches to such questions are emerging, which combine nonexperimental statistical methods with experimental designs.

One approach uses instrumental variables analysis to examine the causal paths between randomization and final outcomes by comparing intervention effects on intermediate outcomes (mediating variables) with those on final outcomes.[31] The other approach uses methods of research synthesis (meta-analysis or multilevel models that pool primary data) with multiple experiments, multiple experimental sites, or both to estimate how intervention effects vary with treatment implementation, sample characteristics, and local context.[32]

Perhaps the most important frontier for randomized experiments in the social sciences is the much-needed expansion of organizational and scientific capacity to implement them successfully on a much broader scale. To conduct this type of research well requires high levels of scientific and professional expertise, which at present exist only at a limited number of institutions. It is therefore hoped that this chapter will contribute to a broader application of this approach to social research.

[30]Two studies that tried to open the black box of treatment effects experimentally are the Riverside, California Welfare Caseload Study, which randomized different caseload sizes to welfare workers (Riccio, Friedlander, and Freedman, 1994) and the Columbus, Ohio, comparison of separate versus integrated job functions for welfare workers (Scrivener and Walter, 2001).

[31]For example, Morris and Gennetian (2003), Gibson, Magnusen, Gennetian, and Duncan (2005), Liebman, Katz, and Kling (2004), and Ludwig, Duncan, and Hirschfield (2001) used instrumental variables with experiments to measure the effects of mediating variables on final outcomes.

[32]Heinrich (2002) and Bloom, Hill, and Riccio (2003) used primary data from a series of experiments to address these issues.

Table 1

Minimum Detectable Effect and Effect Size for Individual Randomization

Sample Size n	Sample Allocation P/(1-P)				
	0.5/0.5	0.6/0.4 or 0.4/0.6	0.7/0.3 or 0.3/0.7	0.8/0.2 or 0.2/0.8	0.9/0/1 or 0.1/0.9
	Minimum Detectable Effect Given σ = $1,000				
50	$ 810	$ 830	$ 880	$ 1,010	$ 1,350
100	570	580	620	710	940
200	400	410	430	500	660
400	280	290	310	350	470
800	200	200	220	250	330
1600	140	140	150	180	230
	Minimum Detectable Effect Size				
50	0.81	0.83	0.88	1.01	1.35
100	0.57	0.58	0.62	0.71	0.94
200	0.40	0.41	0.43	0.50	0.66
400	0.28	0.29	0.31	0.35	0.47
800	0.20	0.20	0.22	0.25	0.33
1600	0.14	0.14	0.15	0.18	0.23

SOURCE: Computations by the author.

NOTE: Minimum detectable effect sizes are for a two-tail hypothesis test with statistical significance of 0.05 and statistical power of 0.80.

Table 2

Minimum Detectable Effect Size for Balanced Group Randomization With ρ = 0.10

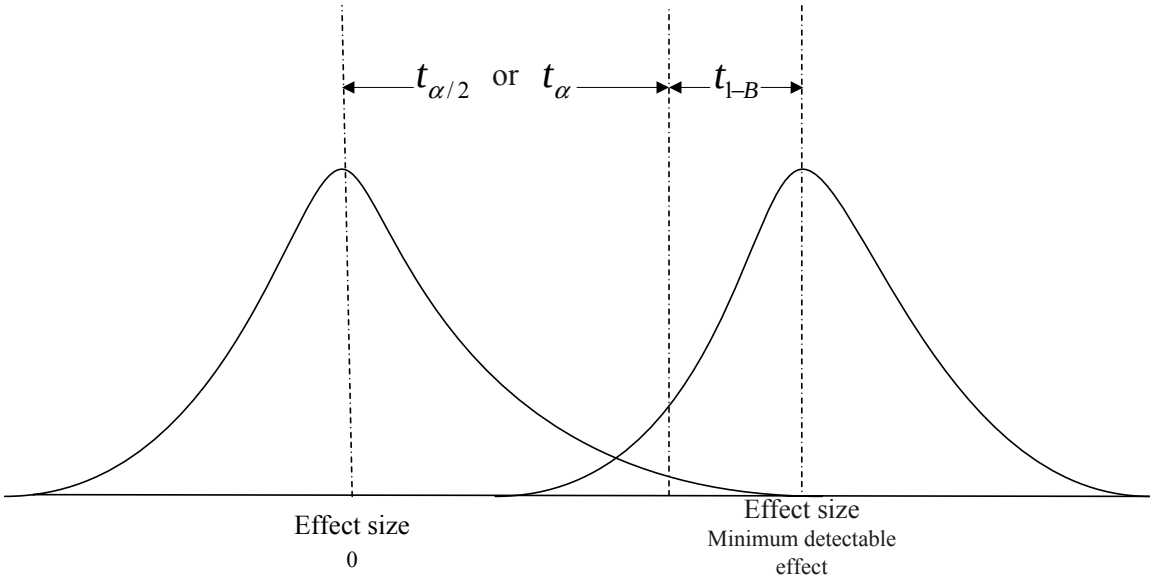
Total Groups Randomized (J)	Group Size (n)				
	10	**30**	**60**	**120**	**480**
	No Covariates				
10	0.88	0.73	0.69	0.66	0.65
30	0.46	0.38	0.36	0.35	0.34
60	0.32	0.27	0.25	0.24	0.24
120	0.23	0.19	0.18	0.17	0.17
480	0.11	0.09	0.09	0.08	0.08
	Group-Level Covariate (R_2^2 = 0.6)				
10	0.73	0.54	0.47	0.44	0.41
30	0.38	0.28	0.25	0.23	0.22
60	0.27	0.20	0.17	0.16	0.15
120	0.19	0.14	0.12	0.11	0.11
480	0.09	0.07	0.06	0.06	0.05

SOURCE: Computations by the author.

NOTE: Minimum detectable effect sizes are for two-tail hypothesis tests with statistical significance of 0.05 and statistical power of 0.80.

Figure 1

The Minimum Detectable Effect Multiplier

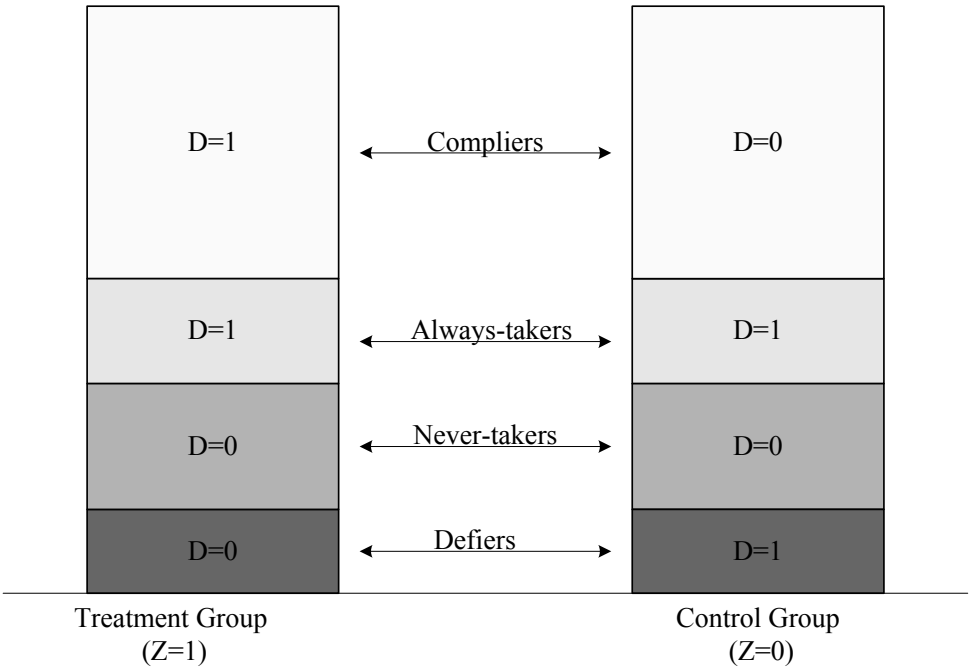


One-tail multiplier = $t_{\alpha} + t_{1-B}$

Two-tail multiplier = $t_{\alpha/2} + t_{1-B}$

Figure 2

A Hypothetical Experiment Including No-Shows and Crossovers



SOURCE: Howard S. Bloom (ed.). 2005. *Learning More from Social Experiments: Evolving Analytic Approaches.* New York: Russell Sage Foundation.

NOTE: D equals 1 if the treatment would be received and 0 otherwise.

References

Aigner, Dennis J. 1985. "The Residential Time-of-Use Pricing Experiments: What Have We Learned?" In Jerry A. Hausman and David A. Wise (eds.), *Social Experimentation.* Chicago: University of Chicago Press.

Angrist, Joshua D. 2005. "Instrumental Variables Methods in Experimental Criminology Research: What, Why and How." *Journal of Experimental Criminology* 2: 1-22.

Angrist, Joshua, Eric Bettinger, Erik Bloom, Elizabeth King, and Michael Kremer. 2002. "Vouchers for Private Schooling in Colombia: Evidence from a Randomized Natural Experiment." *The American Economic Review* 92(5): 1535-58.

Angrist, Joshua, Guido Imbens, and Don Rubin. 1996. "Identification of Causal Effects Using Instrumental Variables." JASA Applications invited paper, with comments and authors' response. *Journal of the American Statistical Association* 91(434): 444-55.

Angrist, Joshua D., and Victor Lavy. 2002. "The Effect of High School Matriculation Awards: Evidence from Randomized Trials." Working Paper 9389. New York: National Bureau of Economic Research.

Banerjee, Abhijit, Shawn Cole, Esther Duflo, and Leigh Linden. 2005. "Remedying Education: Evidence from Two Randomized Experiments in India." Working Paper 11904. Cambridge, MA: National Bureau of Economic Research.

Bell, Stephen, Michael Puma, Gary Shapiro, Ronna Cook, and Michael Lopez. 2003. "Random Assignment for Impact Analysis in a Statistically Representative Set of Sites: Issues from the National Head Start Impact Study." *Proceedings of the August 2003 American Statistical Association Joint Statistical Meetings* (CD-ROM). Alexandria, VA: American Statistical Association.

Bloom, Dan, and Charles Michalopoulos. 2001. *How Welfare and Work Policies Affect Employment and Income: A Synthesis of Research*. New York: MDRC.

Bloom, Howard S. (ed.). 2005a. *Learning More from Social Experiments: Evolving Analytic Approaches.* New York: Russell Sage Foundation.

Bloom, Howard S. 2005b. "Randomizing Groups to Evaluate Place-Based Programs." In Howard S. Bloom (ed.), *Learning More from Social Experiments: Evolving Analytic Approaches*. New York: Russell Sage Foundation.

Bloom, Howard S. 1995. "Minimum Detectable Effects: A Simple Way to Report the Statistical Power of Experimental Designs." *Evaluation Review* 19(5): 547-56.

Bloom, Howard S. 1984. "Accounting for No-Shows in Experimental Evaluation Designs." *Evaluation Review* 8(2): 225-46.

Bloom, Howard S., Johannes M. Bos, and Suk-Won Lee. 1999. "Using Cluster Random Assignment to Measure Program Impacts: Statistical Implications for the Evaluation of Education Programs." *Evaluation Review* 23(4): 445-69.

Bloom, Howard S., Carolyn J. Hill, and James A. Riccio. 2003. "Linking Program Implementation and Effectiveness: Lessons from a Pooled Sample of Welfare-to-Work Experiments." *Journal of Policy Analysis and Management* 22(4): 551-75.

Bloom, Howard S., and James A. Riccio. 2005. "Using Place-Based Random Assignment and Comparative Interrupted Time-Series Analysis to Evaluate the Jobs-Plus Employment Program for Public Housing Residents." *Annals of the American Academy of Political and Social Science* 599 (May): 19-51.

Bloom, Howard S., Lashawn Richburg-Hayes, and Alison Rebeck Black. 2005. "Using Covariates to Improve Precision: Empirical Guidance for Studies that Randomize Schools to Measure the Impacts of Educational Interventions." Working Paper. New York: MDRC.

Bloom, Howard S., JoAnn Rock, Sandra Ham, Laura Melton, and Julieanne O'Brien. 2001. *Evaluating the Accelerated Schools Approach: A Look at Early Implementation and Impacts on Student Achievement in Eight Elementary Schools.* New York: MDRC.

Bloom, Howard S., Larry L. Orr, George Cave, Stephen H. Bell, Fred Doolittle, and Winston Lin. 1997. "The Benefits and Costs of JTPA Programs: Key Findings from the National JTPA Study." *The Journal of Human Resources*, vol. 32, no. 3, Summer.

Borman, Geoffrey D., Robert E. Slavin, A. Cheung, Anne Chamberlain, Nancy Madden, and Bette Chambers. 2005. "The National Randomized Field Trial of Success for All: Second-Year Outcomes." *American Educational Research Journal* 42: 673-96.

Boruch, Robert F. 1997. *Randomized Experiments for Planning and Evaluation.* Thousand Oaks, CA: Sage Publications.

Boruch, Robert F., and Ellen Foley. 2000. "The Honestly Experimental Society: Sites and Other Entities as the Units of Allocation and Analysis in Randomized Trials." In Leonard Bickman (ed.), *Validity and Social Experimentation: Donald Campbell's Legacy*, vol. 1. Thousand Oaks, CA: Sage Publications.

Box, George E.P., J. Stuart Hunter, and William G. Hunter. 2005. 2nd ed. *Statistics for Experimenters: Design Innovation and Discovery.* New York: John Wiley and Sons.

Bryk, Anthony S., and Stephen W. Raudenbush. 1988. "Heterogeneity of Variance in Experimental Studies: A Challenge to Conventional Interpretations." *Psychological Bulletin* 104(3): 396-404.

Cochran, William G., and Gertrude M. Cox, 1957. *Experimental Designs*. New York: John Wiley and Sons.

Cochrane Collaboration. 2002. "Cochrane Central Register of Controlled Trials Database." Available at the Cochrane Library Web site: www.cochrane.org (accessed September 14, 2004).

Cohen, Jacob. 1977/1988. *Statistical Power Analysis for the Behavioral Sciences*. New York: Academic Press.

Cook, Thomas H., David Hunt, and Robert F. Murphy. 2000. "Comer's School Development Program in Chicago: A Theory-Based Evaluation." *American Educational Research Journal* (Summer).

Cornfield, Jerome. 1978. "Randomization by Group: A Formal Analysis." *American Journal of Epidemiology* 108(2): 100-02.

Cox, D. R. 1958. *Planning of Experiments*. New York: John Wiley and Sons.

Donner, Allan, and Neil Klar. 2000. *Design and Analysis of Cluster Randomization Trials in Health Research*. London: Arnold.

Duflo, Esther, and Rema Hanna. 2005. "Monitoring Works: Getting Teachers to Come to School." Working Paper 11880. Cambridge, MA: National Bureau of Economic Research.

Fisher, Ronald A. 1925. *Statistical Methods for Research Workers*. Edinburgh: Oliver and Boyd.

Fisher, Ronald A. 1935. *The Design of Experiments*. Edinburgh: Oliver and Boyd.

Flay, Brian R. 2000. "Approaches to Substance Use Prevention Utilizing School Curriculum Plus Social Environment Change." *Addictive Behaviors* 25(6): 861-85.

Gail, Mitchell H., Steven D. Mark, Raymond J. Carroll, Sylvan B. Green, and David Pee. 1996. "On Design Considerations and Randomization-Based Inference for Community Intervention Trials." *Statistics in Medicine* 15: 1069-92.

Gennetian, Lisa A., Pamela A. Morris, Johannes M. Bos, and Howard S. Bloom. 2005. "Constructing Instrumental Variables from Experimental Data to Explore How Treatments Produce Effects." In Howard S. Bloom (ed.), *Learning More from Social Experiments: Evolving Analytic Approaches.* New York: Russell Sage Foundation.

Gibson, C., Katherine Magnusen, Lisa Gennetian, and Greg Duncan. 2005. "Employment and Risk of Domestic Abuse among Low-Income Single Mothers. *Journal of Marriage and the Family* 67: 1149-68.

Greenberg, David H., and Mark Shroder. 1997. *The Digest of Social Experiments*. Washington, DC: Urban Institute Press.

Hedges, Larry V., and Eric C. Hedberg. 2005. "Intraclass Correlation Values for Planning Group Randomized Trials in Education." Working Paper WP-06-12. Evanston, IL: Northwestern University, Institute for Policy Research.

Heinrich, Carolyn J. 2002. "Outcomes-Based Performance Management in the Public Sector: Implications for Government Accountability and Effectiveness." *Public Administration Review* 62(6): 712-25.

Kane, Thomas. 2004. "The Impact of After-School Programs: Interpreting the Results of Four Recent Evaluations." Working Paper. New York: W.T. Grant Foundation.

Kemple, James J., and Jason Snipes. 2000. *Career Academies: Impacts on Students' Engagement and Performance in High School*. New York: MDRC.

Kempthorne, Oscar. 1952. *The Design and Analysis of Experiments*. Malabar, FL: Robert E. Krieger Publishing Company.

Kish, Leslie. 1965. *Survey Sampling*. New York: John Wiley.

Kling, Jeffrey R., Jeffrey B. Liebman, and Lawrence F. Katz. Forthcoming. "Experimental Analysis of Neighborhood Effects." *Econometrica*.

Kremer, Michael. 2003. "Randomized Evaluations of Educational Programs in Developing Countries: Some Lessons." *American Economic Review* 93(2): 102-06.

Liebman, Jeffrey B., Lawrence F. Katz, and Jeffrey R. Kling. 2004. "Beyond Treatment Effects: Estimating the Relationship Between Neighborhood Poverty and Individual Outcomes in the MTO Experiment." IRS Working Paper 493 (August). Princeton, NJ: Princeton University, Industrial Relations Section.

Lindquist, E.F. 1953. *Design and Analysis of Experiments in Psychology and Education*. Boston: Houghton Mifflin Company.

Lipsey, Mark W. 1990. *Design Sensitivity: Statistical Power for Experimental Research.* Newbury Park, CA: Sage.

Lipsey, Mark W. 1988. "Juvenile Delinquency Intervention." In Howard S. Bloom, David S. Cordray, and Richard J. Light (eds.), *Lesson from Selected Program and Policy Areas.* San Francisco: Jossey-Bass.

Ludwig, Jens, Greg J. Duncan, and Paul Hirschfield. 2001. "Urban Poverty and Juvenile Crime: Evidence from a Randomized Housing-Mobility Experiment." *The Quarterly Journal of Economics* 116(2): 655-80.

Marks, Harry M. 1997. *The Progress of Experiment: Science and Therapeutic Reform in the United States, 1900-1990*. Cambridge: Cambridge University Press.

McCall, W.A. 1923. *How to Experiment in Education*. New York: MacMillan.

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

Morris, Pamela, and Lisa Gennetian. 2003. "Identifying the Effects of Income on Children's Development: Using Experimental Data." *Journal of Marriage and the Family* 65(3): 716-29.

Munnell, Alicia (ed.). 1987. *Lessons from the Income Maintenance Experiments*. Boston: Federal Reserve Bank of Boston.

Murray, David M. 1998. *Design and Analysis of Group-Randomized Trials*. New York: Oxford University Press.

Murray, David M., and Jonathan L. Blitstein. 2003. "Methods to Reduce the Impact of Intraclass Correlation in Group-Randomized Trials." *Evaluation Review* 27(1): 79-103.

Murray, David M., Peter J. Hannan, David R. Jacobs, Paul J. McGovern, Linda Schmid, William L. Baker, and Clifton Gray. 1994. "Assessing Intervention Efforts in the Minnesota Heart Health Program." *American Journal of Epidemiology.* 139(1): 91-103.

Murray, David M., and Brian Short. 1995. "Intraclass Correlation Among Measures Related to Alcohol Use by Young Adults: Estimates, Correlates and Applications in Intervention Studies." *Journal of Studies on Alcohol* 56(6): 681-94.

Myers, Jerome L. 1972. *Fundamentals of Experimental Design*. Boston: Allyn and Bacon.

Newhouse, Joseph P. 1996. *Free for All? Lessons from the RAND Health Insurance Experiment*. Cambridge, MA: Harvard University Press.

Nye, Barbara, Larry V. Hedges, and Spyros Konstantopoulos. 1999. "The Long-Term Effects of Small Classes: A Five-Year Follow-Up of the Tennessee Class Size Experiment." *Education Evaluation and Policy Analysis* 21(2): 127-42.

Olds, David L., John Eckenrode, Charles R. Henderson, Jr., Harriet Kitzman, Jane Powers, Robert Cole, Kimberly Sidora, Pamela Morris, Lisa M. Pettitt, and Dennis Luckey. 1997. "Long-Term Effects of Home Visitation on Maternal Life Course and Child Abuse and Neglect." *The Journal of the American Medical Association* 278 (8): 637-43.

Orr, Larry L. 1999. *Social Experiments: Evaluating Public Programs with Experimental Methods.* Thousand Oaks, CA: Sage Publications.

Orr, Larry L., Judith D. Feins, Robin Jacob, Erik Beecroft, Lisa Sanbomatsu, Lawrence F. Katz, Jeffrey B. Liebman, and Jeffrey R. Kling. 2003. *Moving to Opportunity: Interim Impacts Evaluation.* Washington, DC: U.S. Department of Housing and Urban Development.

Peirce, Charles S., and Joseph Jastrow. 1884/1980. "On Small Differences of Sensation." Reprinted in Stephen M. Stigler (ed.), *American Contributions to Mathematical Statistics in the Nineteenth Century*, vol. 2. New York: Arno Press.

Puma, Michael, Stephen Bell, Ronna Cook, Camilla Heid, and Michael Lopez. 2006. *Head Start Impact Study: First Year Impact Findings.* (Prepared by Westat, Chesapeake Research Associates, The Urban Institute, American Institutes for Research, and Decision Information Resources, June.) Washington, DC: U. S. Department of Health and Human Services, Administration for Children and Families, Office of Planning, Research, and Evaluation.

Raudenbush, Stephen, W. 1997. "Statistical Analysis and Optimal Design for Group Randomized Trials." *Psychological Methods* 2(2): 173-85.

Raudenbush, Stephen W., and Anthony S. Bryk. 2002. *Hierarchical Linear Models: Applications and Data Analysis Methods,* 2nd ed. (Thousand Oaks, CA: Sage Publications).

Raudenbush, Stephen W., Andres Martinez, and Jessaca Spybrook. 2005. "Strategies for Improving Precision in Group-Randomized Experiments." New York: William T. Grant Foundation.

Riccio, James, Daniel Friedlander, and Stephen Freedman. 1994. *Benefits, Costs, and Three-Year Impacts of a Welfare-to-Work Program.* New York: MDRC.

Robins, Philip K., and Robert G. Spiegelman (eds.). 2001. *Reemployment Bonuses in the Unemployment Insurance System: Evidence from Three Field Experiments.* Kalamazoo, MI: W.E. Upjohn Institute for Employment Research.

Rosenbaum, Paul R., and Donald B. Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70(1): 41-55.

Schochet, Peter A. 2005. Statistical Power for Random Assignment Evaluations of Education Programs. Princeton, NJ: Mathematica Policy Research.

Schochet, Peter A. 2006. *National Job Corps Study and Longer-Term Follow-Up Study: Impact and Benefit-Cost Findings Using Survey and Summary Earnings Records Data.* Princeton, NJ: Mathematica Policy Research.

Scrivener, Susan, and Johanna Walter, with Thomas Brock and Gayle Hamilton. 2001. *National Evaluation of Welfare-to-Work Strategies: Evaluating Two Approaches to Case Management: Implementation, Participation Patterns, Costs, and Three-Year Impacts of the Columbus Welfare-to-Work Program.* Washington, DC: U.S. Department of Health and Human Services, Administration for Children and Families, and Office of the Assistant Secretary for Planning and Evaluation; and U.S. Department of Education, Office of the Under Secretary and Office of Vocational and Adult Education.

Sherman, Lawrence W., and David Weisburd. 1995. "General Deterrent Effects of Police Patrol in Crime 'Hot Spots': A Randomized Control Trial." *Justice Quarterly* 12(4): 625-48.

Shultz, Paul T. Forthcoming. "School Subsidies for the Poor: Evaluating the Mexican Progresa Poverty Program." *Journal of Development Economics*.

Siddiqui, Ohidul, Donald Hedeker, Brian R. Flay, and Frank B. Hu. 1996. "Intraclass Correlation Estimates in a School-Based Smoking Prevention Study: Outcome and Mediating Variables, by Sex and Ethnicity." *American Journal of Epidemiology* 144(4): 425-33.

Sikkema, Kathleen J., Jeffrey A. Kelly, Richard A. Winett, Laura J. Solomon, V.A. Cargill, R.A. Roffman, T.L. McAuliffe, T.G. Heckman, E.A. Anderson, D.A. Wagstaff, A.D. Norman, M.J. Perry, D.S. Crumble, and M.B. Mercer. 2000. "Outcomes of a Randomized Community-Level HIV Prevention Intervention for Women Living in 18 Low-Income Housing Developments." *American Journal of Public Health* 90(1): 57-63.

Teruel, Graciela M., and Benjamin Davis. 2000. *Final Report: An Evaluation of the Impact of PROGRESA Cash Payments on Private Inter-Household Transfers*. Washington, DC: International Food Policy Research Institute.

Ukoumunne, O.C., M.C. Gulliford, S. Chinn, J.A.C. Sterne and P.F.J. Burney. 1999. "Methods for Evaluating Area-Wide and Organisation-Based Interventions in Health and Health Care: A Systematic Review. *Health Technology Assessment* 3(5): 1-99.

Van Helmont, John Baptista. 1662. *Oriatrik or, Physick Refined: The Common Errors Therein Refuted and the Whole Art Reformed and Rectified.* London: Lodowick-Lloyd. Available at the

James Lind Library Web site: www.jameslindlibrary.org/trial_records/ 17th_18th_Century/van_helmont/van_helmont_kp.html (accessed January 3, 2005).

Wald, Abraham. 1940. "The Fitting of Straight Lines If Both Variables Are Subject to Error." *Annals of Mathematical Statistics* 11(September): 284-300.

Walker, Robert, Lesley Hoggart, Gayle Hamilton, and Susan Blank. 2006. *Making Random Assignment Happen: Evidence from the UK Employment Retention and Advancement (ERA) Demonstration.* Research Report 330. London: Department for Work and Pensions.

Earlier MDRC Working Papers on Research Methodology

Using Covariates to Improve Precision Empirical Guidance for Studies That Randomize Schools to Measure the Impacts of Educational Interventions
2005. Howard Bloom

Sample Design for an Evaluation of the Reading First Program
2003. Howard Bloom

Intensive Qualitative Research Challenges, Best Uses, and Opportunities
2003. Alissa Gardenhire, Laura Nelson

Exploring the Feasibility and Quality of Matched Neighborhood Research Designs
2003. David Seith, Nandita Verma, Howard Bloom, George Galster

Can Nonexperimental Comparison Group Methods Match the Findings from a Random Assignment Evaluation of Mandatory Welfare-to-Work Programs?
2002. Howard Bloom, Charles Michalopoulos, Carolyn Hill, Ying Lei

Using Instrumental Variables Analysis to Learn More from Social Policy Experiments
2002. Lisa Gennetian, Johannes Bos, Pamela Morris

Using Place-Based Random Assignment and Comparative Interrupted Time-Series Analysis to Evaluate the Jobs-Plus Employment Program for Public Housing Residents
2002. Howard Bloom, James Riccio

Measuring the Impacts of Whole School Reforms
Methodological Lessons from an Evaluation of Accelerated Schools
2001. Howard Bloom

A Meta-Analysis of Government-Sponsored Training Programs
2001. David Greenberg, Charles Michalopoulos, Philip Robins

Modeling the Performance of Welfare-to-Work Programs
The Effects of Program Management and Services, Economic Environment, and Client Characteristics
2001. Howard Bloom, Carolyn Hill, James Riccio

A Regression-Based Strategy for Defining Subgroups in a Social Experiment
2001. James Kemple, Jason Snipes

Explaining Variation in the Effects of Welfare-to-Work Programs
2001. David Greenberg, Robert Meyer, Charles Michalopoulos, Michael Wiseman

Extending the Reach of Randomized Social Experiments
New Directions in Evaluations of American Welfare-to-Work and Employment Initiatives
2001. James Riccio, Howard Bloom

The Politics of Random Assignment: Implementing Studies and Impacting Policy
2000. Judith Gueron

Assessing the Impact of Welfare Reform on Urban Communities
The Urban Change Project and Methodological Considerations
2000. Charles Michalopoulos, Joannes Bos, Robert Lalonde, Nandita Verma

Building a Convincing Test of a Public Housing Employment Program Using Non-Experimental Methods
Planning for the Jobs-Plus Demonstration
1999. Howard Bloom

Estimating Program Impacts on Student Achievement Using "Short" Interrupted Time Series
1999. Howard Bloom

Using Cluster Random Assignment to Measure Program Impacts
Statistical Implications for the Evaluation of Education Programs
1999. Howard Bloom, Johannes Bos, Suk-Won Lee

About MDRC

MDRC is a nonprofit, nonpartisan social and education policy research organization dedicated to learning what works to improve the well-being of low-income people. Through its research and the active communication of its findings, MDRC seeks to enhance the effectiveness of social and education policies and programs.

Founded in 1974 and located in New York City and Oakland, California, MDRC is best known for mounting rigorous, large-scale, real-world tests of new and existing policies and programs. Its projects are a mix of demonstrations (field tests of promising new program approaches) and evaluations of ongoing government and community initiatives. MDRC's staff bring an unusual combination of research and organizational experience to their work, providing expertise on the latest in qualitative and quantitative methods and on program design, development, implementation, and management. MDRC seeks to learn not just whether a program is effective but also how and why the program's effects occur. In addition, it tries to place each project's findings in the broader context of related research — in order to build knowledge about what works across the social and education policy fields. MDRC's findings, lessons, and best practices are proactively shared with a broad audience in the policy and practitioner community as well as with the general public and the media.

Over the years, MDRC has brought its unique approach to an ever-growing range of policy areas and target populations. Once known primarily for evaluations of state welfare-to-work programs, today MDRC is also studying public school reforms, employment programs for ex-offenders and people with disabilities, and programs to help low-income students succeed in college. MDRC's projects are organized into five areas:

- Promoting Family Well-Being and Child Development
- Improving Public Education
- Raising Academic Achievement and Persistence in College
- Supporting Low-Wage Workers and Communities
- Overcoming Barriers to Employment

Working in almost every state, all of the nation's largest cities, and Canada and the United Kingdom, MDRC conducts its projects in partnership with national, state, and local governments, public school systems, community organizations, and numerous private philanthropies.