# Randomizing Groups to Evaluate Place-Based Programs

**Howard S. Bloom**
**MDRC**

**Draft**

**March 2, 2004**

This document is being made available to participants in a methodological workshop at the annual conference of the Society for Research on Adolescence. Please do not distribute the document without permission from the author (howard.bloom@mdrc.org). It is a chapter in a book manuscript that is under revision for publication by the Russell Sage Foundation. The manuscript is entitled ***Learning More From Social Experiments: Evolving Analytic Approaches*** and was edited by Howard S. Bloom.

## Introduction

Many important social interventions aim to change whole communities or organizations. Examples of such *place-based* initiatives include community improvement programs, school reforms, and employer-based efforts to retain workers. Because such programs are designed to affect the behavior of groups of interrelated people rather than disparate individuals, it is generally not feasible to measure their effectiveness in an *experiment,* that is, by randomly assigning each individual to the program or to a control group. By randomizing at the level of groups such as neighborhoods, schools, or companies, however, researchers can still reap most of the methodological benefits afforded by random assignment.

Perhaps the earliest application of place-based random assignment was Gosnell's (1927) study of ways to increase voter turnout. After dividing each of 12 local districts in Chicago into two parts, he randomly chose one part of each district as a target for a series of hortatory mailings and used the other part as a control group. This research was conducted a decade before Ronald A. Fisher, the father of randomized experiments, published his landmark book on the use of randomization to study cause and effect (Fisher 1947/1937).[1] But it was not until about 20 years ago that evaluators began to use group randomization with any frequency. Given that its application was confined mostly to the field of health research, it is no accident that the

[1] Charles S. Peirce and Joseph Jastrow put individual randomization to its earliest known use in a study of minimum perceivable differences in the weights of physical objects and in their later studies of mental telepathy (Peirce and Jastrow 1885; for discussion, see Hacking 1988). Fisher (1926) first wrote about randomization in an article that focused on agricultural experiments.

only textbooks on group randomization ever published, one by Donner and Klar (2000) and the other by Murray (1998), focus on evaluating health programs.[2]

The use of group randomization to study the effects, or *impacts,* of social policies is now spreading to many fields (for a review, see Boruch and Foley 2000). Over the past decade, it has been used to evaluate "whole-school" reforms (Cook, Hunt, and Murphy 2000), school-based teacher training programs (Blank et al. 2002), community health promotion campaigns (Murray et al. 1994a), school-based smoking, drinking, and sex prevention programs (Flay 2000), community employment initiatives (Bloom and Riccio 2002), police patrol innovations (Sherman and Weisburd 1995), family planning programs (Smith et al. 1997), rural poverty reduction initiatives (Teruel and Davis 2000), HIV prevention programs (Sikkema et al. 2000), and group medical practice interventions (Leviton et al. 1999; Eccles et al. 2001). To help foster more frequent and better-informed use of group randomization, this chapter explores the rationale, nature, and consequences of place-based programs; the role, design, and implementation of group randomization evaluations of such programs; and the statistical properties and substantive implications of these evaluation designs.

[2] In his textbook on experimental design in psychology and education — one of the first to be published on the subject — Lindquist (1953) provides an excellent overview of group randomization.

## Reasons for Place-Based Evaluation

There are five main reasons why a research team might choose to study a program in a place-based design using group randomization. The first three depend on features of the program to be evaluated:

1. The effects of the program have the potential to "spill over" to a substantial degree among participants or from participants to nonparticipants.
2. The program's services are delivered most efficiently when targeted at specific locations.
3. The program is designed to address a spatially concentrated problem or situation.

Because such interventions are place-based in nature, it makes sense to evaluate them in a place-based research design. This means conducting random assignment at the level of groups of people who live, work, or receive services in the places being examined. The other two reasons for place-based evaluation relate to the difficulties of implementing random assignment experiments in real-world settings:

1. Using a place-based design will reduce political opposition to randomization.
2. Maintaining the integrity of the experiment requires the physical separation of program group members from control group members.

These situations involve programs that, though not inherently place-based, are studied more readily when random assignment is conducted in a place-based way.

In the next section, I expand on each of these reasons, describing first its conceptual basis and then giving one or more programmatic examples of it. When possible, I also present examples of evaluations that have used group randomization for one or more of the reasons listed above. In some of these real-world examples, group randomization was used for more than one reason, and so the use of an example to illustrate a given reason should not be taken to imply that this reason was the only relevant one.

### Containing Spillover Effects

For decades, scholars have stressed the importance of tailoring evaluations to the theories underlying the programs to be studied. This emphasis is variously referred to as “theory-driven” evaluation (Chen and Rossi 1983), “theory-based” evaluation (Cook, Hunt, and Murphy 2000), and “theory of change” evaluation (Connell and Kubisch 1998). The primary theoretical reason for place-based evaluation is *spillover effects,* which occur when the outcomes for some program participants influence those for other participants or for people who are not participating in the program.

Spillover effects can reflect interdependencies among different actors for a single outcome, independencies among different outcomes for a single actor, or both. Finding a job, for example, might spill over in a variety of ways. It might enable one to help friends or family members find jobs too. It might improve one’s mental health. Or it might enable one to help others find jobs, thereby improving one’s own and others’ mental health. If spillover effects are expected to be an important product or by-product of a program, then an evaluation of the program should account for them.

Spillover effects are recognized in many fields. For example, they play a central role in public finance theory, where they are referred to as “externalities” (Musgrave and Musgrave 1973). Externalities occur when a good or service consumed by one individual or group produces benefits or costs for others. For example, education creates direct benefits for its recipients and indirect benefits for society (a positive externality). The use of gasoline to fuel automobiles generates transportation benefits for drivers and passengers and imposes pollution costs on others (a negative externality).

Despite their theoretical and practical importance, spillover effects are difficult to accommodate in a causal model of individual behavior. Thus, when program impacts on individuals are estimated, spillover effects are usually ignored or assumed not to exist. When Rubin’s (1980) well-known Stable Unit Treatment Value Assumption (SUTVA)[3] holds — at least approximately — one can measure program impacts by studying individuals separately. When SUTVA does not hold, however, one must shift to a higher level of aggregation to internalize the spillover. This higher level is often defined spatially, that is, with respect to place.

### Spillover effects on a single outcome

The first type of spillover occurs when an outcome for one or more people affects the same outcome for other people. Social scientists have developed many causal models to explain how these spillover effects can occur. For example:

[3] Cox (1958, p. 19) refers to the absence of this condition as *interference between different units.*

- *Game theory models* seek to explain how one individual's response to a situation influences others' responses based on their preferences and how each individual's decision affects others' decisions. These models have been used to explore the occurrence of transitions in the racial composition of neighborhoods (Schelling 1971).
- *Network theory models* seek to explain how the flow of information among associated individuals influences their behavior. These models have been used to explain how employment is promoted through family and social connections (Granovetter 1973).
- *Peer group models* seek to explain how individuals' norms and behaviors are shaped by the norms and behaviors of the people with whom they associate. These models have been used to study how smoking, drinking, drug abuse, and violent behavior spread through a group, a neighborhood, or a school.
- *Microeconomic models* seek to explain how supply and demand conditions link consumers' purchasing decisions with producers' production decisions to determine the quantities and prices of goods, services, capital, and labor. These models have been used to explain how employment programs for one group of people can hurt others' employment prospects by reducing the number of job vacancies (Garfinkel, Manski, and Michalopoulos 1992).
- *Macroeconomic models* of income determination seek to explain how private investment, production, consumption, and savings decisions — combined

with government tax and spending policies — cause an increase in one group's income to ripple through an economy in successively larger waves. Such models are used to forecast economic growth (Branson 1972).

- *Chaos theory models*, which are mathematical representations of nonlinear dynamic systems, provide a general analytic framework for examining unstable equilibria (Kellert 1993), which are another tool for studying spillover effects.

Embedded within the models and theories above are two key features of spillover — *feedback effects* and *thresholds*. Feedback effects are changes in individual actions caused by previous individual actions. The feedback is positive if, for instance, increased smoking among some adolescents promotes smoking among their peers. The feedback is negative if, for instance, higher employment for one group worsens the job prospects of other groups.

Thresholds represent situations where behaviors change dramatically beyond a certain point (Granovetter 1978). The level at which this occurs is often called a "tipping point." Many examples of tipping have been identified, including racial transitions in neighborhoods, outbreaks of crime, and epidemics of disease. Compelling as the anecdotal evidence for this phenomenon is (Gladwell 2000),[4] only limited statistical evidence for it is available because of methodological difficulties associated with gathering such evidence.[5]

[4] Gladwell (2000) describes tipping in a wide range of contexts, including fashion (how Hush Puppy shoes sprang back into vogue), the food and entertainment industries (how restaurants and celebrities

(continued)

### Spillover effects among different outcomes

In the second type of spillover, one outcome affects another. According to Myrdal's (1944) principle of "cumulation," for example, intense social interactions among members of society make it possible for small changes on one dimension to produce large cumulative changes on other dimensions. He applied this principle to the relationship between white prejudice against nonwhites and the economic circumstances of nonwhites. More recently, Wilson (1996) posited that high rates of joblessness among adults in a community limit young people's exposure to positive role models and routine modes of living, which, in turn, increases the likelihood of antisocial and illegal behavior among adolescents in that community.

Although spillover effects across outcomes have been the topic of much theorization, little hard evidence about them exists. For example, the extensive literature examining the ef-

fall into and out of favor), criminal behavior (how crime rates plummeted in New York City during the 1990s), and transportation safety (how a few graffiti artists can spark an outbreak of subway crime).

[5] Galster, Quercia, and Cortes (2000) use U.S. Census data to estimate threshold effects for neighborhood characteristics such as the poverty rate, the unemployment rate, and the school dropout rate. Their findings are a rare example of statistical evidence about this phenomenon.

fects of neighborhoods on children is inconclusive because of conceptual, measurement, and statistical problems that remain to be solved.[6]

### Spillover effects and saturation programs

It is particularly easy to see how the two types of spillover effects just described could occur in the case of programs designed to "saturate" an area with services targeted at its entire population. A current application of the saturation approach, the Jobs-Plus Community Revitalization Initiative for Public Housing Families (Bloom and Riccio 2002) provides employment and training services, financial incentives to make work pay, and community supports for work to working-age adults in selected public housing developments in six U.S. cities. The program's designers hypothesized that, by exposing a high percentage of residents to this rich mix of services and activities, Jobs-Plus would induce a critical mass of participants to become employed, which, in turn, would motivate others to follow suit. This is an example of a spillover effect on the same dimension (employment) between groups of individuals (between employed and unemployed residents of the Jobs-Plus developments). The designers also hoped that, by substantially reducing the local concentration of unemployment, Jobs-Plus would have beneficial effects on the neighborhoods' physical and social environment. This is an example of a spillover effect across dimensions (from employment to outcomes such as crime and the housing vacancy

[6] Tienda (1991) identifies conceptual problems that hinder the study of neighborhood effects. Jencks and Mayer (1990) survey the empirical literature on such effects. Brooks-Gunn, Duncan, and Aber (1997) propose a comprehensive agenda for future research on the topic.

rate) both within the same group and between groups of individuals (among Jobs-Plus participants and between Jobs-Plus participants and other residents of the housing developments). Because producing these spillover effects is an explicit goal of the Jobs-Plus model — hence the "Plus" in its name — the program evaluation was planned to account for them by randomly assigning entire housing developments to the program or to a control group.

## Delivering Services Effectively

Another reason for operating programs that focus on groups of people defined by their location (instead of a dispersed set of individuals) is that place can be an effective platform for service delivery. For example, a program may capitalize on economies of spatial concentration, or it may aim to change the practices and cultures of existing organizations.

### Achieving economies of spatial concentration

Two major ways that place-based initiatives can achieve economies of spatial concentration are by benefiting from physical proximity to target group members and by leveraging existing channels of communication. Locating a program near its target group may enhance recruitment efforts by raising its profile; may reduce psychological barriers to participation by enabling people to participate in familiar territory; may reduce the costs of transportation — both in terms of money and time — to and from the program; and may enable staff to operate the program more effectively by exposing them directly to problems and possibilities in their clients' day-to-day lives.

By concentrating outreach in a few locations instead of dispersing it, some programs can make better use of both formal and informal channels of communication. For example, concentrated outreach can facilitate more comprehensive, coordinated, and frequent use of local media to heighten awareness of a problem being addressed, to publicize how a program will help solve the problem, and to inform target group members about how to participate. Saturating local media with a program's message may also stimulate further communication by word-of-mouth. In addition, it is easier to make direct personal contact with target group members when they are located in a small area. When outreach is concentrated spatially, it may be necessary to randomize entire areas — and, by implication, the groups they represent — so as to separate individuals who are supposed to receive the treatment being tested from those who are not. This is why lifestyle interventions to reduce health risk factors based on media outreach and information campaigns, for example, have been designed and tested as randomized place-based experiments (Murray et al. 1994b; Murray and Short 1995).

### Inducing organizational change

Some programs are designed explicitly to change the practices of existing organizations. For example, whole-school reforms are designed to transform the way that primary or secondary schools function by changing the timing, staffing, style, culture, and curriculum of entire schools. It is much easier to evaluate such initiatives by randomly assigning schools rather than individual students within a school. Two examples of this approach are the completed evaluation of the School Development Program (Cook, Hunt, and Murphy 2000) and an ongoing evaluation of Success for All (Slavin 2002).

A second example of programs for inducing organizational change are employer-based initiatives aimed at reducing turnover among employees. In attempting to improve procedures for training, supervising, and counseling employees, such programs focus on entire firms, not only individuals. They may include direct services to help employees meet the demands of their jobs more effectively and special training to help supervisors manage their employees better. Random assignment of firms is now being used to evaluate an employer-based program designed to reduce turnover among low-wage workers in the health care industry in Cleveland, Ohio (Miller and Bloom 2002).

A third example is programs designed to increase physicians' adoption of clinically proven innovations and to reduce their use of practices that have been shown to have harmful side effects. Although the ultimate goal of such initiatives is to change individual behavior, their focus is on transforming medical practices in entire organizations, such as hospitals and group medical practices. Such programs might provide educational activities to groups of physicians who work together, embed audit and feedback procedures within patient information systems for these groups, or stimulate other organizational changes to expedite diffusion of improved medical practices. Grimshaw et al. (in press) present a systematic review of 100 studies that randomly assigned physician groups or medical practices to evaluate interventions focused on these organizational units.

### Tackling Local Problems

In some cases, the nature of the problem being addressed or the test being conducted makes place a natural target.

### Nature of the Problem: When Locus is the Focus

For social problems whose spatial distribution is uneven, place-based solutions make sense. For example, because most crimes are concentrated geographically, crime reduction strategies are often targeted at specific locations. Over the past three decades, scholars of policing have focused on whether preventative patrol — an inherently place-based activity — can reduce crime (Sherman and Weisburd 1995). The first random assignment test of this approach was the Kansas City Preventative Patrol Experiment (Kelling et al. 1974). This landmark study randomly assigned 15 police beats to receive different patrol intensities and compared the subsequent crime rates across beats. The results of the study, which suggested that higher patrol intensities did not produce lower crime rates, had a major effect on police thinking and practice for many years thereafter. In the late 1980s, however, a major study that randomly varied police patrol intensities across 110 "hot spots," or small areas of concentrated crime, in Minneapolis found highly targeted intensive police patrol to be effective (Sherman and Weisburd 1995).

### Nature of the Test: When Programs Are Evaluated at Scale

The ultimate test of a program — especially if it provides an entitlement intended to benefit everyone eligible for it — is what would happen if it were implemented at full scale. Thus, it is important not only to measure the direct effects of the program on its participants but also to find out what its full-scale implementation would mean with respect to spillover effects (which in this context are often referred to as "system effects"), administration, and costs. Achieving the latter objective requires full implementation of the program in selected locations.

One of the most contentious issues in the debate about educational vouchers designed to promote school choice is potential system effects. Although researchers have measured the direct effects of school vouchers on small samples of students who were chosen to receive them through lotteries in three U.S. cities (Peterson et al. 2002), it is unclear how they would influence broader outcomes such racial and economic segregation if they were implemented at full scale. The only direct evidence on this issue is Ladd and Fiske's (2000) nonexperimental study of changes that occurred after New Zealand instituted a nationwide school choice program.

Possible system effects were also a major concern in a series of studies of housing allowances (a form of rental assistance for low-income people) conducted in the United States during the 1970s (Kennedy 1988). While the direct impacts, administrative feasibility, and costs of housing allowances were assessed in individual-level random assignment studies, their effects on the prices and quantities of low-cost housing were measured using nonexperimental analyses of changes in local housing markets following implementation of housing allowance entitlement programs.

The Youth Incentive Entitlement Pilot Projects, an initiative that guaranteed jobs to all interested 16- to 19-year-olds in 17 U.S. cities from 1978 to 1980, likewise included a component to measure system effects (Gueron 1984). The project used individual-level random assignment to examine the program's administrative feasibility and to measure its impacts on unemployment and school success among the 76,000 young people who volunteered to participate. But it also conducted a nonexperimental analysis of the program's impacts on the youth labor market based on full implementation of the program in four of the cities in the experimental study.

Although none of these full-scale tests involved randomizing places, it is possible to imagine doing so. Nevertheless, given the vastness of such an enterprise, its practical application is probably limited to a small number of very important programs.

### Facilitating Randomization

A very different type of reason for testing a program in a place-based experiment is to facilitate political acceptance of randomization by offsetting ethical concerns about "equal treatment of equals."[7] Random assignment treats sample members equally in the sense that each one has an equal chance of being offered the program. This fact is often overlooked, however, because, after randomization, program group members have access to the program while control group members do not.

Place-based randomization is generally easier to "sell" than individual randomization in at least three ways. It can assuage the political concerns of policymakers and program managers, who often cannot accept random assignment of individuals within their organizations but might be open to randomization across organizations. It can circumvent legal restrictions that prohibit programs from treating individuals in the same political jurisdiction differently but that do not prohibit them from treating different jurisdictions differently.[8] And it can capitalize on

[7] This principle, which is central to the theory of taxation in public finance, is often referred to as *horizontal equity* (Musgrave, 1959, pp. 160-161).

[8] Teruel and Davis (2000) describe such a legal restriction on a large-scale program to reduce rural poverty in Mexico.

the fact that much program funding is allocated at the level of political jurisdictions, which opens the door to assigning new funding to jurisdictions on a random basis — at least when funds are so limited that not all jurisdictions will receive them.

### Avoiding Control Group Contamination

One of the greatest threats to the methodological integrity of a random assignment research design is the possibility that some control group members will be exposed to the program, thus reducing the treatment contrast. Such *control group contamination* is especially likely in the case of programs that provide promotional messages and information. For example, if individual students in a school are randomly assigned to a personalized antismoking program, they will most likely share some of the information they glean from the program with peers who were randomly assigned to the control group. The resulting attenuation of the treatment contrast makes it difficult to interpret impact estimates.

One way to guard against this problem is to spatially separate the program group from the control group using place-based randomization. Randomly assigning homerooms, instead of individual students, to an antismoking program or to a control group can limit the extent to which program and control group members share information. However, for some types of programs, even this degree of separation may not be adequate, and it might be better to randomize larger entities, such as schools.

Similarly, in an experiment testing ways to induce physicians to use proven new medical procedures, randomization of individuals might undermine the treatment contrast because physicians often share information in the course of working together in group practices. Thus, it might be preferable to randomize group practices. But if the group practices share privileges at the same hospitals, it might be better to randomize hospitals.

The first major evaluation of the children's television program *Sesame Street* is an illuminating example of control group contamination and how to avoid it (Bogatz and Ball 1971). In the first year of the evaluation, each eligible household in five U.S. cities was randomly assigned to a group that was encouraged to watch *Sesame Street* or to a control group that was not. When the data on who had watched the program were analyzed, it was discovered that most control group members had also watched *Sesame Street* and thus received the treatment being tested. Therefore, in the next phase of the study, which was conducted in two other cities, the program and control groups were separated spatially. This separation was accomplished in one of the new cities by installing free cable television in all households with eligible young children who lived in groups of street blocks selected randomly from a larger pool. The remaining households, which did not receive free cable service, served as a control group. Because all the households were in low-income areas where cable television was prohibitively expensive at the time and because *Sesame Street* was only available through cable in those areas, very few households in the control group blocks could watch the program. Thus, place-based randomization greatly reduced the likelihood of control group contamination.

## Properties of Group Randomization

Having explored some of the most important reasons for using group randomization, I now examine the statistical properties of the approach.[9] As I will discuss in detail, impact estimates based on group random assignment, like those based on individual random assignment, are unbiased. But estimates based on group randomization have less — often, a lot less — statistical precision than those based on individual randomization. The relationship between these two types of randomization is thus analogous to that between cluster sampling and random sampling in survey research (Kish 1965). I begin the discussion of statistical properties by introducing the basic model of program impacts measured in a group randomization design.

### Model of Program Impacts

Consider a situation in which there are *J* groups of *n* individual members each. Assuming that a proportion *P* of these groups are randomly assigned to the program under study and that the rest (proportion 1 – *P*) are randomly assigned to a control group, the program's impact on outcome *Y* can be represented as:

$$Y_{ij} = \alpha + B_0 T_{ij} + e_j + \varepsilon_{ij}\,, \qquad (1)$$

where:

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

[9] For details, see Raudenbush (1997), Murray (1998), or Donner and Klar (2000).

$\alpha$ = the mean outcome for the control group

$B_0$ = the true program impact

$T_{ij}$ = 1 for program group members and 0 for control group members

$e_j$ = the error component for group j

$\varepsilon_{ij}$ = the error component for individual i from group j

The true program impact, $B_0$, is the difference between the mean outcome for program group members and what this mean outcome would have been in the absence of the program. The sample-based estimate of the impact, $b_0$, is the difference between the mean outcome for the program group and the mean outcome for the control group. The random error for this estimator has two components — $e_j$ for group differences and $\varepsilon_{ij}$ for individual differences — that are assumed to have independent and identical distributions with means of 0 and variances of $\tau^2$ (for $e_j$) and $\sigma^2$ (for $\varepsilon_{ij}$). Often referred to as a multilevel, hierarchical, random coefficients, or mixed model, Equation 1 can be estimated using widely available software.[10]

[10] Among the software packages that can perform this kind of analysis are HLM, SAS Proc Mixed, Stata gllamm6, MLWiN, and VARCL.

### Bias and Precision of Impact Estimators

Because randomization is the basis for the analysis, the expected value of the impact estimator is the true program impact. Because randomization was conducted at the group level, the standard error of the impact estimator is the following:[11]

$$SE(b_0)_{GR} = \sqrt{\frac{1}{P(1-P)}}\sqrt{\frac{\tau^2}{J}+\frac{\sigma^2}{nJ}} \qquad (2)$$

If, instead of randomizing the *J* groups to the program or the control group, one had randomized their *nJ* members individually, the expected value of the program impact estimator would still be the true impact, but its standard error would be the following:

$$SE(b_0)_{IN} = \sqrt{\frac{1}{P(1-P)}}\sqrt{\frac{\tau^2}{nJ}+\frac{\sigma^2}{nJ}} \qquad (3)$$

Thus, unless $\tau^2$ equals 0, the standard error for group randomization is larger than its counterpart for individual randomization.

The magnitude of the difference between $SE(b_0)_{GR}$ and $SE(b_0)_{IN}$ depends on the relationship between $\tau^2$ and $\sigma^2$ and the size of each randomized group. The relationship between $\tau^2$

[11] Equation 2 is based on a related expression in Raudenbush (1997, p. 176).

and $\sigma^2$ is usually expressed as an intraclass correlation (Fisher 1925), $\rho$, which equals the proportion of the total population variance ($\tau^2 + \sigma^2$) across groups as opposed to within groups:

$$\rho = \frac{\tau^2}{\tau^2 + \sigma^2} \tag{4}$$

Equations 2, 3 and 4 imply that the ratio between the standard error for group randomization and that for individual randomization — given a fixed total number of individual sample members — is a *group effect multiplier* that can be expressed as:

$$GEM = \sqrt{1+(n-1)\rho} \tag{5}$$

This group effect multiplier is the same as the well-known design effect in cluster sampling (Kish 1965).

Equation 5 indicates that, for a given total number of individuals, the standard error for group randomization increases with the size of the randomized groups, *n,* and with the intraclass correlation, $\rho$. Given that the intraclass correlation reflects the group effect (which is what inflates the standard error), it should not be surprising that the standard error increases with it. The group size comes into play here, too, because larger groups imply fewer groups to be randomized — and thus a larger margin for random error.

Table 4.1 illustrates these relationships. First, note that if the intraclass correlation is 0 (that is, if there is no group effect), the group effect multiplier is 1, and the standard errors for group randomization and individual randomization are the same. Next, note that large groups imply relatively large standard errors, even when the intraclass correlation is small. For example, if $\rho$ equals 0.01, randomizing $J$ groups of 500 people each will produce standard errors 2.48 times those produced by separately randomizing $500J$ individuals. Thus, randomizing public housing developments to evaluate a saturation employment program (Bloom and Riccio 2002) or randomizing communities to evaluate a health promotion campaign (Murray et al. 1994a) can produce large standard errors for program impact estimators.

[Table 4.1 around here]

Because the intraclass correlation captures the degree to which the outcome is stratified by randomized group, its value varies with the type of outcome (for example, academic performance, employment, or health risks) and the type of group (for example, schools, communities, or hospitals). The limited empirical literature on this issue suggests that, for numerous outcome measures and policy domains, intraclass correlations generally range between 0.01 and 0.10 and are concentrated between 0.01 and 0.05.[12] Furthermore, it appears that groups that represent small areas or organizational units (such as census tracts or classrooms) usually have lar-

[12] Murray et al. (1994b) and Siddiqui et al. (1996) report intraclass correlations for measures of adolescent smoking grouped by schools. Murray and Short (1995) report intraclass correlations for measures of adolescent drinking grouped by community. Campbell, Mollison, and Grimshaw (2001) and Campbell, Grimshaw, and Steen (2000) report intraclass correlations for measures of physician practices and patient outcomes grouped by hospitals and by physician groups. Ukoumnuune et al. (1999) report other kinds of intraclass correlations. I thank Jeremy Grimshaw, Brian Flay, and David Murray for directing me to these sources.

ger intraclass correlations — that is, are more homogeneous — than larger groups (such as municipalities or schools).

For intraclass correlations in the middle of the range that is typically observed, group randomization affords much less precision than individual randomization. For example, given an intraclass correlation of 0.05 and groups of 50 individuals each, the standard error of an impact estimator for group randomization is 1.86 times its counterpart for individual randomization.

As Table 4.1 underscores, the benefits of group randomization can come at a high cost with regard to the standard errors of impact estimates. The table also illustrates the importance of properly accounting for grouping when computing standard errors. If one computed the standard errors in a group randomization design as if individuals had been randomized, the results would understate the true standard errors substantially, thereby giving one a false sense of confidence in the impact estimates. As Cornfield (1978, p. 101) aptly observed, “Randomization by group accompanied by an analysis appropriate to randomization by individual is an exercise in self-deception.”

## Implications for Sample Size

Equation 2 indicates how five factors determine the standard errors of program impact estimators based on group randomization. Two of these factors, $\tau^2$ and $\sigma^2$, reflect the underlying variation in the outcome of interest, which must be taken as given. The other three factors — *n, J,* and *P* — reflect the size of the evaluation sample and its allocation to the program and control groups, which are research design choices. In this section, I examine the effects of sample size (*n* and *J*) on precision.

### Using Minimum Detectable Effects to Measure Precision

When examining the precision of an experimental design, it is often helpful to express this property in terms of the smallest program effect that could be detected with confidence. Formally, a *minimum detectable effect* is the smallest true program effect that has a probability of $1 - \beta$ of producing an impact estimate that is statistically significant at the $\alpha$ level (Bloom 1995). The appendix to this chapter demonstrates that this parameter, which is a multiple of the impact estimator's standard error, depends on the following factors: whether a one-tailed t-test (for program effects in the predicted direction) or a two-tailed t-test (for any program effects) is to be performed; the level of statistical significance to which the result of this test will be compared ($\alpha$); the desired statistical power ($1 - \beta$, the probability of detecting a true effect of a given size or larger); and the number of degrees of freedom of the test, which — assuming a two-group experimental design and no covariates — equals the number of groups randomized minus 2 ($J - 2$).

Table 4.2 shows how the minimum detectable effect multiplier (and thus the minimum detectable effect) for one-tailed and two-tailed t-tests varies with the number of groups to be randomized, assuming use of the conventional statistical significance criterion of 5 percent and a statistical power level of .80. This pattern reflects how the t distribution varies with the number of degrees of freedom available. This feature of the t distribution, well known for a century, is not important for most studies based on individual randomization because they typically have many degrees of freedom.[13] When small numbers of groups are randomized and, thus, very few degrees of freedom are available, however, this pattern has very important implications for research design.

[Table 4.2 around here]

The minimum detectable effect is smaller for one-tailed tests than for two-tailed tests because, other things being equal, the statistical power of one-tailed tests is greater than that of two-tailed tests. This, too, is of less concern in studies based on individual randomization because of their much greater statistical power. In group randomization, however, the question of whether to use a one-tailed or a two-tailed test often deserves special consideration. And when small numbers of groups are randomized, the need for statistical power might tip the balance in favor of one-tailed tests.[14]

[13] This relationship was established formally in a paper by William Sealy Gosset published under his pseudonym "Student" (1908), but it was first brought to the attention of empirical researchers by Fisher (1925).

[14] The primary argument for using one-tailed tests in program evaluation is that such analyses are mainly intended to inform decisions about whether or not to support a program. Because it usually makes sense to support a program only if it produces beneficial effects, a one-sided alternative hypothesis — and thus a one-tailed test — is generally indicated. This rationale is quite different from the standard one in the social sciences, where one-tailed tests are recommended only when there are strong a priori reasons for expecting an effect in one direction and the purpose of statistical inference is to test theories rather than to inform program-related decisions.

Because program impact estimates are frequently reported in standardized form as effect sizes — where an effect size equals the impact estimate divided by the control group's standard deviation on the outcome measure[15] — it is useful to express precision as a *minimum detectable effect size.* A study with a minimum detectable effect size of 0.25, for example, can detect with confidence a true program impact equal to 0.25 standard deviation.

To assess the minimum detectable effect size for a research design, one needs a basis for deciding how much precision is needed. From an economic perspective, this basis might be whether the design can detect the smallest effect that would enable a program to break even in a benefit-cost sense. From a political perspective, it might be whether the design can detect the smallest effect that would be deemed important by the public or by public officials. From a programmatic perspective, it might be whether the study can detect an effect that, judging from the performance of similar programs, is likely to be attainable. Smaller minimum detectable effects imply greater statistical precision.

Although there is no standard basis for assessing the minimum detectable effect size, one widely used classification is Cohen's (1988/1977). He proposed that minimum detectable effect sizes of roughly 0.20, 0.50, and 0.80 be considered small, medium, and large, respectively. Lipsey (1990) provided empirical support for this characterization by examining the actual distribution of 102 mean effect size estimates reported in 186 meta-analyses, together representing 6,700 studies with 800,000 sample members. Consistent with Cohen's scheme, the

[15] This standardized metric is often used in meta-analyses to synthesize findings across outcomes and studies (Hedges and Olkin 1985).

lower third of this distribution ranges from 0.00 to 0.32, the middle third ranges from 0.33 to 0.55, and the upper third ranges from 0.56 to 1.20.

### How Sample Size Affects Minimum Detectable Effects

Now consider how the minimum detectable effect size for group randomization, $MDES(b_0)_{GR}$, varies with the number and size of the groups randomized, given the intraclass correlation and the proportion of the groups randomly assigned to the program, *P.* Equation 6, which is derived in this chapter's appendix, represents this relationship as follows:

$$MDES(b_0)_{GR} = \frac{M_{J-2}}{\sqrt{J}}\sqrt{\rho+\frac{1-\rho}{n}}\sqrt{\frac{1}{P(1-P)}}\ , \qquad (6)$$

where $M_{J-2}$ is the minimum detectable effect multiplier in Table 4.2.

The number of groups randomized influences precision through $M_{J-2}$, which varies appreciably only for small numbers of randomized groups and also as a function of $1/\sqrt{J}$. Hence, for many potential applications, the minimum detectable effect size declines in roughly inverse proportion to the square root of the number of groups randomized.

The size of the groups randomized often makes far less difference to the precision of program impact estimators, especially given a moderate to high intraclass correlation. This is because the effect of group size is proportional to $\sqrt{\rho+\frac{1-\rho}{n}}$. For example, if $\rho$ were equal to 0.05, the values of $\sqrt{\rho+\frac{1-\rho}{n}}$ for randomized groups of 50, 100, 200, and 500 individuals each would be approximately 0.26, 0.24, 0.23, and 0.23, respectively. Thus, even a tenfold increase in the size of the groups randomized makes very little difference to the precision of program impact estimators.

Table 4.3 lists values for the minimum detectable effect sizes implied by a wide range of sample sizes and intraclass correlations. These findings are for experiments where $P = .50$. Other things being equal, higher intraclass correlations imply larger minimum detectable effect sizes. For example, compare 1.77, 2.04, and 2.34 in the upper left corner of each panel of the table (for $n = 10$ and $J = 4$); and then compare 0.59, 1.22, and 1.71 in the upper right corner of each panel (for $n = 500$ and $J = 4$). Moreover, increasing the number of groups randomized reduces the minimum detectable effect size. For $n = 10$ and $\rho = 0.01$, for example, increasing the number of groups from 4 to 20 reduces the minimum detectable effect size from 1.77 to 0.44. Scanning the columns within each panel in the table shows that this general result holds independent of the number of groups and the intraclass correlation.

[Table 4.3 around here]

Finally, for a given total number of sample members, increasing group size improves the precision of impact estimates by much less than does increasing the number of groups randomized. For example, for $\rho = 0.01$, the minimum detectable effect size for 4 groups of 10 individuals each is 1.77. Whereas doubling the size of each randomized group reduces this parameter to 1.31, doubling the number of groups randomized reduces it to 0.78. In fact, the size of the groups to be randomized often has almost no influence on precision. For example, for $\rho = 0.05$, increasing the size of each randomized group from 50 to 500 individuals reduces the minimum detectable effect size very little; and for $\rho = 0.10$, the reduction is negligible.

In summary, then, randomizing groups instead of individuals puts precision at a premium. And randomizing more groups almost always boosts precision more than does randomizing larger groups.

## Implications for Sample Allocation

Now consider how *sample allocation* — that is, the proportion of groups randomized to the program rather than to the control group — affects the precision of program impact estimators.

### Balanced Versus Unbalanced Allocations

Virtually all research methodology textbooks prescribe a balanced allocation of sample members to the program and control groups ($P = 1 - P = .50$) because under conditions of *homoscedasticity* — that is, when the variance of the outcome measure is the same for the program group as it is for the control group — balanced allocation maximizes the statistical precision of impact estimators.[16] Generally overlooked, however, is the fact that precision erodes quite slowly as sample allocation departs from balance. Hence, there is more latitude than is commonly thought for using unbalanced allocations when the homoscedasticity assumption is a

[16] This discussion makes the simplifying assumption that each randomized group has the same number of individual members.

reasonable approximation.[17] This latitude can enable researchers to capitalize on such opportunities to increase precision as the availability of public administrative records, which can be used to construct large control groups at low cost. It can also facilitate randomization by allowing for the use of small control groups, which reduces the number of individuals who must be denied access to a program and can thereby reduce political resistance to the approach.

Decisions about sample allocation are more complicated under conditions of *heteroscedasticity* — that is, when the variances of the outcome measure are not the same for the program and control groups. This situation arises when a program produces impacts that vary across individuals or groups.[18] For example, the impacts of whole-school reforms on student achievement may be larger for some types of students or for some types of schools than for others. In such cases, a balanced sample allocation provides greater methodological protection because it is more robust to violations of the assumptions of homoscedasticity.

### When the Variances Are Equal

The findings discussed so far in this chapter assume that $\tau^2$ and $\sigma^2$ are the same for the program group as for the control group. Equation 6 indicates that when this is the case, the mini-

[17] Bloom (1995) demonstrates this latitude for individual randomization, while Liu (2002) demonstrates it for group randomization.

[18] Bryk and Raudenbush (1988) argue that one should expect program impacts to vary across individuals and that this variation provides an opportunity for learning how individuals respond to programs.

minimum detectable effect size is proportional to $\sqrt{\frac{1}{P(1-P)}}$. This expression is minimized when $P$ equals 0.5, as is the corresponding minimum detectable effect size. The same expression can be used to demonstrate that, given a fixed sample size ($n$ and $J$), precision changes very little until one approaches extreme imbalance. To see this, note that $\sqrt{\frac{1}{P(1-P)}}$ equals 2.00, 2.04, 2.18, 2.50, and 3.33 when $P$ is 0.5, 0.6, 0.7, 0.8, and 0.9, respectively. The pattern is the same when $P$ is 0.5, 0.4, 0.3, 0.2, and 0.1, respectively. And it holds regardless of the number of groups randomized, the size of the groups randomized, and the degree of intraclass correlation.

Table 4.4 illustrates the point more concretely. The first column lists sample allocations ranging from $P = .10$ to $P = .90$. The next two columns present the minimum detectable effect sizes for each sample allocation, given two hypothetical sets of values of $n$, $J$, and $\rho$. The fourth column displays the ratio between the minimum detectable effect size for each sample allocation and the minimum detectable effect size for a balanced allocation; thus, when $P = .50$, this ratio is 1.00. As the table illustrates, the minimum detectable effect size changes very little until $P$ drops below .20 or exceeds .80.

[Table 4.4 around here]

## When the Variances Are Unequal

If a program creates impacts that vary across individuals or randomized groups, this can increase or reduce the individual or group variances relative to those for control group members. Consider how this might happen in the context of educational programs. Some programs may have larger than average effects on students who are initially weaker than average. If sufficiently pronounced, this tendency can reduce the individual error variance, $\sigma^2$, for members of the program group. The opposite result may occur if programs have larger than average effects on students who are initially stronger than average. Similarly, school-level responses to programs might vary, thereby reducing or increasing $\tau^2$ for the program group relative to the control group.

For balanced sample allocations, simulations and analytical proofs have demonstrated that statistical tests that assume equal variances for the program and control groups are valid even if the variances are unequal.[19] This is not true for unbalanced allocations, where the size of the inferential error depends on the relationship between the relative sizes of the program and control groups and the relative sizes of their variances.[20]

As a precaution in unbalanced allocation designs, one can estimate the program group variance and the control group variance separately and test the statistical significance of the difference between them. If the difference is statistically significant, the impact analysis can proceed using separate variance estimates. If the difference is not statistically significant, the impact analysis can proceed with a single, pooled variance estimate.

[19] Gail et al. (1996) use Monte Carlo simulations to illustrate this fact for parametric t-tests and non-parametric permutation tests. Kmenta's (1971) expression for the effect of heteroscedasticity in a bivariate regression can be used to prove the same point.

[20] See Gail et al. (1996) and Kmenta (1971).

In practice, however, given the small numbers of groups in a typical group randomization design, there are usually very few degrees of freedom with which to derive separate estimates of $\tau^2$. As a result, statistical tests of the significance of the difference in $\tau^2$ tend to have very little power. One might therefore opt to skip such tests and simply not assume that $\tau^2$ is the same for the program group as for the control group. Doing away with the homoscedasticity assumption does not circumvent the problem of limited degrees of freedom, however, because the resulting impact estimate is based on two separate estimates of $\tau^2$, each of which uses some of the degrees of freedom in the sample. Furthermore, as the imbalance between the number of program group members and the number of control group members increases, the number of degrees of freedom for the program impact estimator can only be approximated and approaches that for the smaller group. This can greatly reduce precision.[21]

The scarcity of degrees of freedom for estimating variances when homoscedasticity does not hold has received virtually no attention in the literature on social experiments, most likely because the vast majority of these experiments call for randomization of individuals rather than groups. In individual designs, a large number of individuals are typically randomized, and the only variance that must be estimated is $\sigma^2$. Thus, there are usually more than enough degrees of freedom to provide separate estimates of $\sigma^2$ for the program group and the control group. Researchers using randomized group designs do not have this luxury. Furthermore, because little is known about how the impacts of programs vary across types of individu-

[21] This finding reflects the number of degrees of freedom for a two-sample difference-of-means test given unequal variances and unbalanced samples (Blalock, 1972).

als and settings, it is not clear how problematic heteroscedasticity is likely to be. At this point in the development of randomized group studies, it therefore seems prudent to use balanced sample allocations whenever possible. Studies with relatively large numbers of groups to be randomized (say, 50 or more) might have greater flexibility in this regard, but even they probably should not depart too much from balance unless the benefits of doing so are compelling.

## Implications for Subgroup Analysis

I now discuss how to analyze a program's impacts for subgroups defined in terms of program characteristics, randomized group characteristics, and individual characteristics. A subgroup analysis addresses two basic questions: What is the impact of the program for each subgroup, and what are the relative impacts of the program across subgroups? The discussion consists of an intuitive explanation followed by a formulaic exposition; readers may skip the latter without losing the thread of the argument. The most striking feature of subgroup analysis in group randomization experiments is the small effect that subdividing the sample on the basis of individual characteristics has on the precision of impact estimators.

### Subgroups Defined by Characteristics of the Program

One way to think about sample subgroups is in terms of variants of the program being tested. For example, in a study of a program for reducing the use of X rays in testing patients for certain medical conditions, one could identify hospitals that implemented the program with high fidelity and hospitals that did not, splitting the program group in two on the basis of this distinction. It is not possible to estimate program impacts for such subgroups experimentally because there is no way to identify their counterparts in the control group. It might be feasible, however, to randomly assign different groups of hospitals to variants of the program for reducing X-ray use and to experimentally compare the outcomes across variants. Indeed, this approach, often referred to as a *multiarm trial,* has been used to test alternative ways of influencing physician practices (Eccles et al. 2001). But because each program variant tested substantially increases the number of groups to be randomized, the approach is probably only feasible for studying small numbers of program variants.

### Subgroups Defined by Characteristics of the Groups Randomized

Subgroups defined by characteristics of the groups randomized can provide valid experimental impact estimates. For example, if schools are randomized, one can observe how impact estimates vary by school size, average past performance, and urban versus suburban location. Likewise, if firms are randomized, one can observe how impact estimates vary by firm size, past employee turnover rates, and industry. These impact estimates are experimental because subdividing the program and control groups according to a characteristic that is determined before randomization (and that therefore could have not been influenced by assignment to the program or control group) creates valid "subexperiments." Hence, the difference between the mean program group outcome and the mean control group outcome in each subexperiment is an unbiased estimator of the program's *net impact* for the subgroup in question. Furthermore, the difference between the net impact estimates for two subgroups is an unbiased estimator of the program's *differential impact* on the subgroups.

Because each subgroup contains only a fraction of the groups that are randomized in the full experiment, however, the precision of subgroup analysis is substantially less than that of full sample analysis. Precision is lost in two ways: The smaller samples of randomized groups used in subgroup analysis both produce larger standard errors and provide fewer degrees of freedom.

To see how this works, consider an experimental sample with two mutually exclusive and jointly exhaustive subgroups, A and B. Assume that *n,* $\tau^2$*,* $\sigma^2$*,* and *P* are the same for both subgroups and for the full sample.[22] Proportion $\Pi_A$ of the randomized groups are in subgroup A and proportion 1 – $\Pi_A$ are in subgroup B. The appendix to this chapter demonstrates that the ratio between the minimum detectable effect size for subgroup A and that for the full sample is:

$$\frac{MDES(b_{0A})_{GR}}{MDES(b_0)_{GR}} = \frac{M_{\Pi_A J-2}}{M_{J-2}}\sqrt{1/\Pi_A} \qquad (7)$$

Equation 7 illustrates the two ways in which moving from the full sample to a subgroup increases the minimum detectable effect size. First, it increases the standard error of the impact estimator by decreasing the sample size — from *J* for the full sample to $\Pi_A J$ for the subgroup.

[22] If $\tau^2$ and $\sigma^2$ are the same for the subgroups as for the full sample, then the subgroups must have the same mean outcome. When this simplification does not hold, $\tau^2$ and $\sigma^2$ are smaller for the subgroups. Equations 7 and 8 may therefore understate the relative precision of subgroup findings. Nevertheless, because the same reduction in variance can be achieved for the full sample by controlling statistically for subgroup characteristics (as discussed later), this issue can be ignored for the moment.

Second, it increases the minimum detectable effect multiplier by decreasing the number of degrees of freedom — from $J - 2$ for the full sample to $\Pi_A J - 2$ for the subgroup.

To illustrate the likely magnitude of these effects on the minimum detectable effect size, consider a hypothetical example where a subgroup contains half the 20 groups that were randomized for an experiment. Hence, $\Pi_A$ equals 0.5, and:

$$\frac{MDES(b_{0A})_{GR}}{MDES(b_0)_{GR}} = \frac{M_8}{M_{18}}\sqrt{1/0.5} = \frac{3.353}{2.985}\sqrt{2} = 1.59$$

In this case, the minimum detectable effect size for subgroup A is 1.59 times that for the full sample.

The implications for differential impacts are more pronounced. The appendix to this chapter demonstrates that the ratio between the minimum detectable effect size of a differential impact estimator for subgroups A and B and that for the net impact estimator for the full sample is:

$$\frac{MDES(b_{0A} - b_{0B})_{GR}}{MDES(b_0)_{GR}} = \frac{M_{J-4}}{M_{J-2}}\sqrt{\frac{1}{\Pi_A(1-\Pi_A)}} \qquad (8)$$

Again, precision is reduced through an increase in the minimum detectable effect multiplier, caused by a decrease in the number of degrees of freedom from $J - 2$ to $J - 4$;[23] and an

[23] Because the differential impact estimator is a four-group "difference of differences of means" based on all randomized groups in the full sample, it has $J - 4$ degrees of freedom.

increase in the standard error, caused by a decrease in the sample size. But in a differential impact analysis, the increase in the minimum detectable effect size that occurs as one moves from the full sample to a subgroup reflects two factors: a smaller sample of randomized groups for each impact estimate and the dual uncertainty produced by taking the difference between the impact estimates for the subgroups. Thus, in the current example, the relative precision of a differential impact estimator is computed as follows:

$$\frac{MDES(b_{0A}-b_{0B})_{GR}}{MDES(b_0)_{GR}}=\frac{M_{16}}{M_{18}}\sqrt{1/((0.5)0.5)}=\frac{3.013}{2.985}\sqrt{4}=2.02$$

### Subgroups Defined by Characteristics of the Individual Sample Members

Subgroups defined by the characteristics of individual sample members can also provide valid experimental impact estimates. Thus, even if schools are the unit of randomization, one can measure program impacts experimentally for different types of students, such as boys or girls, whites or nonwhites, and previously high-performing students or previously low-performing students. If at least some students in every school in the sample have the characteristic of interest, one can proceed as if a separate subexperiment had been conducted solely on students in the subgroup. In this case, the only statistical difference between the subexperiment and the full experiment is the number of students per school.

The implications for precision of subgroups defined this way are entirely different from those already discussed. To see this, recall that the size of groups to be randomized has much less influence on precision than does the number of groups to be randomized and that, in some cases, group size hardly matters at all. This phenomenon determines the precision of impact estimates for subgroups of individuals. For example, assuming random assignment of schools, it is possible that the precision of net impact estimates for boys and girls separately will be almost the same as that for boys and girls together. Furthermore, as discussed below, the precision of an estimator for the differential impact on boys as opposed to girls can be greater than that of the estimator for the net impact on boys and girls together.

Consider two mutually exclusive and jointly exhaustive subgroups, I and II, defined by an individual characteristic such as gender. Assume that, in each randomized group, a proportion $\Pi_I$ of the individuals are in subgroup I and proportion $1 - \Pi_I$ are in subgroup II. Also assume that $\tau^2$ and $\sigma^2$ are the same for the two subgroups and for the full sample.

The appendix to this chapter demonstrates that the ratio between the minimum detectable effect size for subgroup I and that for the full sample is:

$$\frac{MDES(b_{0I})_{GR}}{MDES(b_0)_{GR}} = \frac{\sqrt{\rho + \frac{1-\rho}{\Pi I n}}}{\sqrt{\rho + \frac{1-\rho}{n}}} \tag{9}$$

Equation 9 illustrates how reducing the size of randomized groups by moving from the full sample to a subgroup defined by an individual characteristic can have little effect on precision. For example, assuming 100 individuals per randomized group in the full sample and an intraclass correlation of 0.05, the minimum detectable effect size for a subgroup net impact when there are 50 individuals in the subgroup per randomized group is 1.077 times that for the full sample — a mere 7.7 percent increase.[24]

The precision for subgroup differential impact estimators is even greater. As this chapter's appendix demonstrates:

$$\frac{MDES(b_{0I}-b_{0II})_{GR}}{MDES(b_0)_{GR}}=\sqrt{\frac{(1-\rho)}{\Pi_I(1-\Pi_I)(1+(n-1)\rho)}} \qquad (10)$$

Thus, with 100 individuals per randomized group in the full sample and 50 individuals per randomized group in each subgroup, the minimum detectable effect size for the differential impact estimator is only 0.80 times that for the full sample net impact estimator. The greater precision of the differential impact estimator derives from the fact that it "differences away" the group error component, $e_j$, and thereby eliminates $\tau^2$.

[24] The situation is even more favorable if the individual characteristic defining the subgroups is correlated with the outcome measure. (For example, boys generally score higher than girls on math tests.) In this case, part of the individual variance component, $\sigma^2$, is related to the subgroup characteristic and does not exist within subgroups. Also, if the subgroup mix varies across groups and the subgroup characteristic is correlated with the outcome, part of the between-group variance, $\tau^2$, is related to the subgroup characteristic and does not exist within subgroups. Because both of these improvements could be obtained for the full sample estimator by controlling statistically for subgroup characteristics as a covariate (as discussed later), they are not implications of performing subgroup analysis in group randomization experiments per se.

## Adjusting for Covariates

The next topic to consider is adjusting for covariates, which increases the precision of impact estimates by reducing the amount of unexplained variation in the outcome of interest. This approach is often used for experiments that randomize individuals. But its role can be even more important in experiments that randomize groups, where precision is more limited and therefore at a higher premium. Furthermore, because the correlations among features of aggregate entities are usually quite high (and typically much higher than the correlations among features of individuals), data on group characteristics can reduce the unexplained variation in the group error term — the binding constraint on precision in a group design — substantially.

### Aggregate Covariates, Individual Covariates, and Lagged Outcomes

In group randomization experiments, the two main types of covariates are aggregate characteristics of the groups randomized and individual characteristics of the group members. Although data on both types of covariates can be obtained in some contexts, it is often possible to collect only aggregate data on group characteristics given available resources.

Another important distinction is whether a covariate is a *lagged outcome* measure, that is, a measure of one of the outcomes of interest before randomization was conducted — or another type of *background characteristic.* In an experimental study of a new approach to reading instruction, for example, students' reading test scores before being randomly assigned to the program or a control group would be a lagged outcome measure, whereas students' sex and age would be background characteristics. Lagged outcome measures, often called *pretests,* are usually the most powerful covariates because they reflect the combined result of all the factors that determined the outcome in the past and that therefore are likely to influence it in the future. Put differently, the best predictor of a future outcome is almost always a past measure of the same outcome. Examples include the ability of past earnings to predict future earnings, of past criminal behavior to predict future criminal behavior, of past test scores to predict future test scores, and of past health status to predict future health status.

To provide a framework for this discussion, Equation 11 adds a single covariate, $X_{ij}$, to the program impact model in Equation 1, yielding:

$$Y_{ij} = \alpha + B_0 T_{ij} + B_1 X_{ij} + e^{*}_{j} + \varepsilon^{*}_{ij} \qquad (11)$$

Although $X_{ij}$ is defined to have a separate value for every member of the experimental sample, it can represent an individual characteristic or a group characteristic. Furthermore, it can represent a lagged outcome measure or another type of background characteristic. Note that Equation 11 assumes that the variance for the program group and the variance for the control group are equal.

The two error terms in Equation 11 — $e^{*}_{j}$ for each randomized group and $\varepsilon^{*}_{ij}$ for each individual sample member — differ from their counterparts in Equation 1 because they represent the unexplained variation between and within randomized groups after controlling for the covariate, $X_{ij}$. Therefore, the random error terms in Equation 11 are referred to as conditional errors, and those in Equation 1 are referred to as unconditional errors.

### Effects on Precision

Raudenbush (1997) derives expressions for the standard errors of impact estimators based on group randomization given a balanced sample allocation, a single group covariate or a single individual covariate, and equal variances for the program and control groups. Equations 12 and 13 below extend his findings to represent balanced or unbalanced allocations (with any value for *P*):

$$SE(b_{*0})_{GR} = \sqrt{\frac{1}{P(1-P)}}\sqrt{\frac{\tau_*^2}{J}+\frac{\sigma^2}{nJ}}\sqrt{1+\frac{1}{J-4}}\text{ , for a single group covariate}\qquad(12)$$

$$SE(b_{*0})_{GR} = \sqrt{\frac{1}{P(1-P)}}\sqrt{\frac{\tau_*^2}{J}+\frac{\sigma_*^2}{nJ}}\sqrt{1+\frac{1}{nJ-4}}\text{ , for a single individual covariate}\qquad(13)$$

Comparing these expressions with Equation 2, which includes no covariate, reveals several important differences. First, consider $\sqrt{1+\frac{1}{J-4}}$ in Equation 12 and $\sqrt{1+\frac{1}{nJ-4}}$ in Equation 13, which have no counterparts in Equation 2. The term in Equation 12 (which is undefined for $J \leq 4$) approaches 1 as the number of groups randomized increases. At 10 groups, the term equals 1.080 and is therefore unimportant for larger samples. Similarly, the term in Equation 13 (which is undefined for $nJ \leq 4$) approaches a value of 1 as the number of sample members increases. At 10 groups of 50 individuals each, it equals 1.001 and is therefore unimportant for almost any sample size that is likely to be used.

More important are the differences between the conditional variances, $\tau^{*2}$ and $\sigma^{*2}$, in Equations 12 and 13 and their unconditional counterparts, $\tau^2$ and $\sigma^2$, in Equation 2.[25] By controlling for some of the unexplained variation between randomized groups, a group characteristic can reduce the group variance from $\tau^2$ to $\tau^{*2}$. By controlling for some of the unexplained variation both within and between groups, an individual characteristic can reduce the group and individual variances from $\tau^2$ and $\sigma^2$ to $\tau^{*2}$ and $\sigma^{*2}$, respectively. In this way, covariates reduce the standard errors of program impact estimators — sometimes by a large amount — at a cost of only one degree of freedom per group characteristic and virtually no degrees of freedom per individual characteristic. Hence, the overall effect on precision of adjusting for a single covariate stems almost solely from its effect on standard errors, except in experiments that randomize very small numbers of groups.

### An Empirical Example: Randomizing Schools

Bloom, Bos, and Lee (1999) published the first empirical analysis of the extent to which using past test scores as covariates can improve the precision of educational program impact estimates based on randomization of schools. Their analysis was not conducted using data from an experimental design but rather the existing administrative records of one urban school district. Specifically, the authors estimated the between-school and within-school variance compo-

[25] Equation 12 includes an unconditional individual variance, $\sigma^2$, instead of a conditional one because the group covariates are constant within the groups and thus cannot explain within-group variation.

nents for the standardized math scores and reading scores of third-graders and sixth-graders in 25 elementary schools in Rochester, New York, in 1991 and 1992.

One type of covariate examined was a "student pretest" representing each student's score in the same subject in the preceding grade. For example, individual fifth-graders' scores were used as a student-level pretest for current sixth-graders. The other type of covariate examined was a "school pretest" representing each school's mean score during the preceding year in the same subject and grade. For example, the mean reading scores of sixth-graders in each school in the preceding year were used as a school-level pretest for current sixth graders.

Table 4.5 summarizes the variance estimates obtained. The top panel in the table lists estimates without covariates, the middle panel lists estimates with school pretests, and the bottom panel lists estimates with student pretests. The first four columns in the table present results for each subject and grade separately, and the last column presents the mean results. The findings clearly demonstrate the predictive power of pretests.

[Table 4.5 around here]

School pretests reduce the school variance for all subjects and grades from a mean of 18.0 to a mean of 4.4 — a dramatic reduction of 76 percent. The corresponding reductions by subject and grade range from 72 percent to 82 percent. School pretests do not affect the student variance because school pretest scores are the same for all students in a given annual cohort at a given school. (The slight variation in the findings with and without a school-level pretest merely reflects random error in the maximum likelihood estimates of the variance components.)

Student pretests reduce the school variance by roughly the same amount as do school pretests, although the pattern is not entirely consistent. In addition, student pretests reduce the student variance from a mean of 95.6 to a mean of 50.3 — or by 47 percent. The corresponding reductions by subject and grade range from 32 percent to 59 percent. The fact that student pretests reduce the student-level variance by proportionally less than school pretests reduce the school-level variance reflects the fact, already mentioned, that correlations tend to be smaller within individuals than within groups.

Because the school-level variance is the binding constraint on precision when schools are randomized, the ability of student pretests to reduce the student variance does not add much, if anything, to the precision of impact estimators except by reducing the school variance. Hence, the precision of program impact estimators is roughly the same for the two types of covariates. This point is clearly illustrated in Table 4.6, which presents the minimum detectable effect size implied by the estimated variances listed in Table 4.5 for a realistic, though hypothetical, education policy example. The example assumes a sample of 60 schools with 60 students per grade. Half the schools are randomly assigned to the program under study, and the other half are randomly assigned to a control group. Without covariates, the minimum detectable effect size ranges from 0.23 to 0.29 and averages 0.27; with school pretests as covariates, it ranges from 0.14 to 0.16 and averages 0.15; and with student pretests as covariates, it ranges from 0.09 to 0.25 and averages 0.16. In short, the power of pretests to improve statistical precision is considerable. Furthermore, school-level aggregate pretests have the advantage of being generally inexpensive to obtain relative to student-level individual pretests while increasing precision by about the same amount.

[Table 4.6 around here]

## Blocking and Matching

Baseline covariates can also be used to *block* or *match* groups before they are randomized. This is often done to reduce the potential for a "bad draw" — a situation in which the groups randomly assigned to the program group differ substantially from the groups randomly assigned to the control group, thus confounding the treatment with other variables. Blocking or matching thus increases the precision of program impact estimators by reducing their standard errors.

### Blocking Before Randomization

Blocking entails stratifying the groups to be randomized into blocks defined by specific combinations of baseline characteristics. After blocking, each of the groups in each block is randomly assigned to the program or to the control group. Ordinarily, the sample allocation is held constant across blocks.[26] In a blocking design, a balanced allocation ensures that the program and control groups each represent each block in the same proportion, which, in turn, guarantees that the program and control groups are identical with respect to the factors that define the blocks. For example, in a study of a reading program being implemented in schools in five different cities, a balanced allocation would involve grouping the schools by city and randomizing half the schools in each city (block) to the program and half the schools in each city to the

[26] Blocked randomization in experiments is analogous to proportionally stratified random sampling in survey research (Kish 1965).

control group. This would ensure that the program group and the control group each contain the same number of schools from each city.

There are two main criteria for defining blocks within which to randomize — face validity and predictive validity. Face validity is the degree to which characteristics that define blocks appear on their face to be important determinants of the outcome measure being used. Thus, when assessing the face validity provided by blocking on a set of characteristics, it is important to ask to what extent ensuring that the program and control groups have the same distributions of these characteristics would lend credibility to the evaluation findings. Blocking with respect to individual demographic characteristics such as age, gender, race, and ethnicity or with respect to aggregate group characteristics such as industry, type of organization, and location can boost face validity.

Predictive validity is the degree to which characteristics that define blocks predict — and thus can be used to control for — random variations in the outcome measure. As noted earlier, the best predictor of future outcomes is usually past outcomes, both for individuals and for groups. Thus, blocking with respect to a baseline measure of past outcomes is usually the best approach.

Given the small numbers of groups to be randomized in most cases and the large numbers of potential blocking factors, constructing blocks often requires making difficult trade-offs. Probably the most difficult trade-off is that between predictive validity and face validity. Although the trade-off is not necessary in principle, it is often necessary in practice — for example, when the need for predictive validity calls for blocking on past outcomes but the need for face validity calls for blocking on demographic characteristics, even if the latter do not add predictive power. Unfortunately, blocking on both characteristics usually reduces the quality of the match for each.

If blocking is used, it must be reflected in the corresponding estimates of program impacts and their standard errors. One simple way to take account of blocking is to define a separate 0/1 indicator variable, $I_k$, for each block except one and to add these variables to the basic impact model in Equation 1, yielding:[27]

$$Y_{ij} = \alpha + B_0 T_{ij} + \Sigma \gamma_k I_k + e_j^{**} \varepsilon_{ij}^{**} \qquad (14)$$

These indicator variables increase the explanatory power, or $R^2$, of the impact model and thereby reduce the standard error of the impact estimator. But they also reduce the number of degrees of freedom for estimating $\tau^2$ and thereby increase the minimum detectable effect. These two countervailing forces on precision must be taken into consideration when deciding whether to block and how many blocks to use.

### Pair-Wise Matching Before Randomization

A form of blocking, pair-wise matching entails stratifying groups into pairs before randomizing them. The best way to achieve predictive validity in matching is to rank the groups from highest to lowest with respect to their values on the baseline characteristic to be used and,

[27] As with any set of mutually exclusive, exhaustive categorical variables, one must always have a “left-out” category if one is to estimate a regression with an intercept.

starting with the pair with the highest values, to randomly assign one member to the program and the other member to the control group. The ranking can alternatively be based on a composite indicator that represents a set of baseline characteristics. Matching ensures that the program and control groups are as similar as possible in terms of the characteristic or characteristics used to identify the pairs.

To estimate program impacts and their standard errors in pair-wise matching, indicator variables for all but one pair should be added to the impact model. As observed in the discussion of blocking, there is a trade-off between the standard error and the minimum detectable effect. Although the indicator variables increase the $R^2$ of the model and thereby reduce the standard error of the impact estimate, they also reduce the number of degrees of freedom for estimating $\tau^2$ and thereby increase the minimum detectable effect multiplier. For example, in a design with 10 randomized groups, there are 8 degrees of freedom ($J - 2$) without matching, but only 4 degrees of freedom ($(J / 2) - 1$) with matching.

Because limited resources usually preclude randomizing more than a small number of groups, the large loss of degrees of freedom produced by matching groups actually reduces the precision of program impact estimates — unless the predictive power of the matching is very high.[28] A full treatment of this issue is beyond the scope of this chapter, but it is worth briefly illustrating the trade-offs involved. Consider the following expressions for the minimum detectable effect of an impact estimator given each of three different approaches: pair-wise matching

[28] Originally reported by Martin et al. (1993), this finding was derived independently by the present author using a different formulation.

with respect to a single group-level baseline characteristic, linear regression adjustment for the same characteristic, and no adjustment. In comparing these expressions, assume a fixed number of groups to be randomized and a fixed number of individuals per group.

$$MDE\ (b_{0m})_{GR} = M_{(J/2)-1}\sqrt{1-R_1{}^2}\,SE\,(b_0)_{GR} \tag{15}$$

given pair-wise matching, where $R_1{}^2$ is the predictive power;

$$MDE\,(b_{0r})_{GR} = M_{J-3}\sqrt{1-R_2{}^2}\sqrt{1+\frac{1}{J-4}}\,SE\,(b_0)_{GR} \tag{16}$$

given regression adjustment, where $R_2{}^2$ is the predictive power; and

$$MDE(b_0)_{GR} = M_{J-2}SE(b_0)_{GR} \tag{17}$$

given no adjustment, where $SE(b_0)_{GR}$ is the standard error of the impact estimator with no adjustment.

For no adjustment versus matching, the trade-off is between increasing the minimum detectable effect multiplier from $M_{J-2}$ to $M_{(J/2)-1}$ and reducing the standard error by a factor of $\sqrt{1-R_1{}^2}$. One way to capture this trade-off is to compute the minimum predictive power of matching, $R^2_{\min}$, that would offset the increased minimum detectable effect multiplier and thereby increase precision. This expression can be obtained by setting the right-hand side of Equation 15 (the minimum detectable effect given matching) less than or equal to the right-hand side of Equation 17 (the minimum detectable effect given no adjustment) and rearranging terms, yielding:

$$R^2_{\min} \geq 1 - \frac{M^2_{J-2}}{M^2_{J/2-1}} \qquad (18)$$

Based on this expression, Table 4.7 presents the minimum required predictive power of pair-wise matching given specific numbers of groups to be randomized. The first column presents results for a two-tailed hypothesis test, and the second column presents results for a one-tailed test.[29] The most striking result is that very high predictive power ( $R^2_{\min}$ ) is required for pair-wise matching to be justified assuming a small sample of randomized groups. For example, with six groups to be randomized, matching must predict 52 percent or 40 percent (for a two-tailed test or a one-tailed test, respectively) of the variation in the outcome measure before it improves the precision of the impact estimator relative to no adjustment for baseline characteristics. This is because, in small samples, even small differences in the number of degrees of freedom imply large differences in the minimum detectable effect multiplier.

[Table 4.7 around here]

### An Empirical Example: Randomizing Firms

I now use an ongoing evaluation of Achieve, an employer-based human resource program for reducing job turnover among low-wage workers in the health care industry, to illustrate how pair-wise matching works (Miller and Bloom 2002). Now being implemented by 22 Cleveland health care firms that volunteered to participate, the Achieve program offers employ-

[29] Differences between findings for two-tailed tests and one-tailed tests are most pronounced where there are small numbers of groups to be randomized and thus very few degrees of freedom.

ees a mix of direct services that include individual job counseling and group informational lunch sessions concerning job-related issues. It also provides indirect services to employees by training their supervisors to deal more effectively with issues that arise in the workplace. Because the program is being implemented on a firm-wide basis, it was not feasible to randomize individual employees. Therefore, group randomization was performed at the level of firms.

Firms were recruited for the study in two waves that took place roughly one month apart, with 8 firms in the first wave and 14 firms in the second wave. To maximize the predictive validity of the evaluation design, the firms in each wave were ranked according to their reported rates of employee turnover during the previous six months; and, in each pair, one firm was randomly assigned to the program, while the other firm was randomized to the control group. When it was discovered that the percentage of employees who were black was much higher in the program group than in the control group for one wave and much lower in the program group than in the control group for the other wave, the original assignment was reversed for one pair in each wave to improve the face validity of the evaluation design.[30]

The relative precision of the three approaches to increasing precision already discussed — pair-wise matching with respect to the baseline turnover rate, linear regression adjustment for the baseline turnover rate, and no adjustment — was analyzed using data on employee turnover rates during the first month after random assignment (a short-term outcome measure). The

[30] In retrospect, the researchers realized that such ad hoc adjustments to a random draw can inadvertently bias an experimental design. A better way to trade off predictive validity against face validity in this situation would have been to randomize the entire matched sample again.

predictive power of matching turned out to be large enough to raise the precision of this approach above that for both the alternatives, providing post hoc justification for the researchers' decision to match.

## Accounting for Mobility

An inescapable fact of life for place-based programs is that people move into and out of their places of residence, work, or study — often at a very high rate. For example, a recent analysis of selected public housing developments in four cities indicates that, on average, 29 percent of people who were residents in a given month had moved out two years later (Verma 2003); unpublished calculations indicate that, on average, 43 percent of the employees in the health care firms participating in the Achieve evaluation left their jobs within a six-month period; and the U.S. Department of Education (2003) estimates that only half of American kindergarteners are still at the same school when they reach the third grade.[31]

### Issues Raised by Mobility

Mobility raises important substantive and methodological issues for place-based programs and evaluations thereof. From a programmatic standpoint, the main issue is that mobility reduces program enrollees' exposure to program services because many of them leave before receiving a substantial "dose" of the treatment. High rates of mobility can thus undermine a place-based program's chance to effect meaningful changes in the outcomes of interest. From

[31] This figure is based on a calculation by the present author.

an evaluation standpoint, mobility has two main implications: It creates a need to conduct program impact analyses from multiple perspectives; and it increases the risk of *selection bias* in program impact estimates, which can arise if the program produces selective mobility into or out of the groups randomized and the resulting program and control samples differ from one other in ways that are related to the outcome of interest. To clarify these issues, it is useful to distinguish a program's impacts on people from its impacts on places.

Impacts on people reflect how a program changes outcomes for individuals targeted by the program. For example, one might ask whether an educational program increases reading achievement levels for students who are exposed to it. Although seemingly precise, this question needs further specification to be meaningful in a context where student mobility is high. One might ask whether the program raises reading achievement levels for all students who are present when the program is launched (whether or not they move away subsequently) or, alternatively, whether it does so for students who remain at their initial school throughout the analysis period. The policy question addressed in the first case is: "What are the impacts of the program on students in general when it is implemented under real-world conditions, which include mobility?" The policy question addressed in the second case is: "What are the impacts of the program on students who remain in one school long enough to receive a substantial dose of the treatment?" Both questions are meaningful and have different priorities for different purposes.

In a group randomization study, a longitudinal methodology — in which outcomes for the same individuals are tracked over time — is the most effective way to measure the impacts of programs on people. Thus, using a longitudinal design to measure the general effects of an educational program on students requires following all of them over a fixed period, even after some of them have left the school. Comparing the achievement-related outcomes for students in the program group with those for students in the control group produces valid experimental estimates of the program's impacts on student achievement. Interpretation of the estimates is complicated, however, by the fact that they represent an average response to what is usually a wide range of degrees of exposure to the program being evaluated.

An alternative approach is to conduct a longitudinal analysis only for "stayers" — that is, students in the program and control groups who remain in their initial school throughout the analysis period — so as to reduce variation in exposure to the program. This produces valid experimental estimates of the impacts of the program assuming that it does not influence mobility. But if the program affects the rates at which new students enter the school, current students leave the school, or both, the outcomes for stayers in the control group are not a valid indicator of what the outcomes would have been for members of the program group without the program. Thus, if a program affects mobility, comparing outcomes for stayers in the program group with those for stayers in the control group can introduce selection bias into impact estimates.

Impacts on places reflect how a program changes aggregate outcomes for locations or organizations targeted by the program. For example, one might ask whether an educational program increases reading achievement levels for schools that are exposed to it. The answer may reflect a mixture of two very different forces. The first is the effect of the program on the achievement of students who would remain in place with or without the program; the second is the effect of the program on student mobility. For example, an educational program could raise a school's achievement levels by improving the performance of students who would attend it with or without the program, by attracting and keeping more high-achieving students, or both.

Impacts on places are best measured in a group randomization study using a "repeated cross-section" methodology in which outcomes are tracked for the same places, rather than for the same individuals, over time. For example, to obtain valid experimental estimates of the impacts of a program on reading achievement for a group of schools, one might compare the reading test scores of successive annual cohorts (repeated cross-sections) of third-graders in the schools in the program group with those of successive annual cohorts of third-graders in the schools in the control group. However, to the extent that the program influences student mobility, it is not clear how to interpret the resulting impact estimates in the absence of further information.

### An Empirical Example: Randomizing Housing Developments

Based on randomization of matched pairs of public housing developments in six U.S. cities, the evaluation of Jobs-Plus mentioned earlier relies on a quasi-experimental method called *comparative interrupted time-series analysis* to measure the effects of this saturation employment initiative on public housing residents (Bloom and Riccio 2002).[32] The program's core elements are state-of-the-art employment-related activities and services, financial incentives designed to make work financially more worthwhile by reducing the rent increases that would otherwise occur when residents' earnings rise, and a range of activities designed to promote a community environment that is supportive of employment. As already discussed, these ele-

[32] For a detailed discussion of interrupted time-series analysis, see Shadish, Cook, and Campbell (2002).

ments are intended to create unusually large employment gains that, in turn, generate spillover effects throughout each participating development.

The Jobs-Plus evaluation is assessing impacts both from the perspective of public housing residents and from the perspective of public housing developments because moves into and out of public housing developments are frequent. The individual perspective focuses on how Jobs-Plus affected the people who were living in the participating developments at a specific point in time. This part of the impact analysis addresses the question: How did Jobs-Plus affect the future experiences of its target population, whether or not they moved away? The housing development perspective for Jobs-Plus, in contrast, focuses on how the program affected the housing developments participating in the study. This part of the impact analysis addresses the question: How did Jobs-Plus affect the conditions in its target environment, given that different people lived there at different times?

Figure 4.1 illustrates the comparative interrupted time-series approach as it is being used to estimate the impacts of Jobs-Plus from both perspectives. The graph at the top of the figure illustrates a hypothetical pattern of employment rates for residents in a Jobs-Plus development during the baseline period (before the program was launched) and during the follow-up period (after the program was launched). If Jobs-Plus increases employment, the rates during the follow-up period should rise above the baseline trend. The analysis focuses on comparing the deviations from the baseline trend in each Jobs-Plus development with those in the control group development with which it was matched before random assignment. The impact of Jobs-Plus on employment rates is estimated as the difference between the two sets of deviations.

[Figure 4.1 around here]

To make the analysis operational from the individual perspective, one must identify the people who resided in the developments at the time that Jobs-Plus was launched and follow their employment rates backward and forward in time, regardless of where they lived before and after the point of the program's launch. To make the analysis operational from the housing development perspective, one must identify the people who resided in the developments at each point during the baseline period and after the program's launch. Outcome data for the analyses from both perspectives are being obtained from public administrative records (Bloom and Riccio 2002).

## Reprise

This chapter lays out a research strategy that leverages the widely accepted scientific principle of randomization to evaluate place-based social programs. For theoretical or practical reasons, place-based programs are targeted at group-level units such as firms, neighborhoods, and schools rather than at individual-level units such as employees, residents, and students. Whereas in such programs it is usually not feasible to randomize individual members of the groups, it is often possible to randomize the groups themselves. Particularly because group randomization is being used with increasing frequency to measure the impacts of social programs, it is important for researchers to understand its special, and sometimes counterintuitive, properties. The key points are summarized below.

- **Precision is at a premium.**

Group randomization provides estimates of program impacts that are unbiased for the same reason that individual randomization does so. But impact estimates based on randomization of groups almost always have much less precision than do their counterparts for randomization of the same number of individuals.

- **The number of groups randomized is usually a more important determinant of precision than is group size.**

In most contexts, resources allow for randomization of only a small number of groups, putting a strong constraint on the precision of program impact estimates. Consequently, increasing the number of groups by a given proportion usually improves precision by a much greater amount than does increasing the number of individuals per group by the same proportion.

- **Covariates can improve the precision of program impact estimates.**

Regression adjustments for a baseline covariate, especially if the covariate is a lagged outcome measure, can substantially increase the precision of program impact estimates. This finding holds both when the covariate is an aggregate characteristic of the groups randomized and when it is an individual characteristic of the group members.

- **Subgroup analyses can have counterintuitive properties.**

Estimates of impacts for subgroups of an evaluation sample often have properties that set them apart from those based on randomization of individuals. For some subgroups defined in terms of individual characteristics, for example, program impact estimates offer almost as much precision as do corresponding impact estimates for a full study sample.

- **To improve precision, the characteristics used in pair-wise matching must have considerable predictive power.**

For an evaluation with a small number of groups to be randomized (say, less than 10), the gains in precision produced by randomizing matched pairs of groups may be offset by the loss of degrees of freedom caused by doing so. Thus, unless the predictive power of matching is substantial, it may reduce precision rather than increase it.

- **Mobility is the Achilles' heel of place-based programs and of group randomization experiments.**

The movement of individuals into and out of randomized groups tends to erode the connection between people and place. This erosion not only reduces the effectiveness of place-based programs by decreasing the target population's degree of exposure to them but can complicate the design, execution, and interpretation of evaluation findings about such programs.

It is hoped that, by highlighting and clarifying these and related analytic issues, the present chapter will facilitate more frequent and effective use of group randomization to evaluate — and thereby to improve — place-based social programs.

## Appendix

# Deriving the Minimum Detectable Effect Size in Group Randomization Experiments

This appendix to Chapter 4 derives expressions for the minimum detectable effect size in experiments using group randomization. Results for estimates of three types of program impacts are presented — net impacts for the full sample, net impacts for subgroups, and differential impacts for subgroups. Subgroups defined by the characteristics of the groups randomized and subgroups defined by the characteristics of individual sample members are considered.

## Results for the Full Experimental Sample

The net impact, $B_0$, of a program on an outcome is defined as the difference between the mean outcome in the presence of the program and the mean outcome in the absence of the program. In a group randomization design with one program group and one control group, the net impact is estimated as $b_0$, the difference between the mean outcomes for these two groups. Assume that $J$ groups of $n$ individuals each are randomly assigned with probability $P$ to the program group and with probability $1 - P$ to the control group. The between-group variance is $\tau^2$, the within-group variance is $\sigma^2$, and the intraclass correlation is $\rho$.

[Figure 4A.1 around here]

Figure 4A.1 illustrates why the minimum detectable effect of a program impact estimator is a multiple $M$ of its standard error (Bloom, 1995). The bell-shaped curve on the left represents the t distribution given that the true impact equals 0; this is the null hypothesis. For a positive impact estimate (presumed for present purposes to reflect a beneficial result) to be statistically significant at the $\alpha$ level for a one-tailed test (or at the $\alpha/2$ level for a two-tailed test), it must fall to the right of the critical t value, $t_{\alpha}$ (or $t_{\alpha/2}$), of this distribution. The bell-shaped curve on the right represents the t distribution given that the impact equals the minimum detectable effect; this is the alternative hypothesis. For the impact estimator to detect the minimum detectable effect with probability $1 - \beta$ (that is, to have a statistical power level of $1 - \beta$), the effect must lie a distance of $t_{1-\beta}$ to the right of the critical t value of the alternative hypothesis and a distance of $t_{\alpha} + t_{1-\beta}$ (or $t_{\alpha/2} + t_{1-\beta}$) from the null hypothesis. Because t values are expressed as multiples of the standard error of the impact estimator, the minimum detectable effect is also a multiple of the impact estimator. Thus:

$$M = t_{\alpha} + t_{1-\beta}, \text{ for a one-tailed test} \tag{A.1}$$

$$M = t_{\alpha/2} + t_{1-\beta}, \text{ for a two-tailed test} \tag{A.2}$$

The t values in these expressions reflect the number of degrees of freedom available for the impact estimator, which for the full sample equals the number of randomized groups minus 2 ($J - 2$). I therefore refer to the multiplier for the full sample as $M_{J-2}$ and to the standard error and minimum detectable effect for the full sample impact estimator given group randomization as $SE(b_0)_{GR}$ and $MDE(b_0)_{GR}$, respectively. The relationship among these terms is the following:

$$MDE(b_0)_{GR} = M_{J-2}SE(b_0)_{GR} \tag{A.3}$$

Because the discussion of precision in the chapter is expressed mainly in terms of the metric of effect size — defined as the program impact divided by the standard deviation of the outcome for the target population — this appendix focuses on the minimum detectable effect size *MDES(b₀)GR*. With group randomization, this standard deviation equals $\sqrt{\tau^2+\sigma^2}$. Hence, the minimum detectable effect size is defined as follows:

$$MDES(b_0)_{GR} = \frac{MDE(b_0)_{GR}}{\sqrt{\tau^2+\sigma^2}} \tag{A.4}$$

Equations A.3 and A.4 imply:

$$MDES(b_0)_{GR} = \frac{M_{J-2}SE(b_0)_{GR}}{\sqrt{\tau^2+\sigma^2}} \tag{A.5}$$

Recall Equation 2 (from the body of the chapter):

$$SE(b_0)_{GR} = \sqrt{\frac{1}{P(1-P)}}\sqrt{\frac{\tau^2}{J}+\frac{\sigma^2}{nJ}} \tag{A.6}$$

And note that the definition of intraclass correlation implies:

$$\tau^2 = \frac{\rho\sigma^2}{1-\rho} \tag{A.7}$$

By substituting Equations A.6 and A.7 into Equation A.5 and simplifying terms, one can express the minimum detectable effect size for the full sample net impact estimator thus:

$$MDES(b_0)_{GR} = \frac{M_{J-2}}{\sqrt{J}}\sqrt{\rho+\frac{1-\rho}{n}}\sqrt{\frac{1}{P(1-P)}} \tag{A.8}$$

## Results for Subgroups Defined by Characteristics of the Groups Randomized

For consistency with the example of subgroup analysis provided in the body of the chapter, consider two mutually exclusive and jointly exhaustive subgroups, A and B, that are defined by the characteristics of groups that were randomized. In an experiment where schools are randomly assigned, the subgroups might be urban schools and suburban schools; in an experiment where firms are randomly assigned, the subgroups might be retail firms and food service firms. Proportion $\Pi_A$ of the groups are in subgroup A, and proportion $1 - \Pi_A$ are in subgroup B. For simplicity, it is assumed of each subgroup that $P$, $n$, $\tau^2$, and $\sigma^2$ (and, by extension, $\rho$) are equal to their counterparts for the full sample.

The net impact for subgroup A, $B_{0A}$, is estimated as the difference between the mean outcome for subgroup members who were randomly assigned to the program group and the mean outcome for subgroup members who were randomly assigned to the control group and is denoted $b_{0A}$. Hence, the minimum detectable effect size for this subgroup can be obtained by substituting the number of randomized groups that it contains, $\Pi_A J$, and its multiplier, $M_{\Pi_A J-2}$, into Equation A.8. The ratio between this result and its counterpart for the full sample is:

$$\frac{MDES(b_{0A})_{GR}}{MDES(b_0)_{GR}} = \frac{M_{\Pi_A J-2}}{M_{J-2}}\sqrt{1/\Pi_A} \tag{A.9}$$

The corresponding ratio for subgroup B can be obtained by replacing $\Pi_A$ in Equation A.9 by $1-\Pi_A$.

The differential impact for the two subgroups, $B_{0A} - B_{0B},$ is estimated as the difference between their net impact estimates, $b_{0A} - b_{0B}$ . Because the differential impact reflects the mean outcome estimates for a total of four groups (the program and control groups in each subgroup), it has $J - 4$ degrees of freedom, and the minimum detectable effect multiplier is $M_{J\text{-}4}$. I now explain how to calculate the standard error for this impact estimator.

First, note that Equation A.6 implies that the variance of the full sample net impact estimator is:

$$VAR(b_0)_{GR} = [\frac{1}{JP(1-P)}][\tau^2 + \frac{\sigma^2}{n}] \tag{A.10}$$

Replacing $J$ in Equation A.10 by $\square_A J$ or by $(1 - \square_A)J$ to represent the number of randomized groups in subgroup A or B yields:

$$VAR(b_{0A})_{GR} = [\frac{1}{\Pi_A JP(1-P)}][\tau^2 + \frac{\sigma^2}{n}] \tag{A.11a}$$

$$VAR(b_{0B})_{GR} = [\frac{1}{(1-\Pi_A)JP(1-P)}][\tau^2 + \frac{\sigma^2}{n}] \tag{A.11b}$$

Note that because subgroups A and B are independent samples, the variance of the difference between their net impact estimates is the sum of their respective variances:

$$VAR(b_{0A}-b_{0B})_{GR}=VAR(b_{0A})_{GR}+VAR(b_{0B})_{GR} \qquad \text{(A.12)}$$

Substituting Equations A.11a and A.11b into Equation A.12 yields:

$$VAR(b_{0A}-b_{0B})_{GR}=[\frac{1}{\Pi_A JP(1-P)}][\tau^2+\frac{\sigma^2}{n}]+[\frac{1}{(1-\Pi_A)JP(1-P)}][\tau^2+\frac{\sigma^2}{n}]$$

$$=[\frac{1}{\Pi_A(1-\Pi_A)JP(1-P)}][\tau^2+\frac{\sigma^2}{n}] \qquad \text{(A.13)}$$

Finally, note:

$$\frac{MDES(b_{0A}-b_{0B})_{GR}}{MDES(b_0)_{GR}}=\frac{M_{J-4}}{M_{J-2}}\frac{SE(b_{0A}-b_{0B})_{GR}}{SE(b_0)_{GR}}\sqrt{\frac{\tau^2+\sigma^2}{\tau^2+\sigma^2}}$$

$$=\frac{M_{J-4}}{M_{J-2}}\frac{SE(b_{0A}-b_{0B})_{GR}}{SE(b_0)_{GR}} \qquad \text{(A.14)}$$

Replacing *SE(b₀)_GR* and *SE(b₀A-b₀B)_GR* in Equation A.14 by the square roots of Equations A.10 and A.11, respectively, and simplifying terms yields:

$$\frac{MDES(b_{0A}-b_{0B})_{GR}}{MDES(b_0)_{GR}}=\frac{M_{J-4}}{M_{J-2}}\sqrt{\frac{1}{\Pi_A(1-\Pi_A)}} \qquad \text{(A.15)}$$

## Results for Subgroups Defined by Individual Characteristics

Again for consistency with the body of the chapter, consider two mutually exclusive and jointly exhaustive subgroups, I and II, that are defined in terms of individual characteristics. In an experiment where schools are randomly assigned, the subgroups might be boys and girls; in an experiment where firms are randomly assigned, the subgroups might be long-term employees and recent hires. Assume that proportions $\Pi_I$ and $1 - \Pi_I$ of the individuals in each randomized group belong to subgroups I and II, respectively. Also, assume of each subgroup that $P$, $\tau^2$, and $\sigma^2$ (and, by extension, $\rho$) are the same as for the full sample.

The net impact for subgroup, $I(B_{0I})$, is estimated as the difference between the mean outcome for subgroup members who were randomly assigned to the program group and the mean outcome for subgroup members who were randomly assigned to the control group and is denoted $b_{0I}$. Because the subgroup contains sample members from all $J$ randomized groups, its net impact estimate has $J - 2$ degrees of freedom, and its minimum detectable effect multiplier is $M_{J-2}$. Replacing $n$ in Equation A.8 by $\Pi_I n$ and taking the ratio between this result and its counterpart for the full sample yields:

$$\frac{MDES(b_{0I})_{GR}}{MDES(b_0)_{GR}} = \frac{\sqrt{\rho + \frac{1-\rho}{\Pi_I n}}}{\sqrt{\rho + \frac{1-\rho}{n}}} \qquad \text{(A.16)}$$

Replacing $\Pi_I n$ in Equation A.16 by $(1 - \Pi_I)n$ produces the corresponding result for subgroup II.

The differential impact for the two subgroups, $B_{0I} - B_{0II}$, is estimated as the difference between their net impact estimates, $b_{0I} - b_{0II}$. Relative to the minimum detectable effect sizes for the net impact estimators, the minimum detectable effect size for this estimator is very small, implying high statistical precision. This is because the group-level error random, $e_j$, is "differenced away" when the differential impact is computed, which, in turn, eliminates $\tau^2$. To demonstrate this finding, note that the net impact estimator for each subgroup is the difference between the mean outcome for its members in the program group and the mean outcome for its members in the control group:

$$b_{0I} = \bar{\bar{Y}}_{PI} - \bar{\bar{Y}}_{CI} \qquad \text{(A.17a)}$$

$$b_{0II} = \bar{\bar{Y}}_{PII} - \bar{\bar{Y}}_{CII} \qquad \text{(A.17b)}$$

Thus, the differential impact can be expressed not only as a difference between program-control differences within the subgroups but as a difference between subgroup I-subgroup II differences within the program and control groups:

$$\begin{aligned} b_{0I} - b_{0II} &= (\bar{\bar{Y}}_{PI} - \bar{\bar{Y}}_{CI}) - (\bar{\bar{Y}}_{PII} - \bar{\bar{Y}}_{CII}) \\ &= (\bar{\bar{Y}}_{PI} - \bar{\bar{Y}}_{PII}) - (\bar{\bar{Y}}_{CI} - \bar{\bar{Y}}_{CII}) \qquad \text{(A.18)} \end{aligned}$$

For each randomized group j, the subgroup difference in mean outcomes is $\bar{Y}_{JI} - \bar{Y}_{jII}$, or $\Delta j$. The variance of this within-group subgroup difference for two independent subgroups is:

$$VAR(\bar{Y}_{jI} - \bar{Y}_{jII})_{GR} = VAR(\Delta j)_{GR} = \frac{\sigma^2}{\Pi_I(1-\Pi_I)n} \qquad \text{(A.19)}$$

Averaging *Δj* across the *PJ* randomized groups in the program or across the (1 – *P*)*J* groups in the control group yields the mean subgroup difference for the program group, $\bar{\bar{\Delta}}_P$, or for the control group, $\bar{\bar{\Delta}}_C$. The variances for these means are:

$$VAR(\bar{\bar{\Delta}}_P)_{GR} = \frac{\sigma^2}{PJ\Pi_I(1-\Pi_I)n} \qquad \text{(A.20a)}$$

$$VAR(\bar{\bar{\Delta}}_C)_{GR} = \frac{\sigma^2}{(1-P)J\Pi_I(1-\Pi_I)n} \qquad \text{(A.20b)}$$

Hence, the variance of the difference between the mean subgroup differences for the program and control groups is:

$$VAR(\bar{\bar{\Delta}}_P - \bar{\bar{\Delta}}_C)_{GR} = VAR(b_{0I} - b_{0II})_{GR} = \frac{\sigma^2}{PJ\Pi_I(1-\Pi_I)n} + \frac{\sigma^2}{(1-P)J\Pi_I(1-\Pi_I)n}$$

$$= \frac{\sigma^2}{P(1-P)J\Pi_I(1-\Pi_I)n} \qquad \text{(A.21)}$$

To state the variance of the full sample net impact estimator in comparable terms, substitute Equation A.7 into Equation A.10, and simplify as follows:

$$VAR(b_0)_{GR} = \frac{\frac{\rho\sigma^2}{1-\rho}+\frac{\sigma^2}{n}}{JP(1-P)} = \frac{\frac{n\rho\sigma^2+(1-\rho)\sigma^2}{n(1-\rho)}}{JP(1-P)} = \frac{\sigma^2(1+(n-1)\rho)}{JP(1-P)n(1-\rho)} \qquad \text{(A.22)}$$

Because the differential impact estimator is equivalent to the difference between the program and control groups with respect to their mean subgroup differences, it uses all $J$ groups and computes two means. Thus, it preserves all $J-2$ degrees of freedom from the full sample and has a minimum detectable effect multiplier of $M_{J-2}$. Consequently:

$$\frac{MDES(b_{0I}-b_{0II})_{GR}}{MDES(b_0)_{GR}} = \frac{M_{J-2}\sqrt{\frac{\sigma^2}{P(1-P)J\Pi_I(1-\Pi_I)n}}}{M_{J-2}\sqrt{\frac{\sigma^2(1+(n-1)\rho)}{JP(1-P)n(1-\rho)}}}$$

$$= \sqrt{\frac{(1-\rho)}{\Pi_I(1-\Pi_I)(1+(n-1)\rho)}} \qquad \text{(A.23)}$$

# References

Blalock, Hubert M., Jr. 1972. *Social Statistics*. New York: McGraw-Hill.

Blank, Rolf K., Diana Nunnaley, Andrew Porter, John Smithson, and Eric Osthoff. 2002. *Experimental Design to Measure Effects of Assisting Teachers in Using Data on Enacted Curriculum to Improve Effectiveness of Instruction in Mathematics and Science Education.* Washington, DC: National Science Foundation.

Bloom, Howard S. 1995. "Minimum Detectable Effects: A Simple Way to Report the Statistical Power of Experimental Designs." *Evaluation Review* 19(5): 547-556.

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): 445-469.

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

Bogatz, Gerry Ann, and Samuel Ball. 1971. *The Second Year of Sesame Street: A Continuing Evaluation.* Princeton: Educational Testing Service.

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

Branson, William H. 1972. *Macroeconomic Theory and Policy*. New York: Harper & Row.

Brooks-Gunn, Jeanne, Greg J. Duncan, and J. Lawrence Aber, editors. 1997. *Neighborhood Poverty: Context and Consequences for Children.* New York: Russell Sage Foundation.

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.

Campbell, M. K., J. M. Grimshaw, and I.N. Steen. 2000. "Sample Size Calculations for Group Randomised Trials." *Journal of Health Services and Policy Research* 5: 12-16.

Campbell, Marion K., Jill Mollison, and Jeremy M. Grimshaw. 2001. "Group Trials in Implementation Research: Estimation of Intragroup Correlation Coefficients and Sample Size." *Statistics in Medicine* 20: 391-399.

Chen, Huey-tsyh, and Peter H. Rossi. 1983. "Evaluating with Sense: The Theory-Driven Approach." *Evaluation Review* 7: 283-302.

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

Connell, James P., and Anne C. Kubisch. 1998. "Applying a Theory of Change Approach to the Evaluation of Comprehensive Community Initiatives: Progress, Prospects, and Problems." In *New Approaches to Evaluating Community Initiatives (Volume 2): Theory, Measurement, and Analysis*, edited by Karen Fulbright-Anderson, Anne C. Kubisch, and James P. Connell. Washington, DC: The Aspen Institute.

Cornfield, Jerome. 1978. "Randomization by Group: A Formal Analysis." *American Journal of Epidemiology* 108 (2): 100-102.

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).

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

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

Eccles, M. P., I. N. Steen, J. M. Grimshaw, L. Thomas, P. McNamee, J. Souter, J. Wilsdon, L. Matowe, G. Needham, F. Gilbert, and S. Bond. 2001. “Effect of Audit and Feedback and Reminder Messages on Primary-Care Referrals: A Randomized Trial.” *Lancet* 357: 1406-1409.

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

Fisher, Ronald A. 1926. “The Arrangement of Field Experiments.” *Journal of Agriculture* 33: 503-513.

Fisher, Ronald A. 1947/1937. *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): 861-885.

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: 1069-1092.

Galster, George C., Roberto G. Quercia, and Alvaro Cortes. 2000. "Identifying Neighborhood Thresholds: An Empirical Exploration." *Housing Policy Debate* 11(3): 701-732.

Garfinkel, Irwin, Charles F. Manski, and Charles Michalopoulos. 1992. "Micro Experiments and Macro Effects." In *Evaluating Welfare and Training Programs,* edited by Charles F. Manski and Irwin Garfinkel. Cambridge: Harvard University Press.

Gladwell, Malcolm. 2000. *The Tipping Point: How Little Things Can Make a Big Difference*. Boston: Little, Brown and Company.

Gosnell, Harold F. 1927. *Getting Out the Vote: An Experiment in the Stimulation of Voting*. Chicago: The University of Chicago Press.

Granovetter, Mark. 1973. "The Strength of Weak Ties." *American Journal of Sociology* 78(6): 1360-1380.

Granovetter, Mark. 1978. "Threshold Models of Collective Behavior." *American Journal of Sociology* 83(6): 1420-1443.

Grimshaw, J. M., R. E. Thomas, G. MacLennan, C. Fraser, C. Ramsay, L. Vale, P. Whitty, M. P. Eccles, L. Matowe, L. Shirran, M. Wensing, R. Disktra, C. Donaldson, and A. Hutchison. In press. "Effectiveness and Efficiency of Guideline Dissemination and Implementation Strategies." *Health Technology Assessment.*

Gueron, Judith M. 1984. *Lessons from a Job Guarantee: The Youth Incentive Entitlement Pilot Projects*. New York: MDRC.

Hacking, Ian. 1988. "Telepathy: Origins of Randomization in Experimental Design." *Isis* 79: 427-451.

Hedges, Larry V., and Ingram Olkin. 1985. *Statistical Methods for Meta-Analysis.* Boston: Academic Press.

Jencks, Christopher, and Susan E. Mayer. 1990. “The Social Consequences of Growing Up in a Poor Neighborhood.” In *Inner-City Poverty in the United States*, edited by Laurence E. Lynn, Jr., and M. McGeary. Washington, DC: National Academy Press.

Kellert, Stephen H. 1993. *In the Wake of Chaos: Unpredictable Order in Dynamical Systems.* Chicago: The University of Chicago Press.

Kelling, G., A. M. Pate, D. Dieckman, and C. Brown. 1974. *The Kansas City Preventative Patrol Experiment: Technical Report.* Washington, DC: The Police Foundation.

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

Kmenta, Jan. 1971. *Elements of Econometrics*. New York: Macmillan.

Ladd, Helen F., and Edward B. Fiske. 2000. *When Schools Compete: A Cautionary Tale.* Washington, DC: Brookings Institution Press.

Leviton, Laura C., R. L. Goldenberger, C. S. Baker, and M. Freda. 1999. “Randomized Controlled Trial of Methods to Encourage the Use of Antenatal Corticosteroid Therapy for Fetal Maturation.” *Journal of the American Medical Association* 281(1): 46-52.

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

Liu, Xiaofeng. 2002. “Statistical Power and Optimum Sample Allocation Ratio for Treatment and Control Having Unequal Costs per Unit of Randomization.” Working paper. University of South Carolina, Department of Psychology.

Martin, Donald C., Paula Diehr, Edward B. Perrin, and Thomas D. Koepsell. 1993. “The Effect of Matching on the Power of Randomized Community Intervention Studies.” *Statistics in Medicine* 12: 329-338.

Miller, Cynthia, and Howard Bloom. 2002. “Random Assignment in Cleveland — Round One.” Internal memorandum. New York: MDRC.

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

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

Murray, David M., Brenda L. Rooney, Peter J. Hannan, Arthur V. Peterson, Dennis V. Ary, Anthony Biglan, Gilbert J. Botvin, Richard I. Evans, Brian R. Flay, Robert Futterman, J. Greg Getz, Pat M. Marek, Mario Orlandi, MaryAnn Pentz, Cheryl L. Perry, and Steven P. Schinke. 1994b. “Intraclass Correlation Among Common Measures of Adolescent Smoking: Estimates, Correlates and Applications in Smoking Prevention Studies.” *American Journal of Epidemiology* 140(11): 1038-1050.

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): 681-694.

Musgrave, Richard A. 1959. *The Theory of Public Finance*. New York: McGraw-Hill.

Musgrave, Richard A., and Peggy B. Musgrave. 1973. *Public Finance in Theory and Practice.* New York: McGraw-Hill.

Myrdal, Gunnar. 1944. *An American Dilemma*. New York: Harper and Row.

Peirce, C. S., and Jastrow J. 1885. "On Small Differences of Sensation." *Memoirs of the National Academy of Sciences for 1884* 3: 75-83.

Peterson, Paul E., Patrick J. Wolf, William G. Howell, and David E. Campbell. 2002. "School Vouchers: Results from Randomized Experiments." Unpublished paper. Cambridge: Kennedy School of Government, Harvard University.

Raudenbush, Stephen W. 1997. "Statistical Analysis and Optimal Design for Group Randomized Trials." *Psychological Methods* 2(2): 173-185.

Schelling, Thomas. 1971. "Dynamic Models of Segregation." *Journal of Mathematical Sociology* 1: 143-186.

Shadish, William R., Thomas D. Cook, and Donald T. Campbell. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inferences.* Boston: Houghton Mifflin.

Sherman, Lawrence W., and David Weisburd. 1995. "General Deterrent Effects of Police Patrol in Crime 'Hot Spots': A Randomized Controlled Trial." *Justice Quarterly* 12(4): 625-648.

Siddiqui, Ohidul, Donald Hedeker, Brian R. Flay, and Frank B. Hu. 1996. "Intra-Class Correlation Estimates in School-based Smoking Prevention Study: Outcome and Mediating Variables by Gender and Ethnicity." *American Journal of Epidemiology* 144(4): 425-433.

Sikkema, Kathleen J., J. A. Kelly, R. A. Winett, L. 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. A. 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: 57-63.

Slavin, Robert E. 2002. "Evidence-Based Education Policies: Transforming Educational Practice and Research." *Educational Researcher* 31(7): 15-21.

Smith, Herbert L., Tu Ping, M. Giovanna Merli, and Mark Hereward. 1997. "Implementation of a Demographic and Contraceptive Surveillance System in Four Counties in North China." *Population Research and Policy Review* 16: 289-314.

"Student" (William Sealy Gosset). 1908. "The Probable Error of a Mean." *Biometrika* 6: 1-25.

Teruel, Graciela M., and B. 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.

Tienda, Marta. 1991. "Poor People and Poor Places: Deciphering Neighborhood Effects on Poverty Outcomes." In *Macro-Micro Linkages in Sociology,* edited by Joan Huber. Newbury Park, Calif.: Sage Publications.

Ukoumunne, O. C., M. C. Gulliford, S. Chinn, J. A. C. Sterne, and P. G. 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).

U.S. Department of Education. 2003. *Early Childhood Longitudinal Study, Kindergarten Class of 1998-99 (ECLS-K).* National Center for Education Statistics Web site. <http://nces.ed.gov/ecls>

Verma, Nandita. 2003. *Staying or Leaving: Lessons from Jobs-Plus About the Mobility of Public Housing Residents and the Implications for Place-Based Initiatives.* New York: MDRC.

Wilson, William Julius. 1996. *When Work Disappears: The World of the New Urban Poor.* New York: Alfred Knopf.

**Table 4.1**

## The Group Effect Multiplier

| Intraclass | Randomized group size (*n*) | | | | | |
|---|---|---|---|---|---|---|
| correlation ($\rho$) | 10 | 20 | 50 | 100 | 200 | 500 |
| 0.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| 0.01 | 1.04 | 1.09 | 1.22 | 1.41 | 1.73 | 2.48 |
| 0.02 | 1.09 | 1.17 | 1.41 | 1.73 | 2.23 | 3.31 |
| 0.03 | 1.13 | 1.25 | 1.57 | 1.99 | 2.64 | 4.00 |
| 0.04 | 1.17 | 1.33 | 1.72 | 2.23 | 2.99 | 4.58 |
| 0.05 | 1.20 | 1.40 | 1.86 | 2.44 | 3.31 | 5.09 |
| 0.06 | 1.24 | 1.46 | 1.98 | 2.63 | 3.60 | 5.56 |
| 0.07 | 1.28 | 1.53 | 2.10 | 2.82 | 3.86 | 5.99 |
| 0.08 | 1.31 | 1.59 | 2.22 | 2.99 | 4.11 | 6.40 |
| 0.09 | 1.35 | 1.65 | 2.33 | 3.15 | 4.35 | 6.78 |
| 0.10 | 1.38 | 1.70 | 2.43 | 3.30 | 4.57 | 7.13 |
| 0.20 | 1.67 | 2.19 | 3.29 | 4.56 | 6.39 | 10.04 |

Note: The group effect multiplier equals $\sqrt{1+(n-1)\rho}$ .

## Table 4.2

## The Minimum Detectable Effect Expressed as a Multiple of the Standard Error

| | Multiplier | |
|---|---|---|
| Number of groups (*J*) | Two-tailed test | One-tailed test |
| 4 | 5.36 | 3.98 |
| 6 | 3.72 | 3.07 |
| 8 | 3.35 | 2.85 |
| 10 | 3.20 | 2.75 |
| 12 | 3.11 | 2.69 |
| 14 | 3.05 | 2.66 |
| 16 | 3.01 | 2.63 |
| 18 | 2.99 | 2.61 |
| 20 | 2.96 | 2.60 |
| 30 | 2.90 | 2.56 |
| 40 | 2.87 | 2.54 |
| 60 | 2.85 | 2.52 |
| 120 | 2.83 | 2.50 |
| infinite | 2.80 | 2.49 |

Note: The group effect multipliers shown here are for the difference between the mean program group outcome and the mean control group outcome, assuming equal variances for the groups, a significance level of .05, and a power level of .80.

**Table 4.3**

**The Minimum Detectable Effect Size**

| | Randomized group size (*n*) | | | | | |
|---|---|---|---|---|---|---|
| Number of groups (*J*) | 10 | 20 | 50 | 100 | 200 | 500 |
| | **Intraclass correlation ($\rho$) = 0.01** | | | | | |
| 4 | 1.77 | 1.31 | 0.93 | 0.76 | 0.66 | 0.59 |
| 6 | 1.00 | 0.74 | 0.52 | 0.43 | 0.37 | 0.33 |
| 8 | 0.78 | 0.58 | 0.41 | 0.33 | 0.29 | 0.26 |
| 10 | 0.67 | 0.49 | 0.35 | 0.29 | 0.25 | 0.22 |
| 20 | 0.44 | 0.32 | 0.23 | 0.19 | 0.16 | 0.15 |
| 30 | 0.35 | 0.26 | 0.18 | 0.15 | 0.13 | 0.12 |
| 40 | 0.30 | 0.22 | 0.16 | 0.13 | 0.11 | 0.10 |
| 60 | 0.24 | 0.18 | 0.13 | 0.10 | 0.09 | 0.08 |
| 120 | 0.17 | 0.13 | 0.09 | 0.07 | 0.06 | 0.06 |
| | **Intraclass correlation ($\rho$) = 0.05** | | | | | |
| 4 | 2.04 | 1.67 | 1.41 | 1.31 | 1.26 | 1.22 |
| 6 | 1.16 | 0.95 | 0.80 | 0.74 | 0.71 | 0.69 |
| 8 | 0.90 | 0.74 | 0.62 | 0.58 | 0.55 | 0.54 |
| 10 | 0.77 | 0.63 | 0.53 | 0.49 | 0.47 | 0.46 |
| 20 | 0.50 | 0.41 | 0.35 | 0.32 | 0.31 | 0.30 |
| 30 | 0.40 | 0.33 | 0.28 | 0.26 | 0.25 | 0.24 |
| 40 | 0.35 | 0.28 | 0.24 | 0.22 | 0.21 | 0.21 |
| 60 | 0.28 | 0.23 | 0.19 | 0.18 | 0.17 | 0.17 |
| 120 | 0.20 | 0.16 | 0.14 | 0.13 | 0.12 | 0.12 |
| | **Intraclass correlation ($\rho$) = 0.10** | | | | | |
| 4 | 2.34 | 2.04 | 1.84 | 1.77 | 1.73 | 1.71 |
| 6 | 1.32 | 1.16 | 1.04 | 1.00 | 0.98 | 0.97 |
| 8 | 1.03 | 0.90 | 0.81 | 0.78 | 0.77 | 0.76 |
| 10 | 0.88 | 0.77 | 0.69 | 0.67 | 0.65 | 0.64 |
| 20 | 0.58 | 0.50 | 0.46 | 0.44 | 0.43 | 0.42 |
| 30 | 0.46 | 0.40 | 0.36 | 0.35 | 0.34 | 0.34 |
| 40 | 0.40 | 0.35 | 0.31 | 0.30 | 0.29 | 0.29 |
| 60 | 0.32 | 0.28 | 0.25 | 0.24 | 0.24 | 0.23 |
| 120 | 0.22 | 0.20 | 0.18 | 0.17 | 0.17 | 0.16 |

Note: The minimum detectable effect sizes shown here are for a two-tailed hypothesis test, assuming a significance level of .05, a power level of .80, and randomization of half the groups to the program.

## Table 4.4

## The Minimum Detectable Effect Size, by Sample Allocation

| Proportion allocated to<br>the program group (*P*) | Example 1 | Example 2 | Ratio to<br>balanced allocation |
|---|---|---|---|
| .10 | 0.91 | 0.29 | 1.67 |
| .20 | 0.68 | 0.22 | 1.25 |
| .30 | 0.59 | 0.19 | 1.09 |
| .40 | 0.55 | 0.18 | 1.02 |
| .50 (balanced) | 0.54 | 0.17 | 1.00 |
| .60 | 0.55 | 0.18 | 1.02 |
| .70 | 0.59 | 0.19 | 1.09 |
| .80 | 0.68 | 0.22 | 1.25 |
| .90 | 0.91 | 0.29 | 1.67 |

Notes: Example 1 is for $n = 20$, $J = 10$, $\rho = 0.05$, and a one-tailed hypothesis test. Example 2 is for $n = 80$, $J = 20$, $\rho = 0.01$, and a one-tailed hypothesis test. Both examples assume that the variances are the same for the program group and the control group.

**Table 4.5**

**Estimated School and Student Variances for Standardized Test Scores**

| | Reading | | Math | | |
|---|---|---|---|---|---|
| Type of covariate | Third grade | Sixth grade | Third grade | Sixth grade | Mean |
| No covariate | | | | | |
| School variance ($\tau^2$) | 19.7 | 12.9 | 18.0 | 21.5 | 18.0 |
| Student variance ($\sigma^2$) | 103.7 | 100.0 | 82.2 | 96.6 | 95.6 |
| School pretest | | | | | |
| School variance ($\tau^{*2}$) | 5.1 | 3.6 | 3.3 | 5.7 | 4.4 |
| Student variance ($\sigma^2$) | 105.5 | 100.5 | 83.2 | 97.4 | 96.7 |
| Student pretest | | | | | |
| School variance ($\tau^{*2}$) | 5.4 | 1.6 | 13.9 | 5.2 | 6.5 |
| Student variance ($\sigma^{*2}$) | 50.1 | 41.2 | 56.3 | 53.6 | 50.3 |

Notes: The results shown are based on individual standardized test scores for 3,299 third graders and 2,517 sixth graders in 25 elementary schools in Rochester, New York, in 1991 and 1992 (Bloom et al. 1999). The student pretest was each student's score in the same subject in the preceding grade. The school pretest was each school's mean score in the same subject and grade in the preceding year.

**Table 4.6**

### Minimum Detectable Effect Sizes for a Balanced Allocation of 60 Schools, Each with 60 Students per Grade

| | Reading | | Math | | |
|---|---|---|---|---|---|
| Type of covariate | Third grade | Sixth grade | Third grade | Sixth grade | Mean |
| No covariate | 0.27 | 0.23 | 0.28 | 0.29 | 0.27 |
| School pretest | 0.15 | 0.14 | 0.14 | 0.16 | 0.15 |
| Student pretest | 0.15 | 0.09 | 0.25 | 0.15 | 0.16 |

Notes: The results shown are based on individual standardized test scores for 3,299 third graders and 2,517 sixth graders in 25 elementary schools in Rochester, New York, in 1991 and 1992 (Bloom et al. 1999). The student pretest was each student's score in the same subject in the preceding grade. The school pretest was each school's mean score in the same subject and grade in the preceding year.

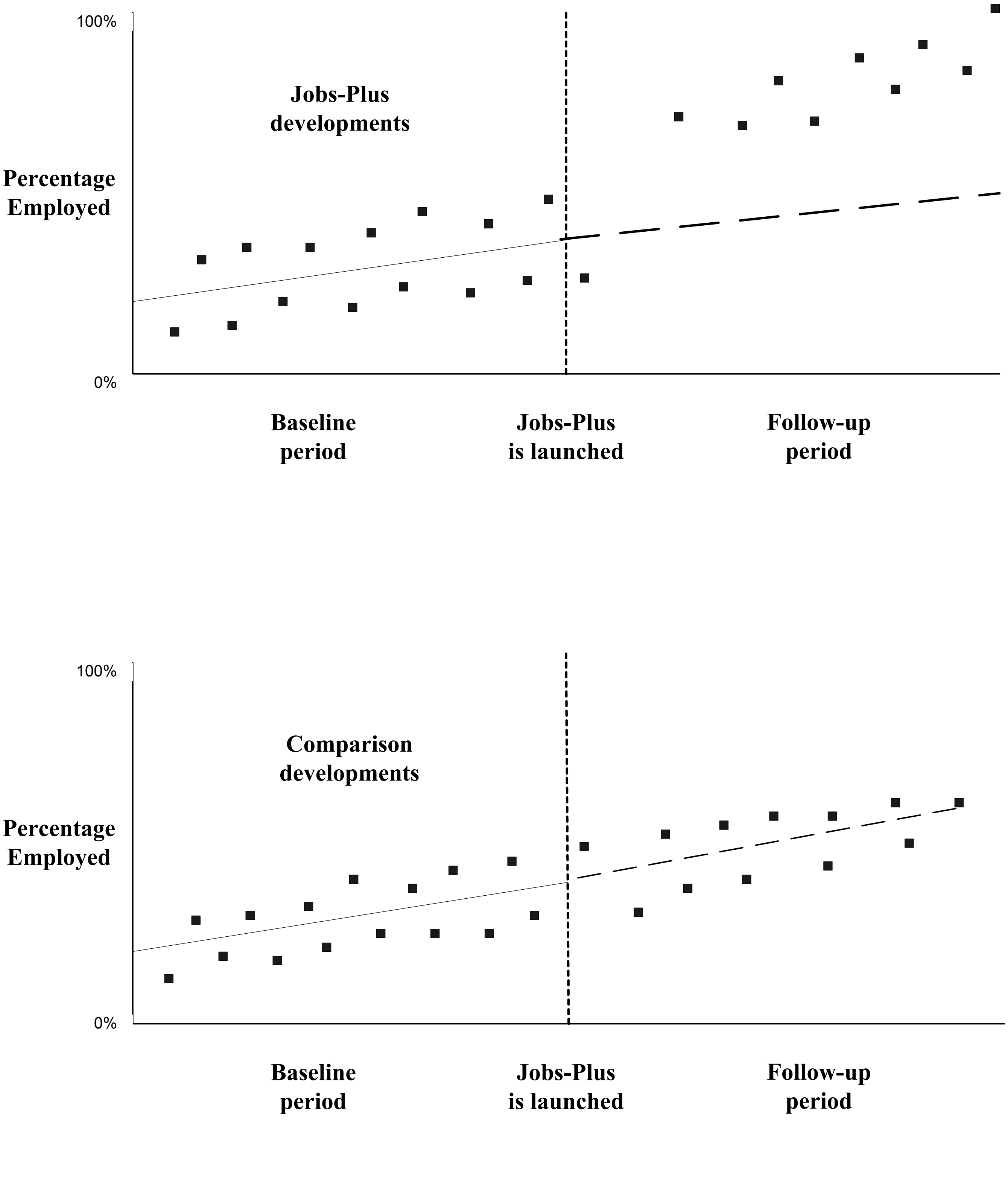
**Table 4.7**

**The Predictive Power Required to Justify Pair-wise Matching**

| Number of randomized groups (*J*) | Required predictive power (incremental $R^2$) | |
|---|---|---|
| | Two-tailed test | One-tailed test |
| 4 | 0.85 | 0.73 |
| 6 | 0.52 | 0.40 |
| 8 | 0.35 | 0.27 |
| 10 | 0.26 | 0.20 |
| 12 | 0.21 | 0.16 |
| 14 | 0.17 | 0.13 |
| 16 | 0.15 | 0.11 |
| 18 | 0.13 | 0.10 |
| 20 | 0.11 | 0.09 |
| 30 | 0.07 | 0.05 |
| 40 | 0.05 | 0.05 |
| 60 | 0.03 | 0.03 |
| 120 | 0.02 | 0.01 |
| infinite | 0.00 | 0.00 |

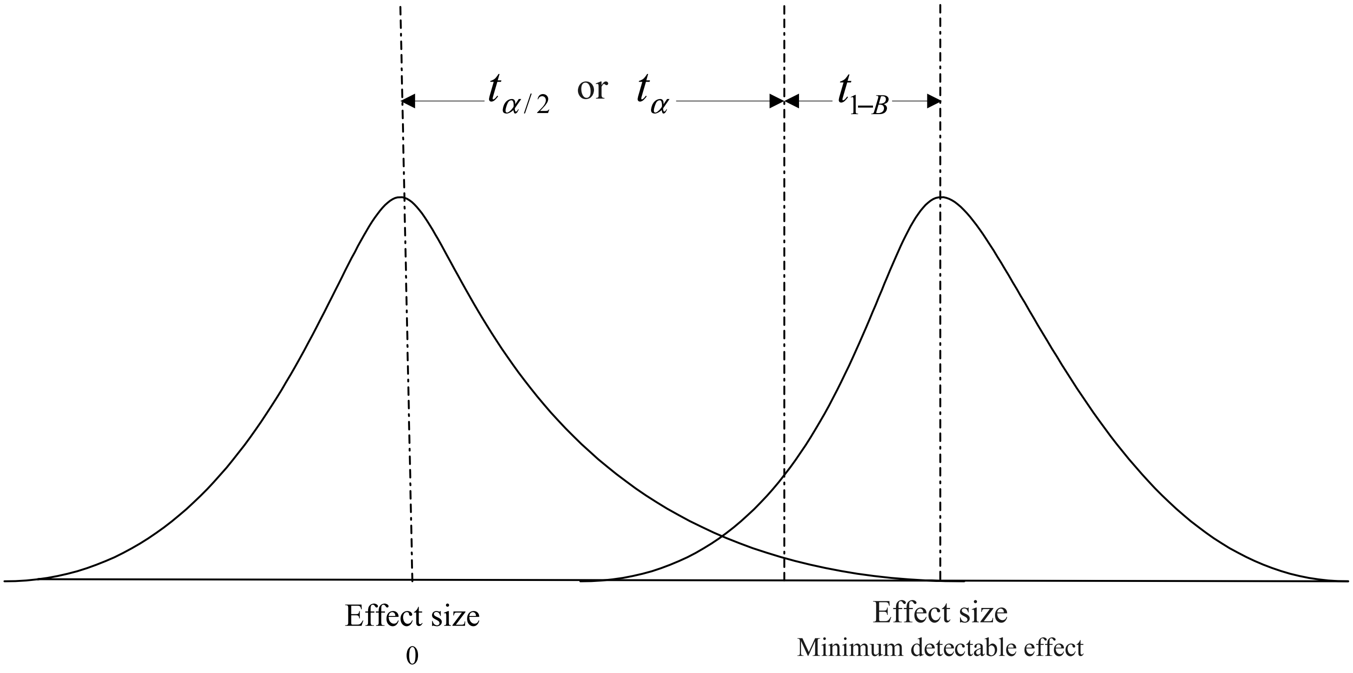
## Figure 4.1

## A Comparative Interrupted Time-Series Analysis of the Impacts of Jobs-Plus on Employment



## Figure 4A.1

## The Minimum Detectable Effect Multiplier



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

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