Chapter 4

Randomizing Groups to Evaluate Place-Based Programs

 Social interventions such as community improvement programs, school reforms, and employer-based efforts to retain workers, whose aim is to change whole communities or organizations, are often called place-based initiatives. Because such programs are designed to affect the behavior of groups of interrelated people rather than individuals, it is generally not feasible to measure their effectiveness in an experiment that randomly assigns individuals to the program or to a control group. By randomizing at the level of groups such as neighborhoods, schools, or companies—also called clusters—researchers can still reap most of the methodological benefits of random assignment.

Perhaps the earliest application of group, or cluster, randomization was Harold E. Gosnell’s (1927) study of ways to increase voter turnout. After dividing each of twelve 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 (1937/1947), the father of randomized experiments, published his landmark book on the use of randomization to study cause and effect.1 Not until about twenty years ago, in the early 1980s, did evaluators begin to use cluster randomization with any frequency. Since its application was confined mostly to research on health, it is no surprise that the only two textbooks on cluster randomization published to date, one by Allan Donner and Neil Khr (2000) and the other by David M. Murray (1998), focus on evaluating health programs.2

The use of cluster randomization to study the effects, or impacts, of social policies is now spreading to many fields (for a review and discussion of the key issues, 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, Hannan et al. 1994), school-based smoking, drinking, and sex prevention programs (Fly 2000), community employment initiatives (Bloom and Riccio 2002), police patrol innovations (Sherman and Weisburd 1995), family-planning programs (Smith et al. 1997), rural health, nutrition, and education 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 foster more frequent and better-informed use of cluster randomization, this chapter explores the rationale, nature, and consequences of place-based programs: the role, design, and implementation of cluster randomization evaluations of such programs; and the statistical properties and substantive implications of these evaluation designs. Many of the points raised in the chapter have been made by other authors in other settings. By putting these points together in new ways, considering them from new perspectives and contexts, providing new examples, and bringing to bear new empirical information, however, I attempt to advance the state of the art.

**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 cluster 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 between 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:

4. Using a place-based design will reduce political opposition to randomization.

5. 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.

**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 (often referred to as system effects) can reflect interdependencies between actors with respect to a single outcome, independencies between outcomes with respect to a single actor, or both. For example, the process of finding a job 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; 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, an evaluation of the program should account for them.

Spillover effects are recognized in many fields. 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 spillovers do exist, however, one must shift to a higher level of aggregation to accommodate them. 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 such spillover effects can occur.

Game-theory models seek to explain how one individual's response to a situation influences others' responses. 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' decisions with producers' 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 total 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 when 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 provides another tool for studying spillover effects.

Embedded within these models and theories are two key features of spillover effects, 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 in 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. The anecdotal evidence for this phenomenon is compelling (Gladwell 2000), but so far only limited statistical evidence for it is available because of methodological difficulties associated with gathering such evidence.

Spillover Effects Between Outcomes  In the second type of spillover, one outcome affects another. According to Gunnar 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 invoked this principle to account for the relationship between white prejudice against nonwhites and the economic circumstances of nonwhites. More recently, William Julius 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 subject of much theorization, little hard evidence about them exists. For example, the extensive literature examining features of neighborhoods and their effects on children remains inconclusive because of conceptual, measurement, and statistical problems (Tienda 1991; Jencks and Mayer 1990; Brooks-Gunn, Duncan, and Aber 1997; for a randomized experiment on this topic, see Kling, Ludwig, and Katz 2005).

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 un-
employment, 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 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 group of dispersed individuals is that place can be an effective platform for service delivery: 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 Spatial concentration of the target-group members benefits place-based initiatives in two major ways: through physical proximity to target-group members and by providing the opportunity to leverage 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 time and money costs of transportation 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 how to participate. Saturating local media with a program’s message may also stimulate word-of-mouth communication. 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 interventions to reduce lifestyle-related health-risk factors that are based on media outreach and information campaigns have been designed and tested as randomized place-based experiments (Murray, Rooney et al. 1994; 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 a primary or secondary school functions by changing the timing, staffing, style, culture, and curriculum of the entire school. It is much easier to evaluate such initiatives by randomly assigning schools to a program rather than individual students within a school to a program. Two examples of this approach are the completed evaluation of the School Development Program (Cook, Hunt, and Murphy 2000) and an ongoing evaluation of the reading program Success for All (Slavin 2002).

Employer-based initiatives aimed at reducing turnover among employees also are designed to change organizations. In attempting to improve procedures for training, supervising, and counseling employees, such programs focus on entire firms, not just on individuals. They may include providing 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 (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. These programs provide education activities, embed audit and feedback procedures in patient information systems, or stimulate other organizational changes to expedite diffusion of improved medical practices. Jeremy M. Grimshaw et al. (2004) presented 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 locus of intervention.
Nature of the Problem: When Locus Is the Focus  Place-based solutions make sense when the focus is on social problems with an uneven spatial distribution. For example, because most crimes are concentrated geographically, crime reduction strategies are often targeted at specific locations. Over 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 fifteen police beats to receive different patrol intensities and compared the subsequent crime rates across beats. 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 one hundred ten “hot spots,” 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 is one that 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, administration, and costs. Doing this requires full implementation of the program in selected locations.

For example, one of the most contentious issues in the debate about 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 as racial and economic segregation if they were implemented at full scale. Perhaps the best evidence bearing on this issue is Helen Ladd and Edward B. Fiske’s (2000) nonexperimental study of changes that occurred after New Zealand instituted a nationwide school choice program and Chang-Tai Hsieh and Miguel Urquiola’s (2003) nonexperimental study of the effects of Chile’s nationwide introduction of school choice by providing vouchers to any student wishing to attend a private school.

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). Whereas 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 by means of nonexperimental analyses of changes in local housing markets after implementation of housing allowance entitlement programs.

The Youth Incentive Entitlement Pilot Projects, an initiative that guaranteed jobs to all interested sixteen-to-eighteen-year-olds in seventeen locations across the United States from 1978 to 1980, likewise included a component to measure system effects (Gueron 1984). The project used a nonexperimental analysis to examine the program’s impacts on unemployment and school success among the 76,000 young people who volunteered to participate and also conducted a nonexperimental analysis of what the program’s impacts on the youth labor market would be if it were fully implemented in four of the cities in the study.

Although none of these full-scale tests involved randomizing places, it is possible to imagine doing so. The vast scale of such an extensive randomized study, however, means that the practical application of the method is probably limited to a small number of exceptionally important programs.

Facilitating Randomization

Another, very different, 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.” In reality, random assignment treats sample members equally in the sense that each 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. And it can capitalize on 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 difference between the group receiving the intervention, or treatment, and the control group. This difference is called the treatment contrast. Such contamination of the control group 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 antisomboking program, they most likely will share some of the information provided through the program with peers who have been randomly assigned to the control group. This second-hand exposure will attenuate the treatment contrast and make it difficult to interpret impact estimates.

One way to head off this problem is to separate the program group from the control group spatially, by means of place-based randomization. For example, randomly assigning homerooms instead of individual students to an antisomking program or to a control group can limit the extent to which program and control-group members share information. For some types of programs, however, even this degree of separation may not be adequate, and it might be better to randomize larger entities, such as schools (for an example, see Flay et al. 1985).

Similarly, in an experiment testing ways to induce physicians to use proven new medical procedures, randomization of individual physicians 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 yet to randomize hospitals (see Campbell, Mollison, and Grimshaw 2001).

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 (Boazt and Ball 1971). In the first year of the evaluation, each eligible household in five U.S. cities was randomly assigned either 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 had received the “treatment” being tested. Therefore, in the next phase of the study, which was conducted in two other cities, the program group and the control group were spatially separated. At that time Sesame Street was available only through cable in those areas, and the separation was accomplished in one of the additional 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, in the control group, did not receive free cable service. All the households were in low-income areas where cable television was prohibitively expensive at the time, so very few households in the control-group blocks were able to watch the program. Thus, place-based randomization greatly reduced the likelihood of control-group contamination.

Statistical Properties of Cluster Randomization

Impact estimates based on cluster random assignment, like those based on individual random assignment, are unbiased (for details, see Raudenbush 1997; Murray 1998; or Donner and Klar 2000). But estimates based on cluster randomization have much less—often much 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).

Model of Program Impacts

Consider a situation in which there are J clusters of n individual members each. Assuming that a proportion P of these clusters 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

\[ Y_{ij} = \alpha + \beta T_{ij} + \epsilon_{ij} \]  

(4.1)

where:

- \( Y_{ij} \) = the outcome for individual i from cluster j
- \( \alpha \) = the mean outcome for the control group
- \( \beta \) = the true program impact
- \( T_{ij} = 1 \) for program-group members and 0 for control-group members
- \( \epsilon_{ij} \) = the error component for individual i from cluster j

The true program impact, \( \beta \), 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, \( e_i \) for cluster differences and \( e_o \) for individual differences, that are assumed to have independent and identical distributions with means of 0 and variances of \( \tau^2 \) (for \( e_i \)) and \( \sigma^2 \) (for \( e_o \)). Various referred to as a multilevel, hierarchical, random coefficients, or mixed model. Equation 4.1 can be estimated by means of widely available software.

**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 cluster level, the standard error of the impact estimator is the following:

\[
SE(b_0)_{CL} = \frac{1}{\sqrt{P(1-P)}} \sqrt{\frac{\tau^2}{nJ} + \frac{\sigma^2}{n}}
\]

(4.2)

If, instead of randomizing the J clusters 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:

\[
SE(b_0)_{IN} = \frac{1}{\sqrt{P(1-P)}} \sqrt{\frac{\tau^2}{nj} + \frac{\sigma^2}{n}}
\]

(4.3)

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

The magnitude of the difference between \( SE(b_0)_{CL} \) and \( SE(b_0)_{IN} \) depends on the relationship between \( \tau^2 \) and \( \sigma^2 \) and the size of each cluster. The relationship between \( \tau^2 \) 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 clusters as opposed to within clusters:

\[
\rho = \frac{\tau^2}{\tau^2 + \sigma^2}
\]

(4.4)

Equations 4.2, 4.3, and 4.4 imply that the ratio between the standard error for cluster randomization and that for individual randomization, given a fixed total number of individual sample members, is a cluster effect multiplier, which can be expressed as:

\[
CEM = \sqrt{1 + (n-1)\rho}
\]

(4.5)

This cluster effect multiplier is the same as the well-known "design effect" in cluster sampling (Kish 1965).\(^{11}\)

Equation 4.5 indicates that, for a given total number of individuals, the standard error for cluster randomization increases with the size of the clusters, \( n \), and with the intraclass correlation, \( \rho \). Given that the intraclass correlation reflects the cluster effect (which is what inflates the standard error), it should not be surprising that the standard error increases with it. The cluster size comes into play here too because, for a given total number of individuals, larger clusters imply that there are fewer clusters to be randomized—and thus a larger margin of random error.

Table 4.1 illustrates these relationships. First, note that if the intraclass correlation is 0 (that is, if there is no cluster effect), the cluster effect multiplier is 1, and the standard errors for cluster randomization and individual randomization are the same. Next, note that large clusters (and thus few clusters to be randomized) imply relatively large standard errors, even when the intraclass correlation is small. For example, if \( \rho \) equals 0.01, randomizing J clusters of five hundred people each will produce standard errors 2.48 times as large as 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, Hannan et al. 1994) can produce large standard errors for program impact estimators because it is usually possible to randomize only a small number of such clusters.

Because the intraclass correlation captures the degree to which the outcome is stratified by cluster, its value varies with the type of outcome (for example, academic performance, employment, or health risks) and the type of cluster (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 clusters that represent small areas or organizational units (such as census tracts or classrooms) usually have larger intraclass correlations— in other words, are more homogeneous— than are larger clusters (such as municipalities or schools).

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

As table 4.1 underscores, the benefits of cluster randomization can
Table 4.1  The Cluster Effect Multiplier

<table>
<thead>
<tr>
<th>Intraclass Correlation (p)</th>
<th>10</th>
<th>20</th>
<th>50</th>
<th>100</th>
<th>200</th>
<th>500</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>0.01</td>
<td>1.04</td>
<td>1.09</td>
<td>1.22</td>
<td>1.41</td>
<td>1.73</td>
<td>2.48</td>
</tr>
<tr>
<td>0.02</td>
<td>1.09</td>
<td>1.17</td>
<td>1.41</td>
<td>1.73</td>
<td>2.23</td>
<td>3.31</td>
</tr>
<tr>
<td>0.03</td>
<td>1.13</td>
<td>1.25</td>
<td>1.57</td>
<td>1.99</td>
<td>2.64</td>
<td>4.00</td>
</tr>
<tr>
<td>0.04</td>
<td>1.17</td>
<td>1.33</td>
<td>1.72</td>
<td>2.23</td>
<td>2.99</td>
<td>4.58</td>
</tr>
<tr>
<td>0.05</td>
<td>1.20</td>
<td>1.40</td>
<td>1.86</td>
<td>2.44</td>
<td>3.31</td>
<td>5.09</td>
</tr>
<tr>
<td>0.06</td>
<td>1.24</td>
<td>1.46</td>
<td>1.98</td>
<td>2.63</td>
<td>3.60</td>
<td>5.56</td>
</tr>
<tr>
<td>0.07</td>
<td>1.28</td>
<td>1.53</td>
<td>2.10</td>
<td>2.82</td>
<td>3.86</td>
<td>5.99</td>
</tr>
<tr>
<td>0.08</td>
<td>1.31</td>
<td>1.59</td>
<td>2.22</td>
<td>2.99</td>
<td>4.11</td>
<td>6.40</td>
</tr>
<tr>
<td>0.09</td>
<td>1.35</td>
<td>1.65</td>
<td>2.33</td>
<td>3.15</td>
<td>4.35</td>
<td>6.78</td>
</tr>
<tr>
<td>0.10</td>
<td>1.38</td>
<td>1.70</td>
<td>2.43</td>
<td>3.30</td>
<td>4.57</td>
<td>7.13</td>
</tr>
<tr>
<td>0.20</td>
<td>1.67</td>
<td>2.19</td>
<td>3.29</td>
<td>4.56</td>
<td>6.39</td>
<td>10.04</td>
</tr>
</tbody>
</table>

Source: Computations by the author.
Note: The cluster effect multiplier equals \( \sqrt{1 + \frac{n}{1-p}} \).

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

Implications for Sample Size

Equation 4.2 indicates how five factors determine the standard errors of program impact estimators based on cluster 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. When designing a cluster randomization study, it is thus necessary to obtain information about \( \tau^2 \) and \( \sigma^2 \), or their relationship as expressed by \( \rho \), by consulting previous research on similar outcomes and groups or by estimating these parameters from existing data. The study by Howard S. Bloom, Johannes M. Bos, and Suk-Won Lee (1999) discussed later in this chapter illustrates how this can be done.

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 statistical 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). This parameter, which is a multiple of the impact estimator's standard error, depends on the following factors (see the chapter appendix for further discussion):

- 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
- The number of degrees of freedom of the test, which—assuming a two-group experimental design and no covariates—equals the number of clusters minus 2, or 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 clusters to be randomized, assuming use of the conventional statistical significance criterion of .05 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 pertinent to most studies based on individual randomization because they typically have many degrees of freedom. When small numbers of clusters are randomized and thus very few degrees of freedom are available, however, this pattern has important implications for research design.

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 cluster 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 clus-
Table 4.2 The Minimum Detectable Effect Expressed as a Multiple of the Standard Error

<table>
<thead>
<tr>
<th>Total Number of Clusters (J)</th>
<th>Multiplier Two-Tailed Test</th>
<th>Multiplier One-Tailed Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>5.36</td>
<td>3.98</td>
</tr>
<tr>
<td>6</td>
<td>3.72</td>
<td>3.07</td>
</tr>
<tr>
<td>8</td>
<td>3.35</td>
<td>2.85</td>
</tr>
<tr>
<td>10</td>
<td>3.19</td>
<td>2.75</td>
</tr>
<tr>
<td>12</td>
<td>3.11</td>
<td>2.69</td>
</tr>
<tr>
<td>14</td>
<td>3.05</td>
<td>2.65</td>
</tr>
<tr>
<td>16</td>
<td>3.01</td>
<td>2.63</td>
</tr>
<tr>
<td>18</td>
<td>2.98</td>
<td>2.61</td>
</tr>
<tr>
<td>20</td>
<td>2.96</td>
<td>2.60</td>
</tr>
<tr>
<td>30</td>
<td>2.90</td>
<td>2.56</td>
</tr>
<tr>
<td>40</td>
<td>2.88</td>
<td>2.54</td>
</tr>
<tr>
<td>60</td>
<td>2.85</td>
<td>2.52</td>
</tr>
<tr>
<td>120</td>
<td>2.82</td>
<td>2.50</td>
</tr>
<tr>
<td>Infinite</td>
<td>2.80</td>
<td>2.49</td>
</tr>
</tbody>
</table>

Source: Computations by the author.

Note: The cluster 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.

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 that of Jacob Cohen (1977/1988), who proposed that minimum detectable effect sizes of roughly 0.20, 0.50, and 0.80 be considered small, medium, and large, respectively. Mark Lipsey (1990) provided empirical support for this characterization by examining the actual distribution of 102 mean effect size estimates reported in 186 meta-analyses that together represented 6,700 studies with 800,000 sample members. Consistent with Cohen’s scheme, the bottom third of this distribution ranges from 0.00 to 0.32, the middle third ranges from 0.33 to 0.55, and the top third ranges from 0.56 to 1.20.

More recently, however, important research has suggested that, at least for education interventions (and perhaps for other types of interventions as well), much smaller effect sizes should be considered substantively important, and thus greater precision might be needed than is suggested by Cohen’s categories. Foremost among the findings motivating these new expectations are those from the Tennessee Class Size Experiment, which indicate that reducing elementary school classes from their standard size of 22 to 26 students to 13 to 17 students increases average student performance by about 0.1 to 0.2 standard deviation (Nye, Hedges, and Konstantopoulos 1999). This seminal study of a major education intervention suggests that even big changes in schools result in what by previous benchmarks would have been considered small effects.

Another important piece of related research is that by Thomas J. Kane (2004), who found that, on average nationwide, a full year of elementary school attendance increases students’ reading and math achievement by only 0.25 standard deviation. Thus, an education intervention that has a positive effect only half as large as this (0.125 standard deviation) seems still to qualify as a noteworthy success. Further reinforcing these findings are results published by the National Center for Educational Statistics (1997) that indicate that, on average nationwide, a year of high school increases reading achievement by about 0.17 standard deviation and math achievement by about 0.26 standard...
deviation. Again, the message is clear: program effects on student achievement of as little as 0.1 to 0.2 standard deviation might be highly policy-relevant.

This research serves to highlight the importance in studies of program effects, including studies based on a cluster randomization design, of careful analysis and thought about how much precision is needed to address the key questions.

**How Sample Size Affects Minimum Detectable Effects**

Now consider how the minimum detectable effect size for cluster randomization, \( MDE(b_{0.01}) \), varies with the number and size of the clusters randomized, given the intraclass correlation and the proportion of clusters randomly assigned to the program, \( P \). Equation 4.6, which is derived in this chapter’s appendix, represents this relationship as follows:

\[
MDE(b_{0.01}) = \frac{M_{0.01}}{\sqrt{J}} \sqrt{\frac{1 - \rho}{n}} \sqrt{\frac{1}{P(1-P)}},
\]  

(4.6)

where \( M_{0.01} \) is the minimum detectable effect multiplier in table 4.2.

The number of clusters randomized influences precision through \( M_{0.01} \); (which varies appreciably only for small numbers of clusters) 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 clusters randomized.

The size of the clusters randomized often makes far less difference to the precision of program impact estimators than does the number of clusters, especially given a moderate to high intraclass correlation. This is because the effect of cluster size is proportional to:

\[
\sqrt{\frac{1 - \rho}{n}}
\]

For example, if \( \rho \) were equal to 0.05, the values of this term for randomized clusters 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 clusters makes 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 top 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 top right corner of each panel (for n = 500 and J = 4). Moreover, increasing the number of clusters reduces the minimum detectable effect size. For n = 10 and ρ = 0.01, for example, increasing the number of clusters from four to twenty 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 clusters and the intraclass correlation.

Finally, for a given total number of sample members, increasing cluster size improves the precision of impact estimates by much less than does increasing the number of clusters. For example, for ρ = 0.01, the minimum detectable effect size for four groups of ten individuals each is 1.77. Whereas doubling the size of each cluster reduces this parameter to 1.31, doubling the number of clusters reduces it to 0.78. In fact, the size of the clusters often has almost no influence on precision. For example, for ρ = 0.05, increasing the size of each cluster from fifty to five hundred individuals reduces the minimum detectable effect size very little: and for ρ = 0.10, the reduction is negligible.

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

Implications for Sample Allocation

Sample allocation—the proportion of clusters randomized to the program rather than to the control group—affects the precision of program-impact estimators in a number of ways.

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—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.8 Generally overlooked, however, is the fact that precision erodes 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 reasonable approximation.8 This latitude enables researchers to capitalize on such opportunities to increase precision as the availability of public administrative records increases, and these 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 increases the number of individuals who can be given access to a program and thus lowers political resistance to the approach.

Decisions about sample allocation are more complicated under conditions of heteroscedasticity—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.9 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 τ² and σ² are the same for the program group as for the control group. Equation 4.6 indicates that when this is the case, the minimum detectable effect size is proportional to:

$$\sqrt{\frac{1}{n}}$$

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 hardly changes that given one approaches extreme imbalance. To see this, note that:

$$\sqrt{\frac{1}{(1-P)}}$$

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 clusters randomized, the size of the clusters, 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 ρ. 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
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.

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 $t^2$. As a result, statistical tests of the significance of the difference in $t^2$ tend to have little power. One might therefore opt to skip such tests and simply not assume that $t^2$ 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 $t^2$, 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.

The scarcity of degrees of freedom for estimating variances when homoscedasticity does not hold has received virtually no attention in the literature on randomized experiments, most likely because the vast majority of these experiments call for randomization of individuals rather than clusters. 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 cluster designs do not have this luxury. Furthermore, because little is known about how the impacts of programs vary across types of individuals and settings, it is not clear how problematic homoscedasticity is likely to be. At this point in the development of randomized cluster studies, it therefore seems prudent to use balanced sample allocations whenever possible. Studies with relatively large numbers of clusters (say, fifty 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

Now consider how to analyze a program's impacts for subgroups defined in terms of program characteristics, cluster characteristics, and in-

---

**Table 4.4 The Minimum Detectable Effect Size, by Sample Allocation**

<table>
<thead>
<tr>
<th>Proportion of Clusters Allocated to the Program (P)</th>
<th>Example 1</th>
<th>Example 2</th>
<th>Ratio to Balanced Allocation</th>
</tr>
</thead>
<tbody>
<tr>
<td>.10</td>
<td>0.91</td>
<td>0.29</td>
<td>1.67</td>
</tr>
<tr>
<td>.20</td>
<td>0.68</td>
<td>0.22</td>
<td>1.25</td>
</tr>
<tr>
<td>.30</td>
<td>0.59</td>
<td>0.19</td>
<td>1.09</td>
</tr>
<tr>
<td>.40</td>
<td>0.55</td>
<td>0.18</td>
<td>1.02</td>
</tr>
<tr>
<td>.50 (balanced)</td>
<td>0.54</td>
<td>0.17</td>
<td>1.00</td>
</tr>
<tr>
<td>.60</td>
<td>0.55</td>
<td>0.18</td>
<td>1.02</td>
</tr>
<tr>
<td>.70</td>
<td>0.59</td>
<td>0.19</td>
<td>1.09</td>
</tr>
<tr>
<td>.80</td>
<td>0.68</td>
<td>0.22</td>
<td>1.25</td>
</tr>
<tr>
<td>.90</td>
<td>0.91</td>
<td>0.29</td>
<td>1.67</td>
</tr>
</tbody>
</table>

*Source: Computations by the author.*  
*Notes: Example 1 is for n = 20, p = 0.05, and a one-tailed hypothesis test. Example 2 is for n = 80, p = 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.*

The table illustrates, the minimum detectable effect size changes little until P drops below .20 or exceeds .80.

### When the Variances Are Unequal

If a program creates impacts that vary across individuals or clusters, the individual or group variances can increase or decrease relative to those for control-group members. Howard S. Bloom et al. (2001) demonstrated this phenomenon in their evaluation of a whole-school reform called Accelerated Schools, and Anthony S. Bryk and Stephen W. Raudenbush (1988) demonstrated it in their reanalyses of two important education experiments. Consider how the phenomenon might arise in the context of education programs. Some programs may have larger-than-average effects on students who are weaker than average initially. If sufficiently pronounced, this tendency can reduce the individual error variance, $\sigma^2$, for members of the program group. The opposite result will 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 $t^2$ 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. 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 (see Gail et al. 1996 and Kmenta 1971).
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? Although often honored in the breach, it is proper research protocol to specify in advance the subgroups for which one will report program impact estimates. Doing so limits the extent to which such analyses become “data-mining” exercises that can generate spuriously significant subgroup differences.

**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 and could split the program group in two on the basis of this distinction. However, 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 multi-arm 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 clusters to be randomized, the approach is probably only feasible for studying small numbers of program variants.

**Subgroups Defined by Cluster Characteristics**

Subgroups defined by characteristics of the clusters 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 cluster 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 that 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 clusters in the full experiment, however, the precision of this type of subgroup analysis is substantially less than that of the full sample analysis. Precision is lost in two ways: the smaller samples of clusters used in subgroup analysis 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 \( \tau^2, \sigma^2, \) and \( P \) are the same for both subgroups and for the full sample. Proportion \( \Pi_A \) of the randomized clusters are in subgroup A and proportion \( 1 - \Pi_A \) are in subgroup B. The ratio between the minimum detectable effect size for subgroup A and that for the full sample is:

\[
\frac{\text{MDES}(b_{0A})_{CL}}{\text{MDES}(b_0)_{CL}} = \frac{M_{1A} - 2}{M_{12}} \sqrt{\frac{1}{\Pi_A}}
\quad (4.7)
\]

(See also the chapter appendix.) Equation 4.7 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. 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 twenty clusters that were randomized for an experiment. Hence, \( \Pi_A \) equals 0.5, and:

\[
\frac{\text{MDES}(b_{0A})_{CL}}{\text{MDES}(b_0)_{CL}} = \frac{M_8}{M_{10}} \sqrt{\frac{1}{0.5}} = \frac{3.35}{2.99} \sqrt{2} = 1.58
\]

In this case, the minimum detectable effect size for subgroup A is 1.58 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:

\[
\frac{\text{MDES}(b_{OA} - b_{OB})_{CL}}{\text{MDES}(b_0)_{CL}} = \frac{M_{J-4}}{M_{J2}} \sqrt{\frac{1}{\Pi_A(1 - \Pi_A)}}
\quad (4.8)
\]

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 \( J - 4, \) and an increase in the stan-
dard 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 clusters 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:

\[ \frac{\text{MDES}(b_{a} - b_{b})_{CL}}{\text{MDES}(b_{b})_{CL}} = \frac{M_{16}}{M_{10}} \sqrt{1/(0.5)(0.05)} = 3.01 \sqrt{1.98} \sqrt{4} = 2.01 \]

Subgroups Defined by Individual Characteristics

Subgroups defined by the characteristics of individual sample mem-
bers can also provide valid experimental impact estimates. Thus, even
if schools are the unit of randomization, one can measure program im-
acts 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 in this way are
entirely different from those already discussed. To see this, recall that
the size of clusters has much less influence on precision than does the
number of clusters and that, in some cases cluster 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 es-
timator for the differential impact on boys as opposed to girls actually
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,
1 and II, defined by an individual characteristic such as gender. Assu-
me that, in each randomized group, a proportion \( \Pi_l \) of the individu-
als are in subgroup I and proportion \( 1 - \Pi_l \) are in subgroup II. Also as-
sume 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 for a simple model
of subgroup differences the ratio between the minimum detectable ef-
fect size for subgroup I and that for the full sample is:

\[ \frac{\text{MDES}(b_{a} - b_{b})_{CL}}{\text{MDES}(b_{b})_{CL}} = \frac{\sqrt{\frac{\rho + 1 - \rho}{\Pi_l n}}}{\sqrt{\frac{\rho + 1 - \rho}{n}}} \]

Equation 4.9 illustrates how reducing the size of randomized clus-
ters by moving from the full sample to a subgroup defined by an in-
dividual characteristic can have little effect on precision. For example,
assuming one hundred individuals per cluster in the full sample and an
intraclass correlation of 0.05, the minimum detectable effect size for a
subgroup net impact when there are fifty individuals in the subgroup
per cluster is 1.08 times that for the full sample—a mere 8 percent
increase.\textsuperscript{22}

The precision for subgroup differential impact estimators can be
even greater. As this chapter’s appendix demonstrates, for a simplified
model of subgroup differences:

\[ \frac{\text{MDES}(b_{a} - b_{b})_{CL}}{\text{MDES}(b_{b})_{CL}} = \frac{\sqrt{(1 - \rho)}}{\Pi_l (1 - \Pi_l)(1 + (n - 1)\rho)} \]

Thus, with one hundred individuals per cluster in the full sample and
fifty individuals per cluster in each subgroup, the minimum de-
tectable effect size for the differential impact estimator is only 80 per-
cent as large as 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 cluster error component, \( \sigma^2 \), and thereby elimi-
nates \( \tau^2 \).

Adjusting for Covariates

Adjusting for covariates increases the precision of impact estimates by
reducing the amount of unexplained variation in the outcome of inter-
est. This approach is often used for experiments that randomize in-
dividuals. But its role can be even more important in experiments that
randomize clusters, where precision is more limited and therefore at a
higher premium. Furthermore, because the correlations among fea-
tures of aggregate entities are usually quite high (typically much
higher than the correlations among features of individuals), data on
cluster characteristics can substantially reduce the unexplained varia-
tion in the group error term—the binding constraint on precision in a
cluster design.
Aggregate Covariates, Individual Covariates, and Lagged Outcomes

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

Another important distinction is whether a covariate is a lagged outcome measure, 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, students’ reading test scores before being randomly assigned to the program or a control group would be a lagged outcome measure, whereas students’ gender 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 4.11 adds a single covariate, $X_{ij}$, to the program impact model in equation 4.1, yielding:

$$Y_{ij} = \alpha + B_0 T_{ij} + B_i X_{ij} + e_{ij} + e^*_{ij}$$

(4.11)

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 cluster characteristic. Furthermore, it can represent a lagged outcome measure or another type of background characteristic. Note that equation 4.11 assumes that the variance for the program group and the variance for the control group are equal.

The two error terms in equation 4.11—$e_{ij}$ for each cluster and $e^*_{ij}$ for each individual sample member differ from their counterparts in equation 4.1 because they represent the unexplained variation between and within clusters after controlling for the covariate, $X_{ij}$. Therefore, the random error terms in equation 4.11 are referred to as conditional errors, and those in equation 4.1 are referred to as unconditional errors.

Effects on Precision

Raudenbush (1997) derived expressions for the standard errors of impact estimators on the basis of cluster randomization given a balanced sample allocation, a single cluster covariate or a single individual covariate, and equal variances for the program and control groups. Equations 4.12 and 4.13 extend his findings to represent balanced or unbalanced allocations (with any value for $P$):

$$SE(b_0)_{\text{CL}} = \sqrt{\frac{1}{P(1-P)} \left[ \frac{\tau^2}{J} + \frac{\sigma^2}{nJ} \right] \left[ \frac{1}{1-\frac{1}{J-4}} \right]}$$

(4.12)

for a single cluster covariate and:

$$SE(b_0)_{\text{CL}} = \sqrt{\frac{1}{P(1-P)} \left[ \frac{\tau^2}{J} + \frac{\sigma^2}{nJ} \right] \left[ \frac{1}{1-\frac{1}{nJ-4}} \right]}$$

(4.13)

for a single individual covariate.

Comparing these expressions with equation 4.2, which assumes no covariate, reveals several important differences. First, consider:

$$\sqrt{1 + \frac{1}{J-4}}$$

in equation 4.12 and:

$$\sqrt{1 + \frac{1}{nJ-4}}$$

in equation 4.13, which have no counterparts in equation 4.2. The term in equation 4.12 (which is undefined for $J \leq 4$) approaches a value of one as the number of groups randomized increases. At ten groups, the term equals 1.08 and is therefore unimportant for larger samples. Similarly, the term in equation 4.13 (which is undefined for $nJ \leq 4$) approaches a value of one as the number of sample members increases. At ten groups of fifty individuals each, it equals approximately 1.00 and is therefore negligible for most sample sizes that are likely to be used.

More important are the differences between the conditional variances, $\tau^2$ and $\sigma^2$, in equations 4.12 and 4.13 and their unconditional counterparts, $\tau^2$ and $\sigma^2$, in equation 4.2.2 By controlling for some of the unexplained variation between clusters, a cluster characteristic can reduce the cluster variance from $\tau^2$ to $\tau^2$. By controlling for some of the unexplained variation both within and between clusters, an individual characteristic can reduce the cluster 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 significant amount, at a cost of only one degree of freedom per cluster characteristic and of 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 few clusters.

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 education program-impact estimates based on randomization of schools. Their analysis used the existing administrative records in twenty-five elementary schools in one urban school district, Rochester, New York, in 1991 and 1992. The authors estimated the between-school and within-school variance components for the standardized math scores and reading scores of third-graders and sixth-graders.

One type of covariate examined was a “student pretest” representing each student’s score in the same subject in the preceding grade; thus, individual second-graders’ and fifth-graders’ scores were used as a student-level pretest to compare with their performance as third- and sixth-graders, respectively. 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; thus, 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.

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 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 school pretests, although the pattern is not entirely consist-
Table 4.6 Minimum Detectable Effect Sizes for a Balanced Allocation of Sixty Schools, Each with Sixty Students per Grade

<table>
<thead>
<tr>
<th>Covariate</th>
<th>Third Grade</th>
<th>Sixth Grade</th>
<th>Third Grade</th>
<th>Sixth Grade</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>No covariate</td>
<td>0.27</td>
<td>0.23</td>
<td>0.28</td>
<td>0.29</td>
<td>0.27</td>
</tr>
<tr>
<td>School pretest</td>
<td>0.15</td>
<td>0.14</td>
<td>0.14</td>
<td>0.16</td>
<td>0.15</td>
</tr>
<tr>
<td>Student pretest</td>
<td>0.15</td>
<td>0.09</td>
<td>0.25</td>
<td>0.15</td>
<td>0.16</td>
</tr>
</tbody>
</table>

Source: Computations by the author using data from Bloom, Bos, and Lee (1999).
Notes: The results shown are based on individual standardized test scores for 3,299 third-graders and 2,517 sixth-graders in twenty-five elementary schools in Rochester, New York, in 1991 and 1992 (Bloom, Bos, and Lee 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.

and averages 0.16. In short, the power of both student and school pretests to improve statistical precision is considerable. Furthermore, school-level aggregate pretests have the advantage of being generally inexpensive to obtain as compared to student-level individual pretests.

One further important property of baseline covariates in a randomized experiment is that missing values for them can be imputed simply without biasing impact estimates. This is because randomization ensures that all baseline covariates are uncorrelated (in expectation) with treatment assignment, which means that their missing values are uncorrelated as well. Therefore, even a simple imputation that sets each missing value of a covariate to its mean for the full experimental sample would make it possible to keep all randomized observations in the impact analysis without creating bias. The only potential cost of not having data on a baseline covariate is a reduction in the precision of the impact estimate due to a reduction in the explanatory power of the covariate.

Blocking and Matching

Baseline covariates can also be used to block, or match, clusters before they are randomized, which means stratifying the clusters to be randomized into blocks defined by specific combinations of baseline characteristics. This is often done to reduce the potential for a “bad draw”—a situation in which the clusters randomly assigned to the program group differ substantially from those randomly assigned to the control group, thus confounding the treatment with other variables. Blocking thus increases the precision of program impact estimators by reducing their standard errors. However, this benefit comes at the cost of increasing the complexity of the impact analysis, which can increase the potential for errors. In addition, blocking reduces the number of degrees of freedom for the impact analysis, which can reduce precision.

**Blocking Before Randomization**

Each of the clusters in a block is randomly assigned to the program or to the control group. Ordinarily the sample allocation is held constant across blocks; this is called balanced allocation. In a blocked design, blocks are allocated constantly, ensuring that each block is represented in the same proportion in the program and control groups. This 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 control group. This procedure 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 does ensuring that the program and control groups have the same distributions of these characteristics 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, for both individuals and clusters. Thus, blocking with respect to a baseline measure of past outcomes is usually the best approach.

Given the small numbers of clusters 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, which though not necessary in principle is often necessary in practice. For example, the need for predictive validity may call for blocking on past outcomes but the need for face validity may call for blocking on demographic characteristics, even if the latter do not add much 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 every block but one and to add these variables to the basic impact model in equation 4.1, yielding:

\[
Y_{it} = \alpha + b_0 T_{it} + \Sigma \gamma_{ik} I_k + \epsilon_{it} + \epsilon_{it}'
\]  

(4.14)

Equation 4.14 specifies the coefficient \( \gamma_k \) for each block as a fixed effect. Doing so removes the outcome differences that exist across blocks from the impact analysis and thereby increases its precision. Statistical inferences (confidence intervals and hypothesis tests) about average impacts obtained from this model thus apply to the specific blocks in the research sample. This approach is most appropriate for samples that have few blocks (and therefore a limited potential for generalization), are chosen opportunistically or in an otherwise idiosyncratic way, or both. To make a broader statistical inference, one could model blocks as random effects, but this strategy adds a new error component to the impact analysis, thereby reducing its precision.

In short, researchers face a vexing trade-off between maintaining the precision of a relatively narrow but well-defined statistical inference (to the set of blocks in their sample) and sacrificing precision to permit a broader statistical inference (to a larger population of blocks). In this author’s opinion, for a small sample of blocks (say, fewer than ten) or an idiosyncratic sample of blocks (chosen in ways that make the sample representative of any readily defined larger population), it is best to use a fixed-effects model to draw statistical inferences and then, using one’s judgment, to attempt to broaden these inferences by describing the range of situations reflected by the current sample. In this way, the generalization is made heuristically rather than statistically, which is consistent with the size and nature of the sample of blocks being used.

The indicator variables in equation 4.14 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 countervailing forces on precision must be taken into account when the decision is made whether to block and how many blocks to use.

**Pairwise Matching Before Randomization**

Pairwise matching entails stratifying clusters into pairs before randomizing them: it is an extreme form of blocking. The best way to achieve predictive validity in matching is to rank the clusters from highest to lowest with respect to their values on the baseline characteristic to be used and, starting with the pair with the highest values, to randomly assign one member to the program and the other member to the control group. Alternatively, the ranking can 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 pairwise 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 reducing the standard error and increasing the minimum detectable effect multiplier. 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 ten clusters, there are eight degrees of freedom (\( 10 - 2 \)) without matching, but only four degrees of freedom \( 10/2 - 1 \) with matching.

Because limited resources often preclude randomizing more than a small number of clusters, the large loss of degrees of freedom produced by matching clusters can reduce the precision of program impact estimates—unless the predictive power of the matching is high (Martin et al., 1998). For example, consider the following expressions for the minimum detectable effect of an impact estimator given each of three different approaches: pairwise matching 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 clusters and a fixed number of individuals per cluster.

\[
MDE(b_{om_{CL}}) = M_{(7/2)-1}\sqrt{1-R^2} \cdot SE(b_{o_{CL}})
\]

(4.15)
given pair-wise matching, where \( R^2 \) is the predictive power;

\[
MDE(b_{om_{CL}}) = M_{(7/2)-1}\sqrt{1-R^2} \cdot \sqrt{1 + \frac{1}{J-4}} \cdot SE(b_{o_{CL}})
\]

(4.16)
given regression adjustment, where \( R^2 \) is the predictive power; and

\[
MDE(b_{o_{CL}}) = M_{(7/2)} \cdot SE(b_{o_{CL}})
\]

(4.17)
given no adjustment, where \( SE(b_{o_{CL}}) \) 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 \( M_{(7/2)-1} \) to \( M_{(7/2)-1} \) and reducing the standard error by a factor of \( \sqrt{1-R^2} \). One way to cap-
ture this trade-off is to compute the minimum predictive power of matching that would offset the increased minimum detectable effect multiplier and thereby increase precision. This expression can be obtained by setting the right side of equation 4.15 (the minimum detectable effect given matching) less than or equal to the right side of equation 4.17 (the minimum detectable effect given no adjustment) and rearranging terms, which results in:

$$R^2_i \geq 1 - \frac{M^2_{i/2}}{M^2_{i/2 - 1}} \quad (4.18)$$

Using this expression, table 4.7 presents the minimum required predictive power of pairwise 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. The most striking result is that very high predictive power ($R^2_{min}$) is required for pairwise matching to be justified, assuming a small sample of randomized groups. For example, for a two-tailed test or a one-tailed test, with six groups to be randomized, matching must predict 52 percent or 40 percent of the variation in the outcome measure, respectively, 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.

To consider the question of whether to match clusters for a given study more fully, it is also necessary to compare the likely precision of a design that matches clusters according to a baseline characteristic or a composite of such covariates with that of a design that randomizes unmatched clusters and controls for their characteristics through covariates in a regression model. Such an analysis can be done by working through the implications of equations 4.15 and 4.16.

**An Empirical Example: Randomizing Firms**

Consider how randomization of matched pairs of groups was used to evaluate Achieve, an employer-based program for reducing job turnover rates among low-wage workers in the health-care industry (Miller and Bloom 2002). Achieve offers employees a mix of direct services that include individual job counseling and group informational lunch sessions about 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, cluster randomization was performed at the level of firms. The first round of the program evaluation was implemented by twenty-two health-care firms in Cleveland that volunteered to participate in the study.

Participating firms were recruited in two waves that occurred roughly one month apart, with eight firms in the first wave and fourteen firms in the second wave. To maximize predictive validity, the firms in each wave were ranked according to their reported rates of employee turnover during the previous six months, with one firm in each pair randomized to the program and the other firm randomized to the control group. When it was discovered that the percentage of black employees in the program groups was much different from that for the control group, the original assignment was reversed for one pair in each wave (in opposite directions) to improve the face validity of the evaluation design. (The researchers later realized that such ad hoc adjustments to a random draw can inadvertently bias an experimental design and that a better way to trade off predictive validity against face validity in this situation would have been to randomize the entire matched sample again.)

Cynthia Miller and Howard Bloom (2002) analyzed the relative precision of three research designs—matching firms based on their past turnover rates, linear regression adjustment for past turnover rates, and taking no account of past turnover rates—using data on employee turnover rates during the first month after random assignment (a short-
term-outcome measure). As it turned out, the predictive power of matching was sufficient to warrant using that approach, thereby providing a post hoc justification for having done so.

Accounting for Mobility

An inescapable fact 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. 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 computations by the present author using data from the U.S. Department of Education (2003) indicate that only half of American kindergarteners are still at the same school when they reach the third grade.

Issues Raised by Mobility

Mobility creates important programmatic and evaluation problems both for programs targeted to individuals and for programs targeted to groups or places. From a programmatic standpoint, the main problem is that mobility reduces enrollees' exposure to the treatment because many leave before receiving a full "dose" of the intervention being tested. High rates of mobility can thus undermine a program's chance to make a meaningful difference in the outcomes of interest.

From an evaluation standpoint, mobility creates two main problems. A general problem is that selective mobility can result in subsequent differences between characteristics of the program group and characteristics of the control group, thereby creating selection bias in program-impact estimates. This problem is often referred to as sample "attrition" (for example, see Rossi, Lipsey, and Freeman 2004, 270).

The second problem created by mobility applies only to place-based programs. Two conceptually and substantively different perspectives from which program impacts can be defined and measured need to be clearly distinguished: impacts on people and impacts on places.

Impacts on people reflect how a program changes outcomes for individuals to whom the program is targeted. For example, one might ask whether an education program increases reading achievement levels for students who are exposed to it. Though seemingly precise, this question needs further specification to be meaningful in a context where student mobility is high. In a high-mobility situation, one might ask whether the program raises reading achievement levels for all students who are present in a school when the program is launched (regardless of whether they move away subsequently) or whether it does so for students who remain at the 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 in different contexts.

In a cluster randomization study, the most effective way to measure the impacts of programs on people is a longitudinal methodology, one in which outcomes for the same individuals are tracked over time. A longitudinal design to measure the general effects of an education 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 of 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," students in the program and control groups who remain in the school throughout the analysis period, so as to reduce variation in exposure to the program. This strategy can produce valid experimental estimates of the impacts of the program assuming that the program does not influence mobility. But if the program affects the types of students who stay in the schools (for example, by improving schools to the extent that families who care more about education are more likely to keep their children in the school than would have been the case otherwise), then stayers in the program schools will differ from stayers in the control schools. Such differences can introduce selection bias into estimates of the program's impacts on student outcomes.

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 education program increases reading achievement levels for schools that are exposed to it. The answer may reflect a mixture of two different forces. The first is the effect of the program on the achievement of students who would remain in the same place with or without the program; the second is the effect of the program on student mobility. For example, an education 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 by both.

Impacts on places are best measured in a cluster randomization study using a "repeated cross-section" methodology, in which outcomes are tracked over time for the same places rather than for the same individuals. 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. To the extent that the program influences student mobility, however, it is not clear how to interpret the resulting impact estimates in the absence of further information.

An Empirical Example: Randomizing Housing Developments

The evaluation of Jobs-Plus mentioned earlier is based on randomization of matched pairs of public housing developments in six U.S. cities and 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; for a detailed discussion of interrupted time-series analysis, see Shadish, Cook, and Campbell 2002). 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 elements are intended to create unusually large employment gains that generate spillover effects throughout each participating development.

The Jobs-Plus evaluation assesses impacts on both public housing residents and public housing developments because moves into and out of public housing developments are frequent. The individual-based portion evaluates the impact of Jobs-Plus on the people who were living in the participating developments at a specific point in time, asking: How did Jobs-Plus affect the future experiences of its target population, whether or not they moved away? The analysis focusing on the housing development asks: 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 used to estimate the impacts of Jobs-Plus from both perspectives. The graph at the top of the figure illustrates a hypothetical pattern of average quarterly earnings 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 earnings, the quarterly levels 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 develop-
ment with those in the control-group development with which it was matched before random assignment. The impact of Jobs-Plus on average earnings is estimated as the difference between the two sets of deviations.

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 earnings levels backward and forward in time, regardless of where they lived before and after the point of the program's launch. To make the analysis from the housing development perspective possible, one must identify the people who resided in the developments during each quarter of the baseline period and after the program's launch and then calculate average earnings in each quarter for the persons who were residents of that quarter. Outcome data for the analyses from both the individual and development perspectives were obtained from public administrative records (Bloom and Riccio 2002).

Summary

This chapter lays out a research strategy that leverages the widely accepted scientific principle of randomization to permit evaluation of 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. In place-based programs, it is usually infeasible to randomize individual members of the groups, but it is often possible to randomize the groups or clusters themselves. Because cluster randomization is being used with increasing frequency to measure the impacts of social programs, it is particularly important for researchers to understand its special, and sometimes counterintuitive, properties. The key features of cluster randomization are the following:

*Precision is at a premium.* Cluster randomization provides estimates of program impacts that are unbiased for the same reason that individual randomization is unbiased. But impact estimates based on randomization of clusters almost always have much less precision than do their counterparts given randomization of the same total number of individuals.

*The number of clusters randomized is usually a much more important determinant of precision than is cluster size.* In most contexts, resources allow for randomization of only a small number of clusters, putting a strong constraint on the precision of program-impact estimates. Consequently, increasing the number of clusters by a given proportion usually improves precision by a much greater amount than does increasing the number of individuals per cluster 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 clusters randomized and when it is an individual characteristic of the cluster members.

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

*To improve precision, the characteristics used in pairwise matching must have considerable predictive power.* For an evaluation with a small number of clusters (say, fewer than ten) to be randomized, the gains in precision produced by randomizing matched pairs of clusters 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 cluster randomization experiments.* The movement of individuals into and out of randomized clusters 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 of such programs.

This chapter has explored in detail the use of cluster-randomized experiments for measuring the impacts of place-based programs. The goal of the chapter was to make clear to readers why the approach has great potential value, when the approach is most appropriate to use, and how to design studies that get the most information possible from it. Given this information it is hoped that future researchers will make more frequent and effective use of cluster randomization to advance the state of the art of evaluation research and thereby improve place-based programs for people.
Appendix: Deriving Expressions for the Minimum Detectable Effect Size for Randomized Cluster Experiments

This appendix derives expressions for the minimum detectable effect size in experiments using cluster 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 clusters 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 cluster 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 clusters of n individuals each are randomly assigned with probability P to the program group and with probability 1 - P to the control group. For both the program group and the control group, the between-cluster variance is \( \tau^2 \), the within-cluster variance is \( \sigma^2 \), and the intraclass correlation is \( \rho \).

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, \( t_{\alpha/2} \) (or \( t_{\alpha} \)), 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, for a one-tailed test,

\[ M = t_{\alpha} + t_{1-\beta}. \]  

(4A.1)

and for a two-tailed test:

\[ M = t_{\alpha/2} + t_{1-\beta}. \]  

(4A.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 clusters minus two (J - 2). The multiplier for the full sample is thus referred to as \( M_{J-2} \). The standard error and minimum detectable effect for the full sample-impact estimator given cluster randomization are referred to as \( SE(b_0)_{CL} \) and \( MDE(b_0)_{CL} \), respectively. The relationship among these terms is the following:

\[ MDE(b_0)_{CL} = M_{J-2}SE(b_0)_{CL}. \]  

(4A.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, \( MDE(b_0)_{CL} \). With cluster randomization, the standard deviation of the outcome for the target population equals \( \sqrt{\tau^2 + \sigma^2} \). Hence, the minimum detectable effect size is defined as follows:

---

Figure 4A.1 The Minimum Detectable Effect Multiplier
MDES(b_0)_{CL} = \frac{MDE(b_0)_{CL}}{\sqrt{\tau^2 + \sigma^2}} \quad (4A.4)

Equations 4A.3 and 4A.4 imply:

\[ MDES(b_0)_{CL} = \frac{M_{1-2}SE(b_0)_{CL}}{\sqrt{\tau^2 + \sigma^2}} \quad (4A.5) \]

Recall equation 4.2:

\[ SE(b_0)_{CL} = \sqrt{\frac{1}{J} \left( \frac{\tau^2}{1} + \frac{\sigma^2}{n} \right)} \quad (4A.6) \]

Note that the definition of intraclass correlation implies:

\[ \tau^2 = \frac{\rho \sigma^2}{1 - \rho} \quad (4A.7) \]

By substituting equation 4A.6 for SE(b_0)_{CL} and equation 4A.7 for \( \tau^2 \) into equation 4A.5 and simplifying terms, one can express the minimum detectable effect size for the full-sample net impact estimator thus:

\[ MDES(b_0)_{CL} = \frac{M_{1-2}}{\sqrt{J}} \sqrt{\frac{1}{n} \left( \frac{1 - \rho}{\rho} + \frac{1}{J} \right)} \quad (4A.8) \]

### Results for Subgroups Defined by Cluster Characteristics

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 clusters. 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 clusters are in subgroup A, and proportion \( 1 - \Pi_A \) are in subgroup B. For simplicity, it is assumed of each subgroup that \( J, n, \tau^2, \) 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 clusters that it contains, \( \Pi_A \), for \( J \), and its multiplier, \( M_{1-2} \) for \( M_{1-2} \), into equation 4A.8. The ratio between this result and its counterpart for the full sample is:

\[ \frac{MDES(b_{0A})_{CL}}{MDES(b_{0})_{CL}} = \frac{M_{1-2}}{M_{1-2}} \sqrt{\frac{\Pi_A}{1 - \Pi_A}} \quad (4A.9) \]

The corresponding ratio for subgroup B can be obtained by replacing \( \Pi_A \) in equation 4A.9 with \( 1 - \Pi_A \).

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 \( M_{1-4} \).

To calculate the standard error for this impact estimator, first note that equation 4A.6 implies that the variance of the full sample net impact estimator is:

\[ \text{VAR}(b_0)_{CL} = \frac{1}{J \rho (1 - \rho)} \left[ \tau^2 + \frac{\sigma^2}{n} \right] \quad (4A.10) \]

Replacing \( J \) in equation 4A.10 with \( \Pi_A J \) or with \( (1 - \Pi_A)J \) to represent the number of clusters in subgroup A or B yields:

\[ \text{VAR}(b_{0A})_{CL} = \frac{1}{\Pi_A J \rho (1 - \rho)} \left[ \tau^2 + \frac{\sigma^2}{n} \right] \quad (4A.11a) \]

\[ \text{VAR}(b_{0B})_{CL} = \frac{1}{(1 - \Pi_A) J \rho (1 - \rho)} \left[ \tau^2 + \frac{\sigma^2}{n} \right] \quad (4A.11b) \]

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:

\[ \text{VAR}(b_{0A} - b_{0B})_{CL} = \text{VAR}(b_{0A})_{CL} + \text{VAR}(b_{0B})_{CL} \quad (4A.12) \]

Substituting equation 4A.11a for \( \text{VAR}(b_{0A})_{CL} \) and equation 4A.11b for \( \text{VAR}(b_{0B})_{CL} \) into equation 4A.12 yields:

\[ \text{VAR}(b_{0A} - b_{0B})_{CL} = \frac{1}{\Pi_A J \rho (1 - \rho)} \left[ \tau^2 + \frac{\sigma^2}{n} \right] + \frac{1}{(1 - \Pi_A) J \rho (1 - \rho)} \left[ \tau^2 + \frac{\sigma^2}{n} \right] \]

\[ = \frac{1}{\Pi_A (1 - \Pi_A) J \rho (1 - \rho)} \left[ \tau^2 + \frac{\sigma^2}{n} \right] \quad (4A.13) \]
Finally, note:

\[
\frac{\text{MDES}(b_{0A} - b_{0I})_{\text{CL}}}{\text{MDES}(b_{0})_{\text{CL}}} = \frac{M_{1-2} \cdot \text{SE}(b_{0A} - b_{0I})_{\text{CL}}}{M_{1-2} \cdot \text{SE}(b_{0})_{\text{CL}}} \frac{\tau^2 + \sigma^2}{\tau^2 + \sigma^2} = \frac{M_{1-2}}{M_{1-2}} \left( \frac{1}{\Pi_A (1 - \Pi_A)} \right) \tag{4A.14}
\]

Replacing \(\text{SE}(b_{0})_{\text{CL}}\) and \(\text{SE}(b_{0A} - b_{0I})_{\text{CL}}\) in equation 4A.14 by the square roots of equations 4A.10 and 4A.11, respectively, and simplifying terms yields:

\[
\frac{\text{MDES}(b_{0A} - b_{0I})_{\text{CL}}}{\text{MDES}(b_{0})_{\text{CL}}} = \frac{M_{1-2}}{M_{1-2}} \left( \frac{1}{\Pi_A (1 - \Pi_A)} \right) \tag{4A.15}
\]

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 \(\tau^2\), 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 \(J\) clusters, its net impact estimate has \(J - 2\) degrees of freedom, and its minimum detectable effect multiplier is \(M_{1-2}\). Replacing \(n\) in equation 4A.8 by \(n\Pi_I\) and taking the ratio between this result and its counterpart for the full sample yields:

\[
\frac{\text{MDES}(b_{0I})_{\text{CL}}}{\text{MDES}(b_{0})_{\text{CL}}} = \frac{\sqrt{\rho + \frac{1 - \rho}{n\Pi_I}}}{\sqrt{\rho + \frac{1 - \rho}{n}}} \tag{4A.16}
\]

Replacing \(n\Pi_I\) in equation 4A.16 with \(n(1 - \Pi_I)\) 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 can be somewhat smaller. This is because in a simple model of subgroup differences, the cluster-level random error, \(\epsilon\), is "difference 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:

\[
b_{0I} = \bar{y}_{PI} - \bar{y}_{CI} \tag{4A.17a}
\]

\[
b_{0II} = \bar{y}_{PII} - \bar{y}_{CII} \tag{4A.17b}
\]

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 and subgroup II within the program and control groups:

\[
u_{0I} - \nu_{0II} = (\bar{y}_{PI} - \bar{y}_{CI}) - (\bar{y}_{PII} - \bar{y}_{CII}) = (\bar{y}_{PI} - \bar{y}_{PII}) - (\bar{y}_{CI} - \bar{y}_{CII}) \tag{4A.18}
\]

For each cluster \(j\), the subgroup difference in mean outcomes is \(\bar{y}_{PI} - \bar{y}_{PII}\), or \(\Delta_j\). The variance of this within-cluster subgroup difference for two independent subgroups is:

\[
\text{VAR}(\bar{y}_{PI} - \bar{y}_{PII})_{\text{CL}} = \text{VAR}(\Delta_j)_{\text{CL}} = \frac{\sigma^2}{\Pi_I(1 - \Pi_I)n} \tag{4A.19}
\]

Averaging \(\Delta_j\) across the \(PJ\) clusters in the program or across the \((1 - P)J\) clusters in the control group yields the mean subgroup difference for the program group, \(\bar{\Delta}_P\), or for the control group, \(\bar{\Delta}_C\). The variances for these means are:

\[
\text{VAR}(\bar{\Delta}_P)_{\text{CL}} = \frac{\sigma^2}{PJ\Pi_I(1 - \Pi_I)n} \tag{4A.20a}
\]

and

\[
\text{VAR}(\bar{\Delta}_C)_{\text{CL}} = \frac{\sigma^2}{(1 - P)\Pi_I(1 - \Pi_I)n} \tag{4A.20b}
\]

Hence, the variance of the difference between the mean subgroup differences for the program and control groups is:
\[
\text{VAR}(\Delta_{\rho} - \Delta_{CL}) = \text{VAR}(b_{\rho} - b_{CL}) = \frac{\sigma^2}{NJ(1 - \Pi_1)n} + \frac{\sigma^2}{(1 - P)NJ(1 - \Pi_1)n}
\]
\[
= \frac{(1 - P)NJ(1 - \Pi_1)n}{\Pi_1(1 - \Pi_1)n}
\]
\[(4A.21)\]

To state the variance of the full-sample net impact estimator in comparable terms, substitute equation 4A.7 for \(\tau^2\) into equation 4A.10, and simplify as follows:

\[
\text{VAR}(b_{\rho}) = \frac{1 - \rho}{JP(1 - P)} = \frac{n(1 - \rho)}{JP(1 - P)n(1 - \rho)} = \frac{\sigma^2(1 + (n - 1)\rho)}{JP(1 - P)n(1 - P)n(1 - \rho)}
\]
\[(4A.22)\]

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 \(J\) clusters 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 \(M_{I-2}\). Consequently:

\[
\text{MDES}(b_{\rho} - b_{CL}) = \frac{\sigma^2}{M_{I-2} \sqrt{JP(1 - P)NJ(1 - \Pi_1)n}}
\]
\[
= \frac{\sigma^2(1 + (n - 1)\rho)}{M_{I-2} \sqrt{JP(1 - P)n(1 - \rho)}}
\]
\[
= \sqrt{\Pi_1(1 - \Pi_1)(1 + (n - 1)\rho)}
\]
\[(4A.23)\]

Notes

1. Charles S. Peirce and Joseph Jastrow (1885) put individual randomization to its earliest known use in a study of minute perceptible differences in the weights of physical objects and in their later studies of mental telepathy (for discussion, see Hacking 1988). Ronald A. Fisher (1925) first wrote about randomization in an article that focused on agricultural experiments.

2. In his textbook on experimental design in psychology and education—one of the first to be published on the subject—E. F. Lindquist (1953) provided an excellent overview of cluster randomization.

3. In the rare instances where the programmatic response to a place-based problem is to induce individuals to move, it is possible to evaluate the program by randomizing individuals. This was done, for example, in the Moving to Opportunity experiment (Kling, Ludwig, and Katz 2005), in which randomly selected residents of public housing projects received rental housing subsidies if they moved to neighborhoods with less concentrated poverty.

4. Donald B. Rubin (1980) referred to the assumed absence of spillover effects as the Stable Unit Treatment Value Assumption. D. R. Cox (1958, 19) referred to the assumed presence of spillover effects as "interference between different units."

5. Malcolm Gladwell (2000) described 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 fall in 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).

6. In one of the few studies that provide statistical evidence for the existence of the tipping phenomenon, George C. Galster, Robert G. Quercia, and Alvaro Cortes (2000) used U.S. Census data to estimate threshold effects for neighborhood characteristics such as the poverty rate, the unemployment rate, and the school dropout rate.

7. This principle, which is central to the theory of taxation in public finance, is often referred to as "horizontal equity" (Musgrave 1959, 160–61).

8. Graciela Teruel and Benjamin Davis (2000) described such a legal restriction on the PROGRESA program, an initiative designed to improve child health, nutrition, and education in rural Mexico.

9. Among the software packages that can perform this kind of analysis are HLM, SAS Proc Mixed, Stata gllamm, MLWIN, and VARCL.


11. Leslie Kish (1965) defined two design effects, one based on the standard errors of cluster sampling versus random sampling, and the other based on the error variances of cluster sampling versus random sampling.

12. This estimated range of intraclass correlations is based on reports by David M. Murray, Brenda L. Rooney et al. (1994) and Ohidul M. Siddiqui et al. (1996) for measures of adolescent smoking clustered by schools; by David Murray and Brian Short (1995) for measures of adolescent drinking clustered by community; by Marion Campbell, Jill Mollison, and Jeremy M. Grimshaw (2001) and Campbell, Grimshaw, and I. N. Steen (2000) for measures of physician practices and patient outcomes clustered by hospitals and by physician groups; and by Obioha C. Ukoumunne et al. (1999) for other kinds of measures.

13. Although William Scyly Gosset established this relationship in a paper published under his pseudonym "Student" (1908), Fisher (1925) first brought it to the attention of empirical researchers.

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

15. This discussion makes the simplifying assumption that each randomized cluster has the same number of individual members. Similar findings can be obtained when the number of individuals varies across clusters, provided that one uses the harmonic mean of the number of individuals per cluster.

17. Anthony S. Bryk and Stephen W. Raudenbush (1988) argued that one should expect program impacts to vary across individuals and that this variation provides an opportunity for learning how individuals respond to programs.

18. Mitchell H. Gail et al. (1996) used Monte Carlo simulations to illustrate this fact for parametric t-tests and nonparametric permutation tests. Jan Kmenta’s (1971, 254–56) expression for the effect of heteroscedasticity in a bivariate regression can be used to prove the same point.

19. 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, 226–28).

20. 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, and equations 4.7 and 4.8 may 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, this issue can be ignored for the moment.

21. Because the differential impact estimator is a four-group “difference of differences of means” based on all clusters in the full sample, it has \( J - 4 \) degrees of freedom.

22. The situation is even more favorable if the individual characteristic defining the subgroups is correlated with the outcome measure, as student gender, for example, is correlated with math test scores (boys generally score higher than girls). 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 clusters and the subgroup characteristic is correlated with the outcome, part of the between-cluster variance, \( \tau^2 \), is related to the subgroup characteristic and does not exist within subgroups. Because both these improvements can be obtained for the full sample estimator by controlling statistically for subgroup characteristics as a covariate, they are not implications of performing subgroup analysis in cluster randomization experiments per se.

23. Equation 4.12 includes an unconditional rather than conditional individual variance because the cluster covariates are constant within each cluster and thus cannot explain within-cluster variation.

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

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

26. Differences between findings for two-tailed tests and one-tailed tests are most pronounced where there are small numbers of clusters to be randomized and therefore few degrees of freedom.

27. At this writing, the final report on Jobs-Plus is being prepared (Bloom, Riccio, and Verma, forthcoming).

28. For simplicity, this section assumes that the cluster-error component is the same for all subgroups in a given cluster. However, the main point of the section still holds, although with less force, if the cluster-error components for different subgroups of individuals differ but are correlated with each other. In this case, the subgroup differential impact estimator “differences away” part (not all) of the cluster-error component.

References


U.S. Department of Education. 2003. Early Childhood Longitudinal Study, Kinder-
Learning More from Social Experiments


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.