

# Evaluation Review

<http://erx.sagepub.com>

---


## Using Cluster Random Assignment to Measure Program Impacts: Statistical Implications for the Evaluation of Education Programs

Howard S. Bloom, Johannes M. Bos and Suk-Won Lee  
*Eval Rev* 1999; 23; 445

The online version of this article can be found at:  
<http://erx.sagepub.com/cgi/content/abstract/23/4/445>

---

Published by:

 SAGE Publications

<http://www.sagepublications.com>

Additional services and information for *Evaluation Review* can be found at:

Email Alerts: <http://erx.sagepub.com/cgi/alerts>

Subscriptions: <http://erx.sagepub.com/subscriptions>

Reprints: <http://www.sagepub.com/journalsReprints.nav>

Permissions: <http://www.sagepub.com/journalsPermissions.nav>

*This article explores the possibility of randomly assigning groups (or clusters) of individuals to a program or a control group to estimate the impacts of programs designed to affect whole groups. This cluster assignment approach maintains the primary strength of random assignment—the provision of unbiased impact estimates—but has less statistical power than random assignment of individuals, which usually is not possible for programs focused on whole groups. To explore the statistical implications of cluster assignment, the authors (a) outline the issues involved, (b) present an analytic framework for studying these issues, and (c) apply this framework to assess the potential for using the approach to evaluate education programs targeted on whole schools. The findings suggest that cluster assignment of schools holds some promise for estimating the impacts of education programs when it is possible to control for the average performance of past student cohorts or the past performance of individual students.*

## **USING CLUSTER RANDOM ASSIGNMENT TO MEASURE PROGRAM IMPACTS**

### **Statistical Implications for the Evaluation of Education Programs**

HOWARD S. BLOOM

JOHANNES M. BOS

*Manpower Demonstration Research Corporation*

SUK-WON LEE

*New York University*

### **THE EVALUATION CHALLENGE**

Over the past several decades, there has been considerable experience with randomized experiments to measure the impacts of social programs (Greenberg and Robins 1986). At the heart of this approach is the random

---

**AUTHORS' NOTE:** *Our thanks to Michelle Moser and Steve Caso for providing us with the data on which this article is based; our thanks to Steve Raudenbush for helping us to better understand the statistical issues addressed by the article; our thanks to Phil Robins, Winston Lin, Judy Gueron, Charles Michalopoulos, and an anonymous referee for their insightful feedback on earlier versions of the article. Please direct requests for further information to [howard\\_bloom@mdrc.org](mailto:howard_bloom@mdrc.org).*



EVALUATION REVIEW, Vol. 23 No. 4, August 1999 445-469  
© 1999 Sage Publications, Inc.

assignment of individuals to a program group, which is supposed to receive program services, or a control group, which is not supposed to receive these services. Only chance determines who is and is not selected to receive services, and each eligible applicant has the same chance of being selected.<sup>1</sup>

This process creates program and control groups with no systematic pre-existing differences—the expected values of all background characteristics, whether measured or not, are the same for both groups. Although in small samples these characteristics may differ by chance due to random sampling error, the margin for error decreases as sample size increases. Hence, the control group experience provides an unbiased estimate of what the program group experience would have been without the program. The difference between these two experiences therefore represents an internally valid estimate of the program impact—what it caused to happen. No other methodology provides the same level of internal validity for program impact estimates (Hollister and Hill 1995).

However, some programs are designed to affect whole groups at once, not separate individuals. For example, school reform initiatives are intended to affect whole schools, comprehensive community initiatives are designed to affect whole communities, and health education programs are often targeted on whole geographic areas (Lasoff, Olson, and Sommerfeld 1994; Connell et al. 1995; Murray et al. 1994). For these programs, it usually is not possible to randomly assign individuals to a program or control group. If a school, a community, or a geographic area is selected for a program, then everyone within the group is potentially affected by the program.

However, it might be possible to randomly assign whole groups or “clusters.” For example, one could randomly select schools for a new education program from among those eligible to participate. The schools not chosen would provide a valid control group because their expected background characteristics would be the same as those for the schools chosen. However, randomly assigning whole schools will produce a smaller effective sample than randomly assigning individual children from these schools. Reducing the effective sample size, in turn, will reduce the statistical power of program impact estimates.

Therefore, random assignment of clusters (“cluster assignment”) can produce impact estimates that are internally valid but may have limited statistical power. As described below, the extent to which clustering reduces statistical power depends on the composition of the clusters involved. This, in turn, depends on how clusters are defined. Cluster assignment therefore may produce adequate statistical power for some purposes but not for others.

Our article examines the statistical power of cluster assignment empirically. We first outline the statistical issues involved and present an analytic

framework for studying them. We then use this framework to assess one potential application of cluster assignment—estimating the impacts of education initiatives. Our findings suggest that cluster assignment holds some promise for this application when it is possible to control for the average performance of past student cohorts or the past performance of individual students. The article concludes by considering the generalizability of our findings and by exploring several problems that can arise when cluster assignment is used in practice.

### STATISTICAL ISSUES

Consider the following situation. A random assignment study takes place in  $J$  sites (schools, communities, geographic areas, etc.), each containing  $n$  sample members (students, residents, etc.). Thus, the sample for the study consists of  $nJ$  individuals. For simplicity, assume that half the sample is assigned to a program group and half to a control group, although our argument holds for any program or control mix. There are three main ways to make this assignment randomly.

- Blocked random assignment would randomly assign half the individuals from each site to the program and half to the control group.
- Cluster random assignment would randomly assign half the sites to the program and half to the control group.
- Simple random assignment would randomly assign half of all individuals to the program and half to the control group, ignoring their sites.

In each case, the difference between the mean outcome for the program group and that for the control group is a valid estimator of the program impact, because its expected value equals the true impact. The main difference in the statistical properties of the three approaches is their statistical power.<sup>2</sup> These differences depend on the extent to which individuals vary within sites and the extent to which sites differ, on average, from each other. Consider the following two extreme cases.

*All variation is between sites.* If the outcome level is the same for all individuals in a site, but mean outcomes vary across sites, then blocked assignment will have no random sampling error. By ensuring that the program and control groups represent each site in the same proportion, blocking will ensure that the groups are identical, regardless of who is selected from each site. In contrast, random sampling error for cluster assignment will be at its

maximum, because program and control group differences will depend entirely on which sites are chosen for each group. Random sampling error for simple random assignment will lie between that for cluster assignment and blocked assignment.

*All variation is within sites.* If the mean outcome is the same for each site but outcomes vary across individuals, then cluster assignment will only reflect sampling error due to who is selected from each site, and thus, it will be equivalent to simple random assignment. Blocked assignment also will be equivalent to simple random assignment because there is no margin for blocking by site to reduce sampling error.

In practice, there usually is variation within and between sites. Hence, blocking by site will reduce random sampling error, and clustering by site will increase it. Therefore, cluster assignment will have the least statistical power for any given total sample size,  $nJ$ .

As noted above, however, for programs that affect whole groups, it usually is not possible to randomly assign individual sample members. Hence, neither blocked random assignment nor simple random assignment is feasible; cluster assignment is the only option available. Thus, it is important to find a way to assess the statistical power of this approach. As a first step toward this end, we restate the program and control group difference of means as the following regression with between-site and within-site error components.

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

where

- $Y_{ij}$  = the outcome for individual  $i$  in site  $j$ ,
- $\alpha$  = the mean outcome for the control population,
- $B_0$  = the true program impact,
- $P_{ij}$  = 1 for individuals subject to the program and 0 for others,
- $e_j$  = the error component for site  $j$ , which is independently and identically distributed with mean zero and variance  $\tau^2$ , and
- $\varepsilon_{ij}$  = the error component for individual  $i$  from site  $j$ , which is independently and identically distributed with mean zero and variance  $\sigma^2$ .

The coefficient,  $B_0$ , is the difference between the mean of outcome  $Y$  for persons subject to the program and the mean of  $Y$  for those not subject to it (the true program impact). The sample-based estimate of  $B_0$  is the program and control group difference of means,  $b_0$ . Random sampling error has two components:  $e_j$ , to represent site-specific differences in mean outcomes, and

$\varepsilon_{ij}$ , to represent individual differences in outcomes within sites. The variance of the site-specific error component is represented by  $\tau^2$ , and the variance of the individual-specific error component is represented by  $\sigma^2$ .

The expected value of the impact estimator,  $b_0$ , equals the true program impact,  $B_0$ . The standard error of the impact estimator for cluster assignment is (Raudenbush 1997),

$$SE(b_0)_{cluster} = \sqrt{\frac{4\tau^2}{J} + \frac{4\sigma^2}{nJ}}. \quad (2)$$

This standard error has a component due to between-site sampling error,  $\frac{4\tau^2}{J}$ , and a component due to within-site sampling error,  $\frac{4\sigma^2}{nJ}$ .

The preceding discussion assumes that a simple program and control group difference of means is used to estimate program impacts. However, most randomized experiments use a regression-adjusted difference of means to reduce the standard error of impact estimates by controlling statistically for background characteristics that are correlated with outcomes. Equation 3 specifies such a regression with a program variable,  $P_{ij}$ , an individual-level background characteristic or covariate,  $X_{ij}$ , and two error components,  $e_j^*$  and  $\varepsilon_{ij}^*$ .

$$Y_{ij} = \alpha + B_0 P_{ij} + B_1 X_{ij} + e_j^* + \varepsilon_{ij}^*. \quad (3)$$

Note that  $X_{ij}$  can be a group characteristic or an individual characteristic, and impact regressions can include any number or mix of these characteristics. Of particular value in this regard are measures of past performance for the same individuals (their pretest scores) or measures of average performance for previous groups (or cohorts) from the same cluster. The true program impact is still  $B_0$ , and the new regression-adjusted impact estimator is  $b_0^*$ . Once again, the expected value of the impact estimator,  $E(b_0^*)$ , equals the true impact,  $B_0$ .

Standard errors for regression-adjusted impact estimators represent generalizations of their counterparts for difference of means estimators. For example, consider the following standard errors for cluster assignment with a regression-adjusted impact estimator using a single group characteristic,  $X_j$ , or a single individual characteristics,  $X_{ij}$ , as the covariate (Raudenbush 1997). For a single group characteristic,

$$SE(b_0^*)_{cluster} = \sqrt{1 + \frac{1}{J-4}} \sqrt{\frac{4\tau^{2*}}{J} + \frac{4\sigma^{2*}}{nJ}}. \quad (4)$$

For a single individual characteristic,

$$SE(b_{0^{**}})_{cluster} = \sqrt{1 + \frac{1}{nJ-4}} \sqrt{\frac{4\tau^{2**}}{J} + \frac{4\sigma^{2**}}{nJ}}. \quad (5)$$

There are two major differences between Equations 4 and 5 for regression-adjusted impact estimators and Equation 2 for a simple treatment and control group difference of means. The first difference involves the inflation factor  $\sqrt{1 + \frac{1}{J-4}}$  or  $\sqrt{1 + \frac{1}{nJ-4}}$ . This factor rapidly approaches 1, as  $J$  (the number of clusters) in Equation 4 and  $nJ$  (the total number of individuals) in Equation 5 increase.

For example, with  $J$  equal to 10 (five program clusters and five control clusters), this inflation factor equals 1.08 in Equation 4. For  $J$  equal to 20, the inflation factor equals 1.03. Thus, for all but very small numbers of clusters, the inflation factor equals roughly 1. Likewise, for  $nJ$  equal to 100 (50 program individuals and 50 control individuals), the inflation factor is 1.005 in Equation 5, and for  $nJ$  equal to 500, it is 1.001. Thus, even for small numbers of clusters, the inflation factor for an individual covariate has virtually no effect on the standard error of the impact estimator.

A second, more important difference between the standard errors for difference of means estimators and their regression-adjusted counterparts under cluster assignment involves the error component variances ( $\tau^2$  and  $\sigma^2$ ,  $\tau^{2*}$  and  $\sigma^{2*}$ , or  $\tau^{2**}$  and  $\sigma^{2**}$ ). To the extent that a covariate "explains" some of the variation in the original error components, the variance of the remaining unexplained error, and thus the standard error of the program impact estimator, decreases accordingly.

A group covariate will have all of its effect on between-cluster error. It cannot explain within-cluster error because it has no within-cluster variation. Hence, a group covariate can reduce  $\tau^2$  but not  $\sigma^2$ . An individual covariate can reduce both between-cluster and within-cluster error variance. However, it might reduce between-cluster variance by less than would be possible for a group characteristic. Hence, it is not clear, a priori, whether a group characteristic or an individual characteristic will provide the most effective covariate for estimating program impacts.

Therefore, when considering cluster assignment to measure the impacts of a program, the following statistical issues should be addressed:

- How does the outcome of interest vary between clusters and within clusters? In other words, what are the values of  $\tau^2$  and  $\sigma^2$ ?
- To what extent can group characteristics and/or individual characteristics reduce these variance components and thus reduce the standard error of program impact estimates?

- Given the likely size of the clusters of interest and the covariates available for a study, how many clusters are required to provide enough statistical power to make the study worthwhile?<sup>3</sup>

Below, we illustrate how to address these questions for studies of schoolwide programs to improve student performance on standardized tests.

### ANALYSIS STRATEGY

Our basic approach is to explore the statistical implications for cluster assignment of the variance component structure of standardized test scores within and between elementary schools in one medium-size city, Rochester, New York. We do not actually compute program impact estimates because we do not examine a specific program. Instead, we infer what the statistical properties of a cluster assignment impact estimator would be if it were applied to a situation such as that of Rochester elementary schools.

Specifically, we use standardized test scores for individual students from 25 Rochester elementary schools to compute the between-school variance,  $\tau^2$ , and the within-school variance,  $\sigma^2$ , separately by the following: grade (for third grade and sixth grade), subject (math and reading), year (1989, 1990, 1991, and 1992), and different “impact estimation models” (covariate specifications).<sup>4</sup> Specific grades and subjects are studied separately because the impacts of educational programs typically are reported by grade. Findings for different years are reported separately to examine their stability over time. Findings for different impact estimation models are reported separately to assess their ability to increase the statistical power of cluster assignment designs.

For each combination of grade, subject, year, and model specification, we estimate  $\tau^2$  and  $\sigma^2$  and use these estimates to project the “minimum detectable effect” and “minimum detectable effect size” for different cluster assignment samples. Intuitively, a minimum detectable effect or a minimum detectable effect size is the smallest effect that a particular research design has a good chance of detecting. The smaller the minimum detectable effect or minimum detectable effect size is, the greater the statistical power of the research design is. A minimum detectable effect is expressed in the original units of the outcome measure (in our case, scale scores from a standardized test), whereas a minimum detectable effect size is expressed as a proportion of the standard deviation of the outcome measure (in our case, the sample standard deviation of individual scale scores).<sup>5</sup> Using these two related metrics, we



address the following question: "How many schools are needed to provide adequate statistical power for a cluster assignment design intended to measure program impacts on student performance?" The following 10 impact estimation models are examined.

#### **BASIC APPROACH (WITH NO COVARIATES)**

Model 1: A program and control group difference of mean test scores

#### **COHORT APPROACHES (WITH GROUP COVARIATE[S] ONLY)**

Model 2: Controlling for the mean test score of different students who were in the same grade in the previous year ( $Y_{jt-1}$ )

Model 3: Controlling for the mean test score of different students who were in the same grade in the previous year and the year before that ( $Y_{jt-1}$  and  $Y_{jt-2}$ )

Model 4: Controlling for the mean test score of different students who were in the same grade 2 years earlier ( $Y_{jt-2}$ )

Model 5: Controlling for the mean test score of different students who were in the same grade 2 and 3 years earlier ( $Y_{jt-2}$  and  $Y_{jt-3}$ )

#### **LONGITUDINAL APPROACHES (WITH INDIVIDUAL COVARIATE[S] ONLY)**

Model 6: Controlling for each individual student's test score in the previous grade ( $y_{ijt-1}$ )

Model 7: Controlling for each individual student's test score two grades earlier ( $y_{ijt-2}$ )

Model 8: Controlling for each student's test score in each of the previous two grades ( $y_{ijt-1}$  and  $y_{ijt-2}$ )

#### **COMBINED APPROACHES (WITH INDIVIDUAL AND GROUP COVARIATES)**

Model 9: Controlling for each individual student's test score in the previous year ( $y_{ijt-1}$ ) and the mean score of different students in the same grade a year earlier ( $Y_{jt-1}$ )

Model 10: Controlling for each individual student's test score 2 years prior ( $y_{ijt-2}$ ) and the mean score of different students in the same grade 2 years earlier ( $Y_{jt-2}$ )

Model 1, a simple treatment and control group difference of means, serves as our point of departure. The other models are versions of Equation 3, with different measures of past student test scores as covariates.

Models 2 through 5 use cohort data for each school/grade (cluster) to control for the mean test scores of its past student cohorts. Hence, these models rely on group covariates to control for "school effects." Model 2 controls for the average performance of each school's most recent past cohort (last year's students in the same grade). Model 3 controls for the average performance of each school's two most recent cohorts. These models could be used to estimate the impacts of a program during its first year of implementation. By comparing their projected statistical power, one can assess the value of obtaining data on 2 years versus 1 year of past school-level performance.

Models 4 and 5 skip over the most recent past cohort (last year's students) and therefore could be used to estimate the impacts of a program during its second year of implementation (when last year's students could have been affected by the program). By comparing the statistical power of Models 4 and 5 to that for Models 2 and 3, it is possible to project the loss of power that will occur if one has to wait a year before estimating the impacts of a program that is slow to startup. This comparison also illustrates how statistical power will erode as one moves from impact estimates for the first year of program follow-up to estimates for the second year.

Models 6, 7, and 8 use individual longitudinal data to control for each student's own past test scores. Hence, they rely on individual covariates to control for individual effects. Model 6 controls for each student's test score during the immediately preceding year, and Model 7 controls for his or her own score 2 years earlier. Model 8 controls for sample members' test scores in both of the past 2 years. Hence, Models 6 and 8 can be used to estimate program impacts during the first year of program implementation, and Model 7 can be used for the second year, in cases in which test scores for a previous year may have been affected by an ongoing program.

Models 9 and 10 complete our analysis by representing combinations of both individual and group covariates. Model 9 lags past individual performance and the performance of previous cohorts by 1 year and, hence, could be used to estimate impacts in the first year of program implementation. Model 10 lags past individual and group performance by 2 years and, hence, could be used to estimate impacts in Year 2 of program implementation.

As indicated above, we do not actually compute program impacts with each model. Instead, we use the results of standardized tests in Rochester to examine the error component structure of each model and thereby infer what its statistical properties would be if it were used to estimate program impacts in a similar environment.

For each impact estimation model, year, subject, and grade,  $\tau^2$  and  $\sigma^2$  were computed using variance components analysis. For Model 1, the variance

components were computed directly from individual test scores. For Models 2 through 10, they were computed in three steps: (a) individual scores were regressed on the appropriate group or individual covariate(s), (b) residuals were computed for this regression, and (c) the SAS VARCOMP procedure was used to estimate the variance components of the residuals (SAS Institute 1989).<sup>6</sup>

### DATA AND SAMPLE

Our outcome data were obtained from the Rochester, New York, school district. They comprise individual scores on Pupil Evaluation Program (PEP) tests for math and reading, which are administered each year to third graders and sixth graders throughout New York State. Rochester scores were available to us for 1989, 1990, 1991, and 1992. The average number of third graders per school in the sample each year ranged from 29 to 121, with a mean of 71. The average number of sixth graders per school ranged from 21 to 96, with a mean of 54.

The PEP test is norm referenced, not criterion referenced. It does not translate into grade equivalents or any other absolute criterion. Hence, its results only have meaning relative to the distribution of scores for a reference group; they do not have meaning in terms of specific identifiable knowledge, ability, or skills.

The reference group we use to interpret PEP test scores is our analysis sample. Table 1 summarizes the distribution of individual PEP test scale scores for this sample by grade, subject, and year. Several points are important to note about these distributions. First, note that although mean scores differ by grade and subject (because they represent different subject matter), they do not differ much over time (suggesting that different versions of the same test were similar and that average student performance did not change much in Rochester during the 4-year period we examined).

Second, note that the standard deviation of individual scores is about 10 or 11 points for all subjects, grades, and years examined. Hence, as explained later, to convert minimum detectable effects (in scale scores) to minimum detectable effect sizes (as a proportion of the standard deviation), we divide the former by roughly 10 or 11.

Third, note how the distribution of scale scores translates into a percentile distribution. For each subject, grade, and year, the difference in scale scores between the 25th and 75th percentiles is 14 to 17 points. In other words, a percentile difference of 50 points reflects a scale score difference of 14 to 17

**TABLE 1: Distribution of Individual PEP Test Scores (full sample)**

	1989	1990	1991	1992
Third-grade math				
Mean	44.3	41.5	44.6	46.0
Standard deviation	10.2	9.7	9.7	10.2
25th percentile	37	35	38	39
50th percentile	45	42	45	47
75th percentile	52	49	52	54
Third-grade reading				
Mean	36.1	36.0	35.6	35.7
Standard deviation	10.8	10.8	11.3	11.1
25th percentile	29	29	28	28
50th percentile	37	36	36	37
75th percentile	44	44	45	44
Sixth-grade math				
Mean	38.0	33.9	34.6	32.6
Standard deviation	11.4	11.4	11.5	11.5
25th percentile	30	25	26	24
50th percentile	37	33	33	31
75th percentile	46	42	42	41
Sixth-grade reading				
Mean	56.5	58.5	56.7	56.1
Standard deviation	11.5	11.5	11.8	12.1
25th percentile	51	52	51	49
50th percentile	59	61	58	57
75th percentile	65	67	66	65

NOTE: PEP = Pupil Evaluation Program. Sample sizes for 1989, 1990, 1991, and 1992, respectively, are, 1,728, 1,748, 1,815, and 1,794 for third-grade math; 1,724, 1,741, 1,806, and 1,797 for third-grade reading; 1,229, 1,398, 1,313, and 1,475 for sixth-grade math; and 1,229, 1,398, 1,314, and 1,468 for sixth-grade reading.

points. This implies that each 1-point difference in scale scores represents a 3-percentile difference. This relationship plays an important role in our interpretation of minimum detectable effects reported later.

Two different samples are used for our analysis: the full sample and a longitudinal subsample. The full sample is used to examine the cohort Models 2 through 5. It contains between 1,724 and 1,815 third graders and between 1,229 and 1,475 sixth graders each year. A subsample of students with test scores for multiple years is used to examine longitudinal approaches. Specifically, we focus on sixth graders with test scores for fourth, fifth, and sixth grade, and on third graders with test scores for first, second, and third grade. It was only possible to obtain these longitudinal data for third graders or sixth graders who took PEP tests in 1991 and 1992.<sup>7</sup>

In the discussion that follows, we first present findings for the cohort approaches (Models 1-5) based on data for the full sample. We then present findings for the longitudinal approaches (Models 6-8) and combined approaches (Models 9 and 10) based on data for the longitudinal subsample. To facilitate a direct comparison of the cohort approaches, the longitudinal approaches and the combined approaches for the same individuals, we also present findings for the cohort approaches for the longitudinal subsample.

### MINIMUM DETECTABLE EFFECTS AND MINIMUM DETECTABLE EFFECT SIZES FOR THE COHORT APPROACHES

By definition, the minimum detectable effect of a study is the smallest true effect that has a  $W\%$  chance of producing an impact estimate that is statistically significant at the  $Z$  level. For a one-tail hypothesis test (to assess program-induced improvement, not just change) at the 0.05 level of statistical significance ( $Z$ ), with 80% statistical power ( $W$ ), the minimum detectable effect is 2.5 times the standard error of an impact estimator.<sup>8</sup> In other words,

$$\text{MDE}(b_0) = 2.5 \text{ SE}(b_0). \quad (6)$$

Substituting Equation 2 into Equation 6 yields,

$$\text{MDE}(b_0) = 5 \sqrt{\frac{\tau^2}{J} + \frac{\sigma^2}{nJ}}. \quad (7)$$

As can be seen, the minimum detectable effect for cluster assignment is

- inversely proportional to the square root of the total number of clusters,  $J$ ;
- inversely but not proportionally related to the number of individuals per cluster,  $n$ ; and
- directly but not proportionally related to the between-cluster and within-cluster variance components,  $\tau^2$  and  $\sigma^2$ .

Minimum detectable effects are a simple way to express statistical power, but to interpret them requires a basis for judging their policy relevance. From a benefit-cost perspective, one might ask whether a proposed study could detect the smallest effect that would make a program break even. From a political perspective, one might ask whether the study could detect the smallest effect that would be deemed as "having made an important difference."

**TABLE 2: Estimated Minimum Detectable Effects for Cohort Approaches (full sample)**

	<i>Third Grade</i>		<i>Sixth Grade</i>		<i>Mean</i>
	<i>Math</i>	<i>Reading</i>	<i>Math</i>	<i>Reading</i>	
Model 1 (no covariates)					
10 Schools	6.6	7.1	7.5	6.3	6.9
20 Schools	4.7	5.0	5.3	4.5	4.9
30 Schools	3.8	4.1	4.3	3.6	4.0
40 Schools	3.3	3.6	3.7	3.2	3.4
60 Schools	2.7	2.9	3.0	2.6	2.8
Model 2 ( $Y_{it-1}$ )					
10 Schools	3.6	4.7	4.9	3.9	4.3
20 Schools	2.4	3.2	3.3	2.6	2.9
30 Schools	2.0	2.6	2.7	2.1	2.3
40 Schools	1.7	2.2	2.3	1.8	2.0
60 Schools	1.4	1.8	1.9	1.5	1.6
Model 3 ( $Y_{it-1}, Y_{it-2}$ )					
10 Schools	3.7	3.5	4.7	4.3	4.0
20 Schools	2.4	2.3	3.0	2.8	2.6
30 Schools	1.9	1.8	2.4	2.2	2.1
40 Schools	1.6	1.5	2.1	1.9	1.8
60 Schools	1.3	1.3	1.7	1.5	1.4
Model 4 ( $Y_{it-2}$ )					
10 Schools	4.3	3.9	5.4	4.5	4.5
20 Schools	2.9	2.6	3.7	3.0	3.1
30 Schools	2.3	2.1	3.0	2.4	2.5
40 Schools	2.0	1.8	2.5	2.1	2.1
60 Schools	1.6	1.5	2.1	1.7	1.7
Model 5 ( $Y_{it-2}, Y_{it-3}$ )					
10 Schools	4.3	3.6	5.0	4.9	4.5
20 Schools	2.8	2.3	3.2	3.1	2.9
30 Schools	2.2	1.9	2.6	2.5	2.3
40 Schools	1.9	1.6	2.2	2.1	2.0
60 Schools	1.5	1.3	1.8	1.7	1.6

NOTE: Based on the mean values of  $\tau^2$  and  $\sigma^2$  for all years of available full-sample data for each model, and assuming 60 students per school (approximately the average grade size for the full sample).

From a programmatic perspective, one might ask whether the study could detect an effect that had a “reasonable chance of being achieved.” Which perspective is applied, and what data are used to inform it will vary from application to application. But as with any measure of statistical power, some such determination must be made in order to interpret it.

Table 2 presents minimum detectable effects for the cluster assignment of 10, 20, 30, 40, and 60 schools (half to a program group and half to a control group), given our estimates of  $\tau^2$  and  $\sigma^2$  and assuming a grade size of 60 students per school. Consider the mean findings in Column 5 for Model 1. For 10 schools (600 students), the minimum detectable effect would be 6.9 points; for 20 schools (1,200 students), it would be 4.9 points; for 30 schools (1,800 students), it would be 4.0 points; for 40 schools (2,400 students), it would be 3.4 points; and for 60 schools (3,600 students), it would be 2.8 points. The situation improves considerably if we control for the performance of at least one recent cohort. For example, under Model 2, the minimum detectable effect would be 4.3, 2.9, 2.3, 2.0, or 1.6 points for 10, 20, 30, 40, or 60 schools, respectively.

Because there is no absolute benchmark for interpreting a specific change in PEP test scores, it is difficult to determine whether the minimum detectable effects for Models 2 through 5 would be adequate for studying a particular educational initiative. One way to do so, however, is to compare this change to the distribution of individual test scores in Table 1.

Table 1 suggests that a 2-point difference in PEP scores represents a 6-percentile difference in the distribution of individual scores. Thus, a minimum detectable effect of 2 points (in the range attainable using 40 to 60 schools) is equivalent to raising the mean performance of program group members by 6 percentile points. For example, it would be equivalent to raising their performance from the 50th to the 56th percentile in the original sample distribution. It seems reasonable to expect a major educational initiative to produce an improvement of at least this magnitude for it to be deemed successful. Hence, a minimum detectable effect in this range might be acceptable for an evaluation of a major educational initiative.

To provide another perspective on the magnitudes of the findings in Table 2, we transformed them into minimum detectable effect sizes by dividing each by the average full-sample standard deviation for the PEP scale scores for the subject and grade involved. These results are presented in Table 3.<sup>9</sup>

Measuring effect size in units of standard deviations is a common way to standardize impact estimates from different studies in order to summarize and compare them. This approach is especially useful for meta-analyses of treatment effectiveness studies that report impacts in different units and for different types of outcomes (Glass, McGaw, and Smith 1981).

Effect size is also a common metric for discussions of statistical power in the behavioral sciences. To provide guidance for researchers, Cohen (1977, 1988) suggests that effect sizes around 0.20 be considered small, those around 0.50 be considered medium-size, and those above 0.80 be considered large.

**TABLE 3: Estimated Minimum Detectable Effect Sizes for Cohort Approaches (full sample)<sup>a</sup>**

	<i>Third Grade</i>		<i>Sixth Grade</i>		<i>Mean</i>
	<i>Math</i>	<i>Reading</i>	<i>Math</i>	<i>Reading</i>	
Model 1 (no covariates)					
10 schools	0.67	0.65	0.65	0.54	0.63
20 schools	0.47	0.46	0.46	0.38	0.44
30 schools	0.38	0.37	0.38	0.31	0.36
40 schools	0.33	0.32	0.33	0.27	0.31
60 schools	0.27	0.26	0.27	0.22	0.26
Model 2 ( $Y_{it-1}$ )					
10 schools	0.37	0.43	0.43	0.33	0.39
20 schools	0.25	0.29	0.29	0.23	0.26
30 schools	0.20	0.23	0.23	0.18	0.21
40 schools	0.17	0.20	0.20	0.16	0.18
60 schools	0.14	0.16	0.16	0.13	0.15
Model 3 ( $Y_{it-1}, Y_{it-2}$ )					
10 schools	0.37	0.32	0.41	0.36	0.37
20 schools	0.24	0.21	0.26	0.23	0.24
30 schools	0.19	0.16	0.21	0.19	0.19
40 schools	0.16	0.14	0.18	0.16	0.16
60 schools	0.13	0.11	0.15	0.13	0.13
Model 4 ( $Y_{it-2}$ )					
10 schools	0.43	0.35	0.47	0.38	0.41
20 schools	0.29	0.24	0.32	0.26	0.28
30 schools	0.24	0.19	0.26	0.21	0.22
40 schools	0.20	0.17	0.22	0.18	0.19
60 schools	0.17	0.13	0.18	0.15	0.16
Model 5 ( $Y_{it-2}, Y_{it-3}$ )					
10 schools	0.44	0.33	0.44	0.42	0.41
20 schools	0.28	0.21	0.28	0.27	0.26
30 schools	0.23	0.17	0.22	0.21	0.21
40 schools	0.19	0.15	0.19	0.18	0.18
60 schools	0.16	0.12	0.16	0.15	0.14

NOTE: Based on the mean values of  $\tau^2$  and  $\sigma^2$  and the sample standard deviation for all years of available full-sample data for each model, and assuming 60 students per school (approximately the average grade size for the full sample).

a. The minimum detectable effect size equals the minimum detectable effect measured in raw Pupil Evaluation Program test scores divided by the standard deviation of the raw scores.

These guidelines have been used by researchers in many fields for many years.

Lipsey (1990) provides an empirical justification for Cohen's effect-size standards based on the distribution of 102 mean effect sizes obtained from



186 meta-analyses of 6,700 studies representing 800,000 sample members. The lower third of this distribution (small effects) ranged from 0.00 to 0.32, the middle third (medium effects) ranged from 0.33 to 0.55, and the upper third (large effects) ranged from 0.56 to 1.20. The majority of meta-analyses in Lipsey's summary represent educational research, and the distribution of effect sizes is about the same for educational research and noneducational research (p. 54). Hence, these findings provide a relevant guide for judging the minimum detectable effect sizes in Table 3.

As can be seen, if it is possible to control for the performance of at least one recent cohort from each school (Models 2-5), the minimum detectable effect size for cluster assignment of 30 to 60 schools (1,800 to 3,600 students, respectively) is about 0.20. This would be considered a small effect, both according to Cohen's (1977, 1988) guidelines and Lipsey's (1990) empirical findings.

### **CORRESPONDING FINDINGS FOR THE LONGITUDINAL APPROACHES**

Our analysis of longitudinal impact estimators focuses on the subsample of Rochester sixth graders and third graders, with individual test scores for 3 consecutive years. As noted above, this information could only be obtained for the 1991 and 1992 samples, and it was available for 85% to 90% of these sample members.<sup>10</sup>

For each subsample, we computed  $\tau^2$  and  $\sigma^2$  for each impact estimation model. We then computed the minimum detectable effects and minimum detectable effect sizes in Table 4. The first two columns in the table report minimum detectable effects measures in PEP test scores; the last two columns present minimum detectable effect sizes computed as a proportion of the standard deviation of PEP test scores. All estimates are for a cluster assignment design, with a total of 40 schools and 60 students per school, with half of the schools randomly assigned to a program and half to a control group. The findings suggest,

- Controlling for past average school performance or past individual student performance markedly reduces the minimum detectable effect and minimum detectable effect size for all but third-grade math scores (discussed below).
- Controlling for more recent past performance reduces the minimum detectable effect and minimum detectable effect size by more than controlling for less recent past performance.

**TABLE 4: Estimated Minimum Detectable Effects and Effect Sizes for Longitudinal and Cohort Approaches (longitudinal subsample)**

	<i>Minimum Detectable Effect and Effect Size for 40 Schools and 60 Students per School</i>			
	<i>Raw PEP Test Score</i>		<i>Effect Size<sup>a</sup></i>	
	<i>1991</i>	<i>1992</i>	<i>1991</i>	<i>1992</i>
Sixth-grade math				
Model 1 (no covariates)	3.9	3.7	0.36	0.34
Longitudinal approaches				
Model 6 ( $Y_{ijt-1}$ )	1.7	2.2	0.16	0.20
Model 7 ( $Y_{ijt-2}$ )	2.7	2.5	0.25	0.23
Model 8 ( $Y_{ijt-1}, Y_{ijt-2}$ )	1.7	2.1	0.16	0.20
Cohort approaches				
Model 2 ( $Y_{jt-1}$ )	2.0	2.3	0.18	0.21
Model 4 ( $Y_{jt-2}$ )	2.9	2.3	0.27	0.21
Model 3 ( $Y_{jt-1}, Y_{jt-2}$ )	2.0	2.2	0.19	0.20
Longitudinal + cohort approaches				
Model 9 ( $Y_{ijt-1}, Y_{jt-1}$ )	1.3	1.9	0.12	0.17
Model 10 ( $Y_{ijt-2}, Y_{jt-2}$ )	2.0	2.0	0.19	0.18
Sixth-grade reading				
Model 1 (no covariates)	3.3	2.8	0.32	0.25
Longitudinal approaches				
Model 6 ( $Y_{ijt-1}$ )	1.4	1.0	0.13	0.09
Model 7 ( $Y_{ijt-2}$ )	1.3	1.8	0.13	0.16
Model 8 ( $Y_{ijt-1}, Y_{ijt-2}$ )	1.2	1.0	0.11	0.09
Cohort approaches				
Model 2 ( $Y_{jt-1}$ )	1.7	1.9	0.17	0.18
Model 4 ( $Y_{jt-2}$ )	2.0	2.2	0.19	0.20
Model 3 ( $Y_{jt-1}, Y_{jt-2}$ )	1.8	2.0	0.17	0.18
Longitudinal + cohort approaches				
Model 9 ( $Y_{ijt-1}, Y_{jt-1}$ )	1.2	1.0	0.11	0.09
Model 10 ( $Y_{ijt-2}, Y_{jt-2}$ )	1.1	1.8	0.11	0.16
Third-grade math				
Model 1 (no covariates)	3.6	3.3	0.37	0.33
Longitudinal approaches				
Model 6 ( $Y_{ijt-1}$ )	3.3	2.8	0.34	0.27
Model 7 ( $Y_{ijt-2}$ )	2.9	2.9	0.30	0.28
Model 8 ( $Y_{ijt-1}, Y_{ijt-2}$ )	2.9	2.6	0.30	0.25
Cohort approaches				
Model 2 ( $Y_{jt-1}$ )	1.7	1.7	0.18	0.17
Model 4 ( $Y_{jt-2}$ )	2.1	2.1	0.21	0.20
Model 3 ( $Y_{jt-1}, Y_{jt-2}$ )	1.6	1.7	0.17	0.16

(continued)

TABLE 4: Continued

	<i>Minimum Detectable Effect and Effect Size for 40 Schools and 60 Students per School</i>			
	<i>Raw PEP Test Score</i>		<i>Effect Size<sup>a</sup></i>	
	<i>1991</i>	<i>1992</i>	<i>1991</i>	<i>1992</i>
Third-grade math				
Longitudinal + cohort approaches				
Model 9 ( $y_{ijt-1}, Y_{jt-1}$ )	2.0	1.5	0.20	0.15
Model 10 ( $y_{ijt-2}, Y_{jt-2}$ )	2.1	2.1	0.22	0.21
Third-grade reading				
Model 1 (no covariates)	3.8	3.5	0.35	0.31
Longitudinal approaches				
Model 6 ( $y_{ijt-1}$ )	2.1	1.8	0.19	0.16
Model 7 ( $y_{ijt-2}$ )	1.8	2.1	0.17	0.19
Model 8 ( $y_{ijt-1}, y_{ijt-2}$ )	1.8	1.8	0.16	0.16
Cohort approaches				
Model 2 ( $Y_{jt-1}$ )	2.2	2.0	0.20	0.18
Model 4 ( $Y_{jt-2}$ )	2.1	1.7	0.19	0.15
Model 3 ( $Y_{jt-1}, Y_{jt-2}$ )	1.7	1.5	0.16	0.13
Longitudinal + cohort approaches				
Model 9 ( $y_{ijt-1}, Y_{jt-1}$ )	1.9	1.6	0.17	0.15
Model 10 ( $y_{ijt-2}, Y_{jt-2}$ )	1.3	1.7	0.12	0.15

NOTE: PEP = Pupil Evaluation Program. Sixth-grade sample sizes for 1991 math and reading and 1992 math and reading are 1,153, 1,153, 1,363, and 1,365, respectively; those for third graders are 1,545, 1,545, 1,754, and 1,754.

a. The minimum detectable effect size equals the minimum detectable effect measured in raw PEP test scores divided by the standard deviation of the raw scores.

- Controlling for 2 years of past performance reduces the minimum detectable effect and minimum detectable effect size by slightly more than controlling for 1 year of past performance.
- Controlling for past individual performance reduces the minimum detectable effect and minimum detectable effect size by slightly more than controlling for past school performance, for all but third-grade math scores (discussed below).
- In general, controlling for both past school performance and past individual performance reduces the minimum detectable effect and minimum detectable effect size by more than controlling for only one of these alternatives.<sup>11</sup>

In general, the greatest increase in statistical power is produced by individual measures of recent past test performance (Model 6). The minimum detectable effect for this approach ranged from about 1 to 2 scale score points, which represents roughly 3 to 6 percentiles (for all but third-grade

math scores, where controlling for past school performance was far more effective). The corresponding minimum detectable effect size ranged from 0.10 to 0.20. Hence, by controlling for recent individual test scores, it is possible for cluster assignment with 40 schools and 60 students per school to detect relatively small improvements in school performance.

The next greatest increase in statistical power was produced by measures of recent past school performance, with a minimum detectable effect around 2 scale score points or 6 percentiles and a minimum detectable effect size ranging from about 0.15 to 0.20. Findings for the other models also suggest an ability to detect fairly small program impacts.

The one apparent anomaly in the findings is that past individual-level performance (Models 6-8) has very little effect on statistical power for estimating program impacts on third-grade math scores. This result was obtained both for 1991 and 1992. One potential explanation for it is that third-grade math tasks differ fundamentally from those for first and second grade. Hence, first- and second-grade math performance may not provide strong predictors for third-grade performance.

### CONCLUSIONS, LIMITATIONS, AND FURTHER ISSUES

This article was motivated by the need for a rigorous way to estimate the impacts of programs designed to affect whole groups. For such programs, it usually is not possible to randomly assign individuals to a program or control group. As an alternative, we explore the possibility of randomly assigning groups or clusters. Although the statistical theory of cluster sampling has been known for many years (Cochran 1963), the properties of cluster assignment for specific applications are not well known. Thus, to explore the feasibility of using this approach requires empirical analysis of its properties for applications being considered.

To facilitate such analyses, we tried to do three things: (a) clarify the statistical issues involved, (b) provide an analytic framework for studying these issues, and (c) use the framework to assess one potential application of cluster assignment—evaluating educational programs targeted on whole schools.

Our empirical results suggest that cluster assignment of schools holds some promise when it is possible to control for either the past performance of individual students (individual effects) or the average performance of recent past student cohorts (school effects). These findings are quite robust. They

hold for two different grades (third grade and sixth grade), two different subjects (math and reading), and four different years (1989, 1990, 1991, and 1992).

On balance, we find that controlling for individual effects improves statistical power by slightly more than controlling for school effects. Controlling for more recent past performance improves statistical power by slightly more than controlling for less recent past performance. Controlling for 2 years of past performance improves statistical power by slightly more than controlling for 1 year of past performance. But most important, all of these approaches improve statistical power substantially.

Consequently, we project that if a good measure of past individual or school performance is available, it might be possible to detect a 3- to 6-percentile improvement in average student performance with cluster assignment of 40 schools and 60 students per school (2,400 students overall). This implies an effect size of roughly 0.10 to 0.20, which by most existing standards suggests adequate statistical power.

Nevertheless, our findings represent only one step toward a better understanding of the strengths and limitations of cluster assignment. To explore this issue further, the following issues must be considered.

### **Do Our Findings Apply to Other Standardized Tests?**

We replicated key portions of our analysis for outcome measures based on individual scores from the California Achievement Test (CAT) in math, administered to Rochester fifth graders during 1989, 1990, 1991, and 1992. These findings were consistent with those reported for third-grade and sixth-grade PEP tests.

### **Will Our Findings Apply When Different Tests Are Used by Schools in Different Years?**

Schools often change the standardized test they use. Hence, the test used for current students might differ from that used for past students. To examine the implications of this possibility, we analyzed current student math performance, controlling for the reading performance of previous cohorts, and vice versa. Our results were quite similar to those presented above. Hence, our findings are not sensitive to how past cohort performance is measured.

### **Can Our Findings Apply to a Study Conducted in More Than One City?**

Our findings reflect the variance components of test scores for students and schools in one city. Hence, they indicate what would happen if these schools were randomly assigned to a program group or control group. But, more than one city might be necessary to recruit enough eligible schools for an impact study. This could add a between-city variance component to test scores. However, blocking schools by city would eliminate this extra variance. For example, one might recruit 8 schools from each of five cities (40 schools) and randomly assign 4 schools from each city to the program and 4 to the control group. Doing so would remove all city-specific differences between the program group and control group. Hence, this variance component would not affect the statistical power of program impact estimates.

### **How Sensitive Is Cluster Assignment to Contamination of the Treatment?**

When individuals are randomly assigned to a program, some may not participate, and some of those assigned to the control group may inadvertently receive program services.<sup>12</sup> Consequently, the difference between program services received by those assigned to the program and those assigned to the control group is diluted, and the measured program impact is attenuated.

When clusters are randomly assigned, these problems can be even more severe. If, for example, in a 20-school evaluation of a reform, 2 of the 10 program schools fail to implement the reform and 2 of the control schools develop a close alternative, one's ability to identify program impacts can be lessened substantially. It is not possible to correct such a problem by eliminating the failed program schools or the control schools that adopt a similar program, because doing so would compromise the experimental research design. For those considering cluster assignment, this means that stringent control of the treatment is essential.

### **How Sensitive Is Cluster Assignment to Experimental Attrition?**

Attrition from a study, or failure to collect follow-up data on some sample members, is potentially a more serious problem than contamination of the treatment. Instead of diluting the treatment contrast, such attrition (also

referred to as experimental mortality) compromises the internal validity of the experimental design because program group members and control group members for whom follow-up data are available may represent nonrandom subsamples of the original program and control groups. Thus, a decision by one school in a multischool study to stop providing data could seriously undermine the quality of a study and the believability of its results.

This problem could be offset somewhat by grouping clusters into blocks of two and randomly assigning one cluster from each block to the program and one to the control group. In this case, if one cluster dropped out of the study, the other cluster from its block could be dropped as well. This would reduce the sample size of the study but would not bias its impact estimates.

### **How Sensitive Is Cluster Assignment to Outliers?**

Another source of problems for cluster assignment is the possibility that unusual circumstances will produce an aberration in the outcomes for a whole cluster. In a study based on individual random assignment, a sample member may win the lottery or go to prison. These rare events could have a strong effect on the outcome for an individual, but they are unlikely to markedly affect the overall mean outcome for program group members or control group members. However, if a philanthropist decides to donate a large sum of money to a control school in an educational experiment, this donation might affect the experience of a large proportion of the sample at once. Even though the standard error of impact estimates accounts for these random events, they could make any single impact estimate unbelievable.

### **Will Our Findings Apply to Other School Systems?**

Our findings represent the variance components of elementary schools in Rochester, New York, between 1989 and 1992. To the extent that the variance components of our sample schools are similar to those in other school systems, our findings will apply elsewhere. To the extent that Rochester is idiosyncratic in this regard, our findings will not apply elsewhere. The only way to answer this question is to replicate our analysis for other school systems.

### **Will Our Findings for Schools Apply to Other Types Of Clusters?**

The variance components of clusters depend on how clusters are defined and the forces causing individuals to group together in them. One likely

determinant is the geographic scope of each cluster. For example, census blocks (small geographic units) are probably more homogeneous and differ more on average from each other than do census tracts (which comprise numerous census blocks). Census tracts are probably more homogeneous and differ more, on average, from each other than do municipalities (which comprise many census tracts). This reflects the way people concentrate geographically. Hence, one cannot apply our findings directly to comprehensive community initiatives, community health education programs, or other programs targeted on different types of clusters. Nevertheless, it is possible to use our analytic framework to examine the statistical properties of cluster assignment for these applications.

## NOTES

1. More complex designs randomly assign sample members to different treatment groups or a control group (e.g. Freedman and Friedlander 1995).

2. These approaches have two other differences that are important to note but lie outside the scope of the present article. One difference represents a limitation of cluster assignment; the other represents an advantage.

The limitation stems from the fact that cluster assignment cannot produce separate experimental impact estimates for each site (cluster), whereas blocked assignment can do so. However, this issue is not germane to the present discussion about situations in which blocked assignment is not possible.

The advantage of cluster assignment stems from its ability to capture program macroeffects. Macroeffects (Garfinkel, Manski, Michalopoulos 1992; Harris 1985) represent changes in the environment at a site caused by a program that influences outcomes for both control group members and program group members. Hence, by comparing outcomes for these two groups, blocked random assignment will miss these effects. To date, however, there is very little evidence about the existence of such macroeffects.

3. Raudenbush (1997) provides a framework for jointly determining the optimal number and size of clusters. Our examples take cluster size (school size) as given, but our approach can be generalized to decisions about the number and size of clusters.

4. There were 34 elementary schools in Rochester, all of which had a third grade, but only 25 of which had a sixth grade. To maintain the same sample of schools for our third-grade and sixth-grade analyses, we only report findings for the 25 schools that had both grades. Further analyses, not reported here, indicate that third-grade findings for all 34 schools are similar to those for the 25 schools in our sample.

5. See Bloom (1995) for a discussion of minimum detectable effects and Cohen (1977, 1988) for a discussion of effect size.

6. SAS PROC MIXED combines the steps that we used into one procedure (SAS Institute 1997), but we kept them separate to reflect the intuition of the process.

7. Scores for first, second, fourth, and fifth grade were obtained from CAT, CT5, and DRP tests.



8. Bloom (1995) illustrates that the minimum detectable effect in terms of the standard normal deviate is  $z0.95$  plus the absolute value of  $z0.20$  for this case.

9. For each subject and grade, we used the mean of the standard deviations for 1989, 1990, 1991, and 1992.

10. For sixth graders, the full samples sizes for 1991 math and reading and 1992 math and reading, respectively, are 1,313, 1,314, 1,475, and 1,468; their counterparts for the longitudinal subsample are 1,153, 1,153, 1,363, and 1,365. For third graders, the corresponding full-sample sizes are 1,815, 1,806, 1,794, and 1,797, and the corresponding longitudinal subsample sizes are 1,545, 1,545, 1,754 and 1,754

11. The most noticeable exception to this rule (for 1991 third-grade math scores) is probably due to sampling error in the variance component estimates. Close examination of this finding indicates that although the total individual error variance was smaller for Model 9 than for Models 2 or 6, the allocation of this total to each variance component was such that the corresponding minimum detectable effect for Model 9 was slightly larger than that for Model 2.

12. Bloom (1984) provides a correction for program group members who do not receive program services, which he refers to as "no-shows" in experimental research. Haynes and Dantes (1987) provide a similar correction for this problem in clinical trials, which they refer to as non-compliance. Bloom et al. (1997) provide a corresponding correction for both no-shows and crossovers (control group members who receive program services).

## REFERENCES

- Bloom, H. S. 1984. Accounting for no-shows in experimental evaluation designs. *Evaluation Review* 8 (2): 225-46.
- Bloom, H. S. 1995. Minimum detectable effects: A simple way to report the statistical power of experimental designs. *Evaluation Review* 19 (5): 547-56.
- Bloom, H. S., Larry L. Orr, S. H. Bell, G. Cave, F. Doolittle, W. Lin, and J. M. Bos. 1997. The benefits and costs of JTPA Title II-A Programs: Key findings from the National Job Training Partnership Study. *The Journal of Human Resources* 32 (3): 549-76.
- Cochran, W. G. 1963. *Sampling techniques*. New York: John Wiley.
- Cohen, J. 1977. *Statistical power analysis for the behavioral sciences*, rev. ed. New York: Academic Press.
- Cohen, J. 1988. *Statistical power analysis for the behavioral sciences*, 2d. ed. Hillsdale, NJ: Lawrence Erlbaum.
- Connell, J. P., A. C. Kubisch, L. B. Schorr, and C. Weiss, eds. 1995. *New approaches to evaluating community initiatives: Concepts, methods and contexts*. Washington, DC: Aspen Institute.
- Freedman, S., and D. Friedlander. 1995. *The JOBS evaluation: Early findings on program impacts in three sites*. U.S. Department of Health and Human Services. Administration for Children and Families, Office of the Assistant Secretary for Planning and Evaluation.
- Garfinkel, I., C. F. Manski, and C. Michalopoulos. 1992. Micro experiments and macro effects. In *Evaluating welfare and training programs*, edited by C. F. Manski and I. Garfinkel. Cambridge, MA: Harvard University Press.
- Glass, G. V., B. McGaw, and M. L. Smith. 1981. *Meta-analysis in Social Research*. Beverly Hills, CA: Sage.

- Greenberg, D., and P. Robins. 1986. The changing role of social experiments in policy analysis. *Journal of Policy Analysis and Management* 5 (2): 340-652.
- Harris, J. 1985. Macroexperiments versus microexperiments for health policy. In *Social experimentation*, edited by J. A. Hausman and D. A. Wise. Chicago: University of Chicago Press.
- Haynes, R. B., and R. Dantes. 1987. Patient compliance and the conduct and interpretation of therapeutic trials. *Controlled Clinical Trials* 8 (1): 12-9.
- Hollister, R. G., and J. Hill. 1995. Problems in the evaluation of community-wide initiatives. In *New approaches to evaluating community initiatives: Concepts, methods, and contexts*, edited by J. Connell, A. C. Kubisch, L. B. Schorr, and C. H. Weiss, 127-72. Washington, DC: Aspen Institute.
- Lasoff, M., L. Olson, and M. Sommerfeld. 1994. School-reform networks at a glance. *Education Week* November 2, 1994.
- Lipsey, M. 1990. *Design sensitivity: Statistical power for experimental research*. Newbury Park, CA: Sage.
- Murray, D. M., P. J. Hannan, D. R. Jacobs, P. J. McGovern, L. Schmid, W. L. Baker, and C. Gray. 1994. Assessing intervention effects in the Minnesota Heart Health Program. *American Journal of Epidemiology* 139 (1): 91-103.
- Raudenbush, S. W. 1997. Statistical analysis and optimal design in cluster randomized trials. *Psychological Methods* 2 (2): 173-85.
- SAS Institute Inc. 1989. Chapter 44: The VARCOMP procedure. In *SAS/STAT user's guide*, Volume 2 Version 6, 1661-7. Cary, NC: SAS Institute.
- SAS Institute Inc. 1997. Chapter 18: The MIXED procedure. In *SAS/STAT Software: Changes and enhancements through release 6.12*, 577. Cary, NC: SAS Institute.

*Howard S. Bloom is Chief Social Scientist for the Manpower Demonstration Research Corporation. He has been principal investigator for numerous large-scale experimental and quasi-experimental evaluation studies and taught research methods and applied statistics at Harvard University and New York University for over 20 years.*

*Johannes Bos is a Senior Research Associate at the Manpower Demonstration Research Corporation. He has played a lead role in a number of major MDRC evaluations of education and training programs for disadvantaged youths, teen parents, and welfare recipients. In addition, he has taught courses in program evaluation and statistics at New York University.*

*Suk-Won Lee is a Ph.D. candidate in the Robert F. Wagner School of Public Service at New York University. His research interests are mainly in the areas of program evaluation, social experimentation, and unemployment policy.*