Psychological Methods
1997, Vol. 2, No. 2, 173–185

1082-989X/97/$3.00

Statistical Analysis and Optimal Design for Cluster Randomized Trials

Stephen W. Raudenbush
Michigan State University

In many intervention studies, therapy outcome evaluations, and educational field trials, random treatment assignment of clusters rather than persons is desirable for political feasibility, logistics, or ecological validity. However, cluster randomized designs are widely regarded as lacking statistical precision. This article considers when and to what extent using a pretreatment covariate can increase experimental precision. To answer this question, the author first optimizes allocation of resources within and between clusters for the no-covariate case. Optimal sample sizes at each level depend on variation within and between clusters and on the cost of sampling at each level. Next, the author considers optimal allocation when a covariate is added. In this case, the explanatory power of the covariate at each level becomes highly relevant for choosing optimal sample sizes. A key conclusion is that statistical analysis that fully uses information about the covariate–outcome relationship can substantially increase the efficiency of the cluster randomized trial, especially when the cost of sampling clusters is high and the covariate accounts for substantial variation between clusters. Recent multilevel studies indicate that these conditions are common.

The advantages of randomized experiments in facilitating causal inference are widely recognized. Randomization probabilistically equates treatment groups on all pretreatment covariates. As a result, a confidence interval for a treatment contrast quantifies uncertainty about the magnitude of causal impact of the treatments. Such unambiguous causal inferences are not possible in quasi-experiments (intervention studies without randomization) because the confounding effects of measured or unmeasured pretreatment influences cannot be eliminated with certainty.

Research reported here was conducted with support from the Project on Human Development in Chicago Neighborhoods with funds from the MacArthur Foundation, the National Institute of Justice and the National Institute of Mental Health. An earlier version of this article was presented at the Manpower Demonstration Research Corporation's Workshop on Alternatives to Individual Random Assignment. I would like to thank Winston Lin for his extremely helpful critique of an earlier draft and Howard Bloom for providing a helpful example.

Correspondence concerning this article should be addressed to Stephen W. Raudenbush, Department of Counseling, Educational Psychology, and Special Education, Michigan State University, 461 Erickson Hall, East Lansing, Michigan 48824.

This article considers design issues that arise in cluster randomized trials, that is, experiments in which clusters of persons rather than persons themselves are assigned at random to treatments. Such clusters might be classrooms, schools, therapy groups, clinics, health maintenance organizations (HMOs), neighborhoods, program sites, or even entire cities or towns. There are many settings in which cluster randomized experiments are more feasible or desirable than experiments based on the randomized assignment of individuals to treatments. First, it may be politically or logistically impossible to assign children to preschools, patients to therapists, or residents to neighborhoods. Yet random assignment of preschools, therapists, or neighborhoods may be quite feasible. Second, even when it is possible to assign persons within clusters to treatments, it may be undesirable because of "diffusion of treatments" (Cook & Campbell, 1979). If half the teachers within a school are assigned to an in-service training program, they may share their new learning with colleagues who do not attend. Third, and perhaps most important, results of demonstration projects based on random assignment of clusters may generalize better to the policy contexts they are designed to inform. The unit of assignment and treatment—the cluster—is often the unit

of reform. After the research is completed, a preschool or a therapist will adopt a new approach for all clients, not for a randomly selected half.

Despite its appeal, researchers have often regarded cluster randomization with suspicion because it complicates the analysis and is assumed to lack statistical precision.

Data Analysis

Walsh (1947) showed that if clusters are the unit of randomization, p values based on conventional analyses will generally be too small. The greater the variability between clusters, the more severe the bias. He showed how to use information about the intracluster correlation to estimate the bias associated with conventional confidence intervals and tests. Kish (1965) related use of the intracluster correlation coefficient in cluster randomized trials to the design effect in sample surveys based on cluster sampling. Cornfield (1978) wrote: "Randomization by cluster accompanied by an analysis appropriate to randomization by individual is an exercise in self-deception" (p. 101).

Variability between clusters can arise for two reasons. First, persons are typically nonrandomly selected into the intact clusters that are the unit of randomization, and these selection effects can create intercluster variation on the outcome. Second, even if no such selection effects are present, the shared experience within a classroom, therapy group, or HMO can create intercluster outcome variation that is extraneous to the treatment effect.

Classical experimental design texts (e.g., Kirk, 1982) recommend the nested analysis of variance (ANOVA) for such designs. Unfortunately, these methods do not apply well when the data are unbalanced (unequal sample sizes per cluster) or when covariates are available. Recently, however, more appropriate analytic methods have become available under the label of *multilevel models* (Goldstein, 1987), *hierarchical linear models* (Bryk & Raudenbush, 1992), or *random coefficient models* (Longford, 1993). Raudenbush (1993) showed how these methods duplicate classical ANOVA results for a variety of designs and generalize application to the more complex data that arise in large-scale field studies.

Statistical Precision

Aversion to cluster randomized designs is based in part on a widespread perception that they lack statistical precision. When the analysis is done correctly, the standard error of the treatment contrast will typically depend more heavily on the number of clusters than on the number of participants per cluster. Studies with large numbers of clusters tend to be expensive. Blair and Higgins (1985) provided a lucid discussion of the influence of the intracluster correlation on precision and power. Murray et al. (1994) provided a thorough review of design options for coping with poor precision in trials for which whole communities are assigned to treatments.

Planning Cluster Randomized Trials

In this article, I argue that sound planning for cluster-randomized trials, based on now-standard hierarchical statistical models, can help ensure the design of affordable studies with adequate statistical precision. Such planning involves careful choice of sample sizes and evaluation of alternative designs.

Choosing Sample Sizes

In planning individualized randomized trials, a key consideration is the total sample size, which will strongly influence experimental precision. Planning of cluster randomized trials is more complex in that two sample sizes—the within-cluster sample size and the total number of clusters—contribute to experimental precision.

In many cluster-based interventions, the sample size per cluster is under the researchers' control. Examples include school-based interventions designed to prevent mental health disorders and related substance abuse (Jones, 1992; Rosenbaum, Flewelling, Bailey, Ringwalt, & Wilkinson, 1994), community-based health-promotion trials (Johnson et al., 1990; Murray et al., 1994), evaluations of residential treatment programs (Friedman & Glickman, 1987), and community-based care (Goldberg, 1994; Lehman, Slaughter, & Myers, 1992). Sampling large numbers of persons per cluster will constrain the number of clusters one can afford to sample, especially when individual-level data collection is expensive as, for example, when individual psychological assessments or biological assays are required. In these cases, I shall approach planning from the perspective of optimal allocation: Choosing the optimal within-cluster sample size is a prelude to deciding on the total number of clusters.

In other cases, cluster size will be immutable as, for example, when married couples are assigned at random to alternative approaches of family therapy. Whether the within-cluster sample size is determined by optimal allocation or by nature, both the within- and between-cluster sample sizes will contribute to precision, but the relative importance of these two sample sizes will depend on whether covariates are used.

Use of Covariates

It is widely known that the precision of an experiment using individual random assignment can be substantially increased by using a pretreatment covariate. The added precision depends on the explanatory power of the covariate. For example, Porter and Raudenbush (1987) gave an example in which adding a single covariate has the same effect on precision as doubling the sample size. In cluster randomized trials, assessing the added value of the covariate is more complex because it depends on the explanatory power of the covariate, size of the variance components, and costs of sampling at each level. An analysis of optimal allocation of resources may reveal that the optimal within-cluster sample size when using the covariate is quite different from the optimal sample size for the design without the covariate.

Optimal Design

In choosing between two designs (e.g., a design without a covariate as compared with a design with a covariate), one proceeds as follows. First, one assesses optimal allocation separately for each design, given assumptions about costs and variance components. Second, one computes the standard error of the treatment contrast for each design. I define the optimal design for a given set of assumptions as the design that yields the smaller standard error.[1] A design is uniformly optimal if, for all plausible assumptions about costs and variance components, it produces the smaller standard error. This approach is readily generalized to comparisons of multiple designs (e.g., using blocking vs. covariance analysis vs. no covariables). In this article, I restrict attention to two designs (i.e., with and without the covariate) for simplicity.

Optimal Allocation and Precision With No Covariate

Let us consider a simple setting in which one wishes to compare an experimental group E and a control group C on a posttreatment, continuous, interval scale outcome variable, Y. It is not possible to assign persons at random to treatments; rather J clusters, each of size n, will be assigned at random into equal size groups of $J/2$. Thus, the total sample size will be nJ with $nJ/2$ persons in E and $nJ/2$ persons in C. In addition, one has access to a prerandomization covariate, X, measured on each of the nJ subjects. A simple linear model for this scenario, ignoring the covariate, is

$$Y_{ij} = \gamma_0 + \gamma_1 S_j + u_j + e_{ij}, \qquad (1)$$

for $i = 1, \ldots, n$ subjects within $j = 1, \ldots, J$ *clusters;* S_j takes on a value of 0.5 for those in the experimental group and −0.5 for those in the control group; γ_0 is the grand mean; γ_1 is the treatment contrast defined as the mean difference between the two groups; it is typically assumed that $u_j \sim N(0, \tau^2)$, $e_{ij} \sim N(0, \sigma^2)$. The u_js are assumed independent for all j and the e_{ij}s are assumed independent for all i and j with u_j and e_{ij} independent of each other. Here τ^2 is the between-cluster variance, and σ^2 is the within-cluster variance.

Equation 1 models the dependence between observations in the same cluster via the random effect, u_j. Thus, the covariance between a pair of observations Y_{ij} and $Y_{i'j}$ (outcomes of two persons i and i' located in the same cluster j) is τ^2, and the correlation between these two observations is the intracluster correlation

$$\rho = \frac{\tau^2}{\tau^2 + \sigma^2}. \qquad (2)$$

As Equation 2 shows, the intracluster correlation is the proportion of variance in the outcome that lies between clusters.

This model assumes *conditional independence,* that is, given the random effect associated with a cluster, the responses of persons within the cluster are independent. This assumption will be unrealistic in some settings. For example, suppose that classrooms were assigned at random to experimental and control conditions, where the experimental treatment involved cooperative learning (Slavin, 1983). By design, the intervention promotes a high degree of interaction among subgroups of students within the experimental but not within the control condition. An appropriate statistical model for such data would represent the

[1] Alternatively, optimality could be defined with respect to maximizing power, but this requires specification of effect size and makes the presentation to follow a bit more complicated.

additional clustering in the experimental classrooms, for example, by use of a three-level hierarchical model (Bryk & Raudenbush, 1992, chap. 8) where cooperative learning groups within experimental classrooms represent a level in the model. Although the implications of such a model for research planning are beyond the scope of this article, I caution the reader to consider the realism of model assumptions; I consider the implications of more complex models for future work on planning later (see Final Remarks section).

In the framework of Equation 1, if $\rho = 0$, an "individual-level analysis" that ignores the clusters will give an appropriate estimate of the treatment effect and the standard error. However, if $\rho > 0$, as pointed out by Walsh (1947), such an analysis would produce a negatively biased estimate of the standard error of the treatment contrast, yielding a liberal test of significance and too-short confidence interval. Even small departures of ρ from zero can produce considerable bias.

The standard analysis for Equation 1 in the case of balanced data is the mixed, two-factor nested ANOVA (Kirk, 1982, p. 460, Table 1). Clusters are the random factor and treatments the fixed factor. Under the assumptions associated with Equation 1, this analysis gives estimates and F tests for the between-cluster variance component and the treatment contrast. The F test for treatments is the ratio of the between-treatment mean square to the between-cluster mean square; the F test for residual cluster variation within treatments is the ratio of the between-cluster mean square to the within-cluster mean square (see Table 1).

If interest is confined to the treatment contrast, a simple t test of group differences with the cluster means as the sample data will duplicate the nested ANOVA in the case of balanced data. Hopkins (1982) provided a lucid discussion of the relationship between nested analyses and aggregated analyses.

Table 1
Analysis of Variance for a Balanced, Two-Group Cluster Randomized Design Without Covariates (Case 1)

Source	*df*	*MS*	Expected *MS*
Treatment	1	$MS_{\text{treatment}}$	$n\tau^2 + \sigma^2 + nJ\gamma_1^2/4$
Clusters within treatment	$J - 2$	MS_{clusters}	$\sigma^2 + n\tau^2$
Persons within clusters	$J(n - 1)$	MS_{persons}	σ^2

Note. Estimates are as follows: $\hat{\gamma}_1 = \bar{Y}_{\cdot\cdot E} - \bar{Y}_{\cdot\cdot C}$; $\hat{\sigma}^2 = MS_{\text{persons}}$; $\hat{\tau}^2 = [MS_{\text{clusters}} - MS_{\text{persons}}]/n$. If $MS_{\text{clusters}} < MS_{\text{persons}}$, $\hat{\tau}^2 = 0$.

Some authors have recommended testing residual variation between clusters first; if such variation is found nonsignificant, the variation within and between clusters is pooled to yield a more powerful test of treatment group differences. This type of preliminary testing, however, is known to yield underestimates of uncertainty about treatments when cluster variation is found nonsignificant (Fabian, 1991). Thus, a Type II error in testing for cluster effects increases the probability of a Type I error in testing treatment effects. In the sequel, I avoid the two-step testing procedure on these grounds; retention of the null hypothesis of no cluster effects may be presumed a Type II error in most cases of practical interest.

Estimates

The restricted maximum likelihood (REML) estimate of the treatment contrast and its variance are

$$\hat{\gamma}_1 = \bar{Y}_{\cdot\cdot E} - \bar{Y}_{\cdot\cdot C}$$
$$\text{Var}(\hat{\gamma}_1) = \frac{4\Delta}{J}, \tag{3}$$

where

$$\Delta = \tau^2 + \sigma^2/n. \tag{4}$$

Thus, the treatment contrast estimate is the difference between the mean outcomes in the two groups with a variance that is efficiently estimated via REML by

$$\hat{\text{Var}}(\hat{\gamma}_1) = \frac{4 * \text{mean square between clusters}}{nJ}. \tag{5}$$

In the case of balanced data (equal ns in all clusters), estimation of Equation 1 by REML duplicates the familiar results of the nested ANOVA, as presented in Table 1 (Rao & Kleffe, 1988, pp. 35–37). When the data are unbalanced, REML estimation requires an iterative procedure because the treatment contrast estimate and the variance components estimates are mutually dependent in the case of unbalanced data (see Raudenbush, 1993, for an extensive discussion of how mixed linear regression models estimated via REML duplicate and generalize classical ANOVA procedures).

Cost of Data Collection

In planning the study, one must consider cost. In particular, T monetary units are available. It is estimated that, once one has sampled a cluster, it costs C_1 to enroll each subject, while the cost of sampling each additional cluster is C_2. Thus, the total cost of the study is

$$T = J(C_1 n + C_2). \tag{6}$$

In many settings, more complex cost functions will be needed. For example, it may be far more costly to sample experimental clusters than control clusters because of large costs associated with implementing the experimental intervention. I consider this issue later (see Final Remarks section), while using the simple cost function of Equation 6 for illustrative purposes.

Optimal Allocation

The cost constraint of Equation 6 defines J as

$$J = \frac{T}{nC_1 + C_2}. \tag{7}$$

Thus the variance of the treatment contrast (Equation 3) is expressible as a function of total resources available, the relative costs of sampling at each level, and n:

$$\text{Var}(\hat{\gamma}_1) = \frac{4(\tau^2 + \sigma^2/n) * (nC_1 + C_2)}{T}. \tag{8}$$

Minimizing Equation 8 with respect to n yields "n(optimal)," that is, the sample size n per cluster that minimizes the sampling variance of the treatment contrast:

$$n(\text{optimal}) = \frac{\sigma}{\tau} * \sqrt{\frac{C_2}{C_1}}. \tag{9}$$

This result duplicates the familiar result of Cochran (1977; see also Snijders & Bosker, 1993) for minimizing the variance of the population mean estimate in a survey using a balanced two-stage cluster sample with simple random sampling of clusters and then simple random sampling of persons within clusters. Equation 9 shows that a large n per cluster is most advisable when the variability within clusters is large in relation to the variability between them and when the cost of sampling additional clusters is large in relation to the cost of sampling persons.

To illustrate, I now set $T = 500$, $C_1 = 1$, and $\sigma^2 + \tau^2 = 1$, so that

$$J = \frac{500}{n(\text{optimal})C_1 + C_2}; \tau^2 = \rho; \sigma^2 = 1 - \rho. \tag{10}$$

Table 2 provides n(optimal), the corresponding J, and the sampling variance of the treatment contrast estimates under various assumptions about intracluster correlations and cost. Intracluster correlations ranging from small to large include .01, .05, .10, .20, and .50. The cost of sampling clusters is viewed as twice, 10 times, or 50 times that of sampling persons within clusters.

Table 2 gives the results that are expected. Small intracluster correlations and expensive sampling of clusters favor large ns. The variance estimates can be put in perspective by noting that if the treatment groups are separated by 0.30 standard deviations, a variance of 0.0225 or smaller would be needed to provide a power of roughly 0.50 to detect the treatment effect at the 5% level of significance. Table 2 therefore assigns an asterisk to those scenarios producing a sampling variance less than 0.0225 as a rough indicator of the designs producing a modicum of precision. It is clear that planners run into serious

Table 2
Optimal Sample Sizes and Corresponding Sampling Variances as a Function of the Intracluster Correlation and Cost Based on an Analysis That Ignores the Covariate

Intracluster correlation (ρ)	Cluster/person cost ratio (C_2)	n(optimal)	J	Var ($\hat{\gamma}_1$)
.01	2	14	31	.0103*
.01	10	31	12	.0138*
.01	50	70	4	.0232
.05	2	6	61	.0133*
.05	10	14	21	.0226
.05	50	31	6	.0522
.10	2	4	80	.0156*
.10	10	9	26	.0304
.10	50	21	7	.0811
.20	2	3	104	.0186*
.20	10	6	31	.0426
.20	50	14	8	.1317
.50	2	1	146	.0233
.50	10	3	38	.0693
.50	50	7	9	.2606

Note. An asterisk indicates a scenario that produces a sampling variance less than 0.0225 as an indicator of the designs producing a modicum of precision.

problems when the intracluster correlation and the cost of sampling clusters are simultaneously large. I now consider how using a covariate can improve things.

Optimal Allocation and Precision With a Covariate

I now add a covariate, X, but now use a hierarchical linear model for multilevel analysis to use all of the information in the covariate to account for variation in the outcome. The analysis may be viewed as a nested analysis of covariance with random effects of clusters and fixed effects of treatment and covariate.

Model

The model can be represented by the mixed linear regression model

$$Y_{ij} = \gamma_0 + \gamma_1 S_j + \gamma_2 X_{ij} + u_j + e_{ij}, \qquad (11)$$

where S_j again takes on a value of 0.5 for clusters in the experimental group and −0.5 for those in the control group; X_{ij} is the covariate measured at the person level; γ_1 again is the treatment contrast; γ_2 is now the regression coefficient for the person-level covariate; and it is assumed that $u_j \sim N(0, \tau^2_{y|x})$, $e_{ij} \sim N(0, \sigma^2_{y|x})$. The symbols τ^2 and σ^2 now both have the subscript $y|x$ to emphasize that both the between-cluster variance and the within-cluster variances are now residual variances conditional on the effects of the covariate X.

Estimation

An assumption of ordinary least squares regression is that the model residuals are independent. Equation 11 fails to satisfy this assumption, because the covariance between residuals of persons in the same cluster will be $\tau^2_{y|x}$. Efficient estimation requires an iterative algorithm, for example, that based on maximum likelihood.

The Appendix describes maximum likelihood estimation for the treatment contrast and derives the variance of the treatment contrast given the covariate and the variance components:

$$\mathrm{Var}\,(\hat{\gamma}_1|X) = \frac{4\Delta_{y|x}}{J}\left[1 + \frac{J(M_{..E} - M_{..C})^2/4}{\Delta_{y|x} SS_{wx}/\sigma^2_{y|x} + SS_{bx}}\right], \qquad (12)$$

where

$$\Delta_{y|x} = \tau^2_{y|x} + \sigma^2_{y|x}/n$$

$$SS_{wx} = \sum_{j=1}^{J/2}\sum_{i=1}^{n}(X_{ijE} - M_{.jE})^2 + \sum_{J/2+1}^{J}\sum_{i=1}^{n}(X_{ijC} - M_{.jC})^2 \qquad (13)$$

$$SS_{bx} = n\left[\sum_{j=1}^{J/2}(M_{.jE} - M_{..E})^2 + \sum_{J/2+1}^{J}(M_{.jC} - M_{..C})^2\right].$$

Here X_{ijE} and X_{ijC} are the covariate values of person i in cluster j of the experimental group and person i in cluster j of the control group, respectively; $M_{.jE}$ and $M_{.jC}$ are the covariate means for cluster j of the experimental group and cluster j of the control groups, respectively; and $M_{..E}$ and $M_{..C}$ are the covariate means for the experimental and control groups, respectively. Also, SS_{wx} is the pooled, within-cluster sum of squares of the covariate, and SS_{bx} is the pooled, within-treatment, between-cluster sum of squares of the covariate. Equation 12 has the same form as the variance of the treatment contrast with the covariate under the model without the covariate (Equation 1) except that (a) the variances are residual variances and (b) there is a correction factor that depends on the distance between the treatment means on the covariate. Thus, the benefit of adding the covariate is greatest when the residual variance, $\Delta_{y|x}$, is much smaller than the unconditional variance, Δ, and when the covariate means of the two treatment groups are similar.

Inferences about the treatment effect under the analysis of covariance are based on the conditional distribution of the outcome given the covariate, X, as in Equation 12. However, in planning research, the difference between covariate means cannot be known in advance. Thus, in planning research, it is natural to treat X as a random variable and to consider the variance of the treatment effect estimator averaged over possible values of the covariate. Treating X as normally distributed produces a useful substitute for Equation 12:

$$\mathrm{Var}\,(\hat{\gamma}_1|X) = \frac{4\Delta_{y|x}}{J}\left(1 + \frac{U}{V\theta + Y}\right), \qquad (14)$$

where

$$U = J(M_{..E} - M_{..C})^2/(4\Delta_x) \sim \chi^2(1)$$
$$Y = SS_{bx}/\Delta_x \sim \chi^2(J-2) \quad (15)$$
$$V = SS_{wx}/\sigma_x^2 \sim \chi^2[J(n-1)].$$

Here $\Delta_x = \tau_x^2 + \sigma_x^2/n$, where τ_x^2 is the between-cluster variance of X, and σ_x^2 is the within-cluster variance of X. Defining $\rho_x = \tau_x^2/(\tau_x^2 + \sigma_x^2)$ as the intracluster correlation on the covariate and $\rho_{y|x} = \tau_{y|x}^2/(\tau_{y|x}^2 + \sigma_{y|x}^2)$ as the residual intracluster correlation of Y (after adjusting for X), one defines

$$\theta = \frac{n\rho_{y|x} + 1 - \rho_{y|x}}{n\rho_x + 1 - \rho_x}. \quad (16)$$

Equation 14 provides a general expression that simplifies in specific interesting cases.

Case 1. X is a cluster-level covariate, so that $SS_{wx} = 0$. Then

$$\text{Var}(\hat{\gamma}_1|X) = \frac{4\Delta_{y|x}}{J}\left[1 + \frac{F(1, J-2)}{J-2}\right], \quad (17)$$

where $F(1, J-2)$ is distributed as F with 1 and $J-2$ degrees of freedom. In this case, the variance of the treatment effect over possible random samples of the covariate is

$$\text{Var}(\hat{\gamma}_1) = E[\text{Var}(\hat{\gamma}_1|X)] = \frac{4\Delta_{y|x}}{J}\left[1 + \frac{1}{J-4}\right]. \quad (18)$$

Case 2. X is a person-level covariate in a study in which persons are assigned at random to clusters and clusters to treatments, with no independent cluster effects. Here $\rho_{y|x} = \rho_x = 0$; $\theta = 1$. Then

$$\text{Var}(\hat{\gamma}_1|X) = \frac{4\sigma_{y|x}^2}{J}\left[1 + \frac{F(1, Jn-2)}{Jn-2}\right], \quad (19)$$

where $F(1, Jn-2)$ is distributed as F with 1 and $Jn-2$ degrees of freedom, so that the variance of the treatment effect over possible random samples of the covariate is

$$\text{Var}(\hat{\gamma}_1) = E[\text{Var}(\hat{\gamma}_1|X)] = \frac{4\sigma_{y|x}^2}{J}\left(1 + \frac{1}{Jn-4}\right). \quad (20)$$

Case 3. Here, $\rho_x = \rho_{y|x}$ with no other restrictions on the parameters. The assumption that the covariate and outcome have the same (or similar) intracluster correlations is probably quite realistic in many settings. Then, we have $\theta = 1$ so that

$$\text{Var}(\hat{\gamma}_1|X) = 4\frac{\Delta}{J}\left(1 + \frac{U}{V+Y}\right). \quad (21)$$

The ratio $U/(V+Y)$ will not be distributed exactly as F because, although U and Y are independent, neither is independent of V. However, the Appendix provides a very accurate approximation (to the order of J^{-5}) of the expectation of Equation 21, namely,

$$\text{Var}(\hat{\gamma}_1) = E[\text{Var}(\hat{\gamma}_1|X)]$$
$$\approx \frac{4\Delta_{y|x}^2}{J}\left\{1 + \frac{1}{Jn-2}\left[1 + \frac{1}{Jn-2}\left(1 + \frac{1}{Jn-2}\left(1 + \frac{12}{Jn-2}\right)\right)\right]\right\}. \quad (22)$$

It is easily shown that Equation 22 converges rapidly to

$$\text{Var}(\hat{\gamma}_1) = E[\text{var}(\hat{\gamma}_1|X)] \approx \frac{4\Delta_{y|x}}{J}\left(1 + \frac{1}{Jn-4}\right). \quad (23)$$

This is the expression for the variance of the treatment contrast that is used below in determining optimal allocation.

Optimal Allocation

If one substitutes for $T/(nC_1 + C_2)$ for J in Equation 23 and then minimizes with respect to n, one obtains, to a close approximation (see the Appendix), n(optimal) as the solution to

$$n(\text{optimal}) = \frac{k_2 + \sqrt{k_2^2 + (1-k_1)\left(\frac{k_2\sigma_{y|x}^2}{k_1\tau_{y|x}^2} + \frac{k_2^2}{k_1}\right)}}{1-k_1}, \quad (24)$$

where $k_1 = C_1/T$ and $k_2 = C_2/T$. Given the negligible magnitude of k_1 and k_2 as J increases, it is readily apparent that the right-hand side of Equation 24 is dominated by $\sqrt{k_2\sigma_{y|x}^2/(k_1\tau_{y|x}^2)}$, which has the same form as in Equation 9.

To illustrate, I use the same costs as before. However, computation of standard errors requires also that I make assumptions about the magnitudes of the within- and between-cluster variance components adjusted for the covariate. Such assumptions will typically be based on past research. In principle, the adjusted between-cluster variance, $\tau_{y|x}^2$, can be either smaller or larger than the unadjusted between-cluster variance, τ^2. This adjusted variance will be larger than the unadjusted variance when the direction of the association between the covariate and the outcome is different at the two levels. At least in educational research, however, experience shows that prior academic attainment and social background are posi-

tively related to valued educational outcomes at both levels. In fact, the relationship between such covariates and the outcome tends to be considerably stronger between clusters than within clusters. In Scotland, prior academic achievement accounts for half the variation in educational attainment within neighborhoods and over 90% of the variation between neighborhoods (Garner & Raudenbush, 1991). In U.S. high schools, socioeconomic status explains only about 7% of the variation within schools but over half the variation between schools (Bryk & Raudenbush, 1992, chap. 4). Prior research can be vital in choosing covariates to maximize the benefits of the multilevel covariance analysis.

Recently, Bloom (1995) has examined past findings from a large-scale evaluation of adult basic education programs to facilitate planning of a cluster randomized trial. He found that the covariate, pretested cognitive achievement, accounted for 73% of the variation between clusters and 48% of the variation within clusters in posttested educational achievement, a finding that appears representative of prior research using cognitive outcomes. Under the same assumptions regarding cost and the unconditional variances as before, but assuming the covariate has the same explanatory power as found by Bloom, one can compute the optimal n, corresponding J, and the variance of the treatment effect estimate. These are presented in Table 3. Note that the optimal n is larger and the corresponding J is smaller when the covariate is used (Table 3) than when the covariate is not used (Table 2). This occurs because the explanatory power is greater between clusters than within clusters, so that the adjusted intracluster correlation is smaller than the unadjusted intracluster correlation. Most important, note the substantial reduction in the variance of the treatment contrast when the covariate is used. As a result, the analysis using the covariate significantly increases the probability of detecting a non-zero treatment effect.

Table 3
Optimal Sample Sizes and Corresponding Sampling Variances as a Function of the Intracluster Correlation and Cost Based on an Analysis With a Covariate

Unconditional intracluster correlation (ρ)	Cluster/person cost ratio (C_2)	n(optimal)	J	Var ($\hat{\gamma}_1$)
.01	2	19	23	.0050*
.01	10	43	9	.0062*
.01	50	97	4	.0094*
.05	2	9	48	.0060*
.05	10	19	17	.0091*
.05	50	43	5	.0186*
.10	2	6	64	.0067*
.10	10	13	22	.0116*
.10	50	29	6	.0274
.20	2	4	85	.0076*
.20	10	9	27	.0152*
.20	50	20	7	.0422
.50	2	2	126	.0085*
.50	10	4	35	.0225*
.50	50	10	8	.0784

Note. This table is constructed with Equations 23 and 24 on the basis of prior research of Bloom (1995) showing that a prior measure of cognitive ability accounted for 73% of the between-cluster variation and 48% of the within-cluster variation in an educational achievement posttest. Appropriate assumptions about the explanatory power of the covariate at each level will be case specific. An asterisk indicates a scenario that produces a sampling variance less than 0.0225 as an indicator of the designs producing a modicum of precision.

Optimal Design

Figure 1 graphically illustrates the concept of optimal design. The figure plots the standard error of the treatment contrast as a function of plausible values of the costs and the variance components. (The standard errors in the figure are based on Bloom's estimates of the total unadjusted variance of 1579). For all plausible values of cost and variance, the standard error associated with the no-covariate analysis exceeds that associated with the covariance analysis. Thus the covariance analysis is uniformly optimal. Of course, this conclusion is based on the assumptions concerning the explanatory power of the covariate at each level.

Relative Efficiency of the Two Designs

One defines the relative efficiency of two unbiased estimators to be the ratio of the reciprocal of their variances. Table 4 gives the relative efficiency of the analysis without the covariate in comparison with the analysis with the covariate (under Bloom's, 1995, estimates of variance components). In each case, the variance is computed at n(optimal) and the corresponding J.

Table 4 shows that the larger the unconditional intracluster correlation and the more expensive it is to sample clusters in relation to persons within clusters, the greater the benefit of using the covariance analysis as opposed to the analysis with no covariate. This will generally be the case when the covariate more effectively explains variation between clusters than within

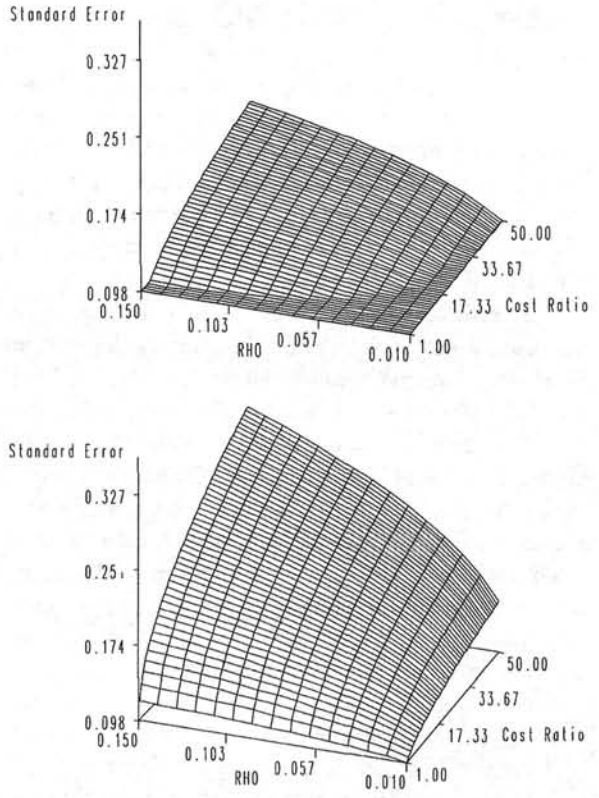

Figure 1. Standard error of treatment contrast: With (Top) and without (Bottom) the covariate. This figure is constructed on the basis of prior research of Bloom (1995) showing that a prior measure of cognitive ability accounted for 73% of the between-cluster variation and 48% of the within-cluster variation in an educational achievement posttest. Appropriate assumptions about the explanatory power of the covariate at each level will be case specific.

clusters as when using cognitive outcomes where schools or neighborhoods are the clusters. In other cases, of course, this relationship may not hold, and the relative efficiencies will not have the pattern of Table 4.

Use of an Aggregated or Between-Cluster Covariate

Suppose that an analyst uses the aggregated outcome and the aggregated covariate to estimate the treatment effect as in Equation 17, *Case 1*. One sees that the efficiency of the aggregated analysis relative to that of the two-level analysis is

$$\frac{1+\dfrac{1}{Jn-4}}{1+\dfrac{1}{J-4}} \approx \frac{J-3}{J-4}, \tag{25}$$

Table 4
Efficiency of the Design That Ignores the Covariate Relative to That of the Design That Uses the Covariate in Two-Level Analysis

Unconditional intracluster correlation (ρ)	Cluster/person cost ratio (C_2)	Relative efficiency
.01	2	.481
.01	10	.450
.01	50	.405
.05	2	.449
.05	10	.404
.05	50	.356
.10	2	.430
.10	10	.381
.10	50	.337
.20	2	.406
.20	10	.358
.20	50	.320
.50	2	.364
.50	10	.325
.50	50	.301

Note. This table is constructed with Tables 2 and 3 on the basis of prior research of Bloom (1995) showing that a prior measure of cognitive ability accounted for 73% of the between-cluster variation and 48% of the within-cluster variation in an educational achievement posttest. Appropriate assumptions about the explanatory power of the covariate at each level will be case specific.

revealing that, while the aggregated analysis can be quite inefficient when J is very small, as is common in community trials (Murray et al., 1994), it becomes nearly efficient as J becomes large.

Contextual Effects Models

The model of Equation 11 is founded on the assumption that only one parameter is needed to represent the relationship between the covariate and the outcome. Researchers often find, however, that the within-cluster and between-cluster relationships between the covariate and outcome can be quite different (e.g., Willms, 1986). Then the needed model is

$$\begin{aligned} Y_{ij} &= \gamma_0 + \gamma_1 S_j + \gamma_2 X_{ij} + \gamma_3 M_{.j} + u_j + e_{ij} \\ &= \gamma_0 + \gamma_1 S_j + \gamma_2 M_{.j} + \gamma_2 (X_{ij} - M_{.j}) \\ &\quad + \gamma_3 M_{.j} + u_j + e_{ij} \\ &= \gamma_0 + \gamma_1 S_j + \gamma_2 (X_{ij} - M_{.j}) + (\gamma_2 + \gamma_3) \\ &\quad M_{.j} + u_j + e_{ij}. \end{aligned} \tag{26}$$

Here γ_2 is the within-cluster coefficient, γ_3 is the contextual coefficient, and $\gamma_2 + \gamma_3$ is the between-cluster coefficient. Clearly, if γ_3 is zero, the within-

and between-cluster coefficients are equal, as is assumed by the basic model (Equation 11). Now averaging the outcome within clusters gives

$$\bar{Y}_{.j} = \gamma_0 + \gamma_1 S_j + (\gamma_2 + \gamma_3)M_{.j} + u_j + \bar{e}_{.j}, \quad (27)$$

where the treatment effect will be adjusted for the between-cluster coefficient, $\gamma_2 + \gamma_3$. Thus, the aggregated analysis (Equation 27) will give the same treatment effect estimate as will the two-level contextual effects model (Equation 26). Given relatively large within-cluster sample sizes, this analysis will be more efficient than the analysis based on Equation 11 when

$$R > \frac{1}{J-3}, \quad (28)$$

where R is the proportion reduction in between-cluster variance associated with estimating $\gamma_2 + \gamma_3$ rather than just γ_2. A more refined assessment of the benefit of incorporating the contextual effect as a means of improving estimation of the treatment effect is a useful topic of further research.

Final Remarks

Although cluster randomized trials, like individual randomized trials, provide unbiased estimates of program impact, they are often regarded as weak in precision. The argument presented here is that careful choice of covariates and sound planning combined with efficient analyses that use all of the information at each level can significantly increase the precision of cluster randomization studies.

I have considered the simple case of a single person-level covariate and have assumed that use of that covariate would explain variation at both levels. It is possible, of course, to use multiple person-level and cluster-level covariates in the same analysis. The number of person-level covariates that can be used is constrained by nJ, while the number of cluster-level covariates is constrained by J.

In many cases, within-cluster sample sizes are under the control of the researcher. It is then useful to determine the optimal within-cluster sample size given plausible assumptions about costs and variance components. In some cases, however, the research setting will fix the within-cluster sample size n. For example, when married couples or identical twins are randomized to therapy, $n = 2$ by definition, and in classroom research it may be logistically necessary to include every member of each classroom in the study. In these cases, of course, the optimal allocation formulas presented here will be unnecessary, but the standard error formulas will still apply with the fixed n substituted for n(optimal). When the fixed sample sizes vary, substituting the harmonic n, $n_{\text{harmonic}} = J/\sum n_j^{-1}$, will give a good approximation to the standard error. In either case, once within-cluster sample sizes are determined, overall precision will depend on the total number of clusters, which is constrained by the resources available for the study. The overall precision may be substantially enhanced by the use of a covariate.

I have also assumed that investigators will have collected some covariate information at the person level. It has become quite routine to collect at least some demographic information. However, a decision to collect effective covariates could significantly increase the cost of data collection. Collecting covariates at the person level would increase C_1, while collecting cluster-level covariates would increase C_2. Such costs will, of course, affect optimal allocation, and one might well imagine a setting where aggregate covariates would be comparatively cheap, increasing the relative efficiency of the aggregated analysis of covariance.

In many cases, only very rough estimates of cost will be available. In this case, it would be wise to compute n (optimal) across a range of plausible cost assumptions and, in each case, to compute the relevant J and standard error (as in Tables 2–3 and Figure 1). It will often be found that the optimal design and standard error will be quite insensitive to variation in plausible assumptions about cost.

There are cases in which the optimal design will be quite unbalanced. Suppose, as described earlier, that the treatment increases social interactions within experimental clusters, creating a new source of dependence within them. In this case, within-classroom variance may be larger in the experimental classrooms than in the control classrooms, implying the need for larger sample sizes within experimental clusters than within control clusters. Also described earlier was a setting in which it is more costly to sample experimental clusters than to sample control clusters, implying an optimal design having more control clusters than experimental clusters. While it is beyond the scope of this article to consider more complex variance and cost functions leading to optimal designs that are unbalanced, the basic approach outlined here can and should be generalized to apply to these circumstances.

It is also important to extend the kind of analysis used here to other designs. For example, one might consider designs involving multiple pretests and fol-

low-ups of persons nested within clusters assigned at random to treatments. In these designs, measures of change (e.g., growth rates or acceleration rates) become the outcomes. Between-cluster variances may be much smaller than for cross-sectional status measures, thus increasing the efficiency of the cluster randomized trial in a setting where it is expensive to sample large numbers of clusters. However, longitudinal follow-up of persons may be expensive. A variant involves repeated cross-sectional samples from the same clusters before and after implementation of the treatment (Feldman & McKinlay, 1994; Murray et al., 1994). The relevant error variation for assessing program effects is the within-cluster variation over time rather than the between-cluster variation, possibly leading to gains in precision when sampling large numbers of clusters is expensive.

Sound planning of cluster randomized trials requires collection of data on variance components and costs of sampling at each level. The levels may involve variation between time points within persons, variation between persons within clusters, variation in cluster means over time, or variation between cluster means cross-sectionally. The levels that are relevant depend on the design options under consideration. Next, the most efficient statistical analysis for each given design option must be chosen. It is then possible to estimate the optimal allocation of resources and power for competing design alternatives in order to make best use of the resources available for the study.

References

Blair, R. C., & Higgins, J. (1985). A comparison of the power of the paired samples rank transform statistic to that of Wilcoxon's signed ranks statistics. *Journal of Educational Statistics, 10,* 368–383.

Bloom, H. (1995, November). *Minimum detectable effects in a cluster randomized experiment.* Presentation at the Manpower Development Research Corporation, New York.

Bryk, A., & Raudenbush, S. (1992). *Hierarchical linear models in social and behavioral research: Applications and data analysis methods.* Newbury Park, CA: Sage.

Cochran, W. (1977). *Sampling techniques* (3rd ed.). New York: Wiley.

Cook, T., & Campbell, D. (1979). *Quasi-experimentation.* New York: Rand McNally.

Cornfield, J. (1978). Randomization by group: A formal analysis. *American Journal of Epidemiology, 108*(2), 100–102.

Fabian, V. (1991). On the problem of interactions in the analysis of variance. *Journal of the American Statistical Association, 86*(414), 362–367.

Feldman, H. A., & McKinlay, S. M. (1994). Cohort versus cross-sectional design in large field trials: Precision, sample size, and a unifying model. *Statistics in Medicine, 13,* 61–78.

Friedman, I., & Glickman, N. W. (1987). Residential program characteristics for completion of treatment by adolescent drug abusers. *Journal of Nervous and Mental Disease, 175*(7), 419–424.

Garner, C., & Raudenbush, S. (1991). Educational attainment: A multi-level analysis of the influence of pupil ability, family, school, and neighborhood. *Sociology of Education, 64*(4), 251–262.

Goldberg, D. (1994). Cost-effectiveness in the treatment of patients with schizophrenia. *Acta Psychiatrica Scandinavia, 39*(382), 89–92.

Goldstein, H. (1987). *Multilevel models in educational and social research.* London: Oxford University Press.

Hopkins, K. (1982). The unit of analysis: Group means versus individual observations. *American Educational Research Journal, 19,* 5–18.

Johnson, C. A., Pentz, M. A., Weber, M. D., Dwyer, J. H., Baer, N., MacKinnon, D. P., Hanson, W. B., & Flay, B. R. (1990). Relative effectiveness of comprehensive community programming for drug abuse prevention with high-risk and low-risk adolescents. *Journal of Consulting and Clinical Psychology, 58,* 447–456.

Jones, B. A. (1992). Collaboration: The case for indigenous community-based organization support of dropout prevention programming and implementation. *Journal of Negro Education, 61,* 496–508.

Kirk, R. (1982). *Experimental design: Procedures for the behavioral sciences (2nd ed.).* Belmont, CA: Brooks/Cole.

Kish, L. (1965). *Survey sampling.* New York: Wiley.

Lehman, A. F., Slaughter, J. G., & Meyers, C. P. (1992). Quality of life experiences of the chronically mentally ill. *Evaluation and Program Planning, 15,* 7–12.

Longford, N. (1993). *Random coefficient models.* Oxford: Clarendon Press.

Murray, D., McKinlay, S., Martin, D., Donner, A., Dwyer, J., Raudenbush, S., & Graubard, B. (1994, August). Design and analysis issues in community trials. *Evaluation Review,* pp. 493–514.

Porter, A. C., & Raudenbush, S. W. (1987). Analysis of covariance: Its model and use in psychological research. *Journal of Counseling Psychology, 34,* 383–392.

Rao, C. R., & Kleffe, J. (1988). *Estimation of variance components and applications.* Amsterdam: Elsevier.

Raudenbush, S. W. (1993). Hierarchical linear models and experimental design. In L. K. Edwards (Ed.), *Applied analysis of variance in behavioral science* (pp. 459–495). New York: Marcel Dekker.

Rosenbaum, D. P., Flewelling, R. L., Bailey, S. L., Ringwalt, C. L., & Wilkinson, D. L. (1994). Cops in the classroom: A longitudinal evaluation of drug abuse resistance education (DARE). *Journal of Research on Crime and Delinquency, 31*(1), 3–31.

Slavin, R. E. (1983). *Cooperative learning.* New York: Longman.

Snijders, T., & Bosker, R. (1993). Standard errors and sample sizes for two-level research. *Journal of Educational Statistics, 18,* 237–259.

Walsh, J. (1947). Concerning the effects of the intra-class correlation on certain significance tests. *Annals of Mathematical Statistics, 18,* 88–96.

Willms, J. D. (1986). Social class segregation and its relationship to student's examination results in Scotland. *American Sociological Review, 51,* 224–241.

Appendix

Approximations for Experimental Precision

The first aim is to derive Equation 12, the variance of the treatment contrast under the two-level model of Equation 11. The model may be formulated in matrix notation as

$$Y_j = X_j\gamma + 1_n u_j + e_j \tag{A1}$$

where

$$Y_j = (Y_{1j}, \cdots, Y_{nj})^T$$

$$e_j = (e_{ij}, \cdots, e_{nj})^T$$

$$1_n = (1, \cdots, 1)^T$$

$$X_j = \begin{pmatrix} 1 & S_j & X_{1j} \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ 1 & S_j & X_{nj} \end{pmatrix} \tag{A2}$$

$$\gamma = (\gamma_0, \gamma_1, \gamma_2)^T.$$

Thus we have, for $\sigma^2_{y|x}$, $\tau^2_{y|x}$ known,

$$\mathrm{Var}(Y_j|X) = V = \tau^2_{y|x} 1_n 1_n^T + \sigma^2_{y|x} I_n$$

$$\hat{\gamma} = \left(\sum_{j=1}^{J} X_j^T V^{-1} X_j \right)^{-1} \sum_{j=1}^{J} X_j^T V^{-1} Y_j \tag{A3}$$

$$\mathrm{Var}(\hat{\gamma}) = \left(\sum_{j=1}^{J} X_j^T V^{-1} X_j \right)^{-1}.$$

Algebraic simplification leads to

$$\mathrm{Var}\begin{pmatrix} \hat{\gamma}_1 \\ \hat{\gamma}_2 \end{pmatrix} = \left[\sigma^{-2}_{y|x} \begin{pmatrix} 0 & 0 \\ 0 & SS_{wx} \end{pmatrix} + (n\tau^2_{y|x} + \sigma^2_{y|x})^{-1} \begin{pmatrix} nJ/4 & nJd/4 \\ nJd/4 & SS_{bx} \end{pmatrix} \right]^{-1} \tag{A4}$$

where $d = M_{\cdot\cdot E} - M_{\cdot\cdot C}$. Of course, $\sigma^2_{y|x}$, $\tau^2_{y|x}$ will not be known; ML estimation of γ requires substitution of ML estimates of $\sigma^2_{y|x}$, $\tau^2_{y|x}$ in Equation A4. Computing Equation A4 yields Equation 12 for the variance of the treatment effect estimate.

The next aim is to derive the approximation of Equation 22. One begins with Equation 14, expressed as

$$\mathrm{Var}(\hat{\gamma}_1|X) = \mathrm{Var}(\hat{\gamma}_1|U, Y, V) = \frac{4\Delta_{y|x}}{J}\left(1 + \frac{U}{V\theta - Y}\right). \tag{A5}$$

The distributional assumptions for U, Y, and V are given by Equation 15. Now, taking the expectation of Equation A5 with respect to U gives

$$\mathrm{Var}(\hat{\gamma}_1|Y, V) = E_U[\mathrm{Var}(\hat{\gamma}_1|U, Y, V)] = \frac{4\Delta_{y|x}}{J}\left(1 + \frac{1}{V\theta + Y}\right). \tag{A6}$$

Next, one takes the expectation of Equation A6 with respect to Y, giving

$$\mathrm{Var}(\hat{\gamma}_1|V) = E_Y[Var(\hat{\gamma}_1|Y, V)] = \frac{4\Delta_{y|x}}{J}\left[1 + E_y\left(\frac{1}{V\theta + Y}|V\right)\right]. \tag{A7}$$

Finding an expression for the expectation in Equation A7 is difficult, but the expectation can be approximated to any degree required by expanding

$$(V\theta + Y)^{-1} = D^{-1} \tag{A8}$$

in a Taylor series, giving to the fifth order

$$E_y(D^{-1}) = D_0^{-1} + 2Y_0D_0^{-3} - 8Y_0D_0^{-4} + 12Y_0(Y_0 + 4)D_0^{-5}, \tag{A9}$$

where

$$D_0 = V\theta + Y_0, \qquad Y_0 = J - 2. \tag{A10}$$

The process must iterate one more time, taking the expectation of Equation A10 with respect to the distribution of V:

$$E_V[E_y(D^{-1})] = E_V(D_0^{-1}) + 2Y_0E_V(D_0^{-3}) - 8Y_0E_V(D_0^{-4}) + 12Y_0(Y_0 + 4)\,E_V(D_0^{-5}). \tag{A11}$$

Again, the expectations in Equation A11 are not expressible in simple form. However, expanding D_0^{-1} in a fifth-degree Taylor series, substituting the resulting expression into Equation A11 (ignoring all terms less than 0 (J^{-3})), and taking the expectation gives

$$\text{Var}\,(\hat{\gamma}_2) = \frac{4\Delta_{ytx}}{J}(1 + \epsilon), \tag{A11A}$$

where

$$\epsilon = \epsilon_0 + 2(V_0\theta^2 + Y_0)\epsilon_0^3 - 8(V_0\theta^3 + Y_0)\epsilon_0^4 + 12[V_0(V_0 + 4)\theta^4 + 2V_0Y_0\theta^2 + Y_0\,(Y_0 + 4)]\epsilon_0^5, \tag{A12}$$

where

$$\epsilon_0 = (V_0\theta + Y_0), \qquad V_0 = J(n - 1). \tag{A13}$$

Setting $\theta = 1$ in Equation A12 gives Equation 22.

Finally, one considers minimization of the variance of the treatment effect estimate based on covariance analysis (Equation 23) subject to the cost constraint (Equation 6). Direct minimization of Equation 23 after substituting Equation 7 for J produces a quartic equation that defies simple solution. However, if one approximates Equation 23 by

$$\text{Var}\,(\hat{\gamma}_1) \approx \frac{4\Delta_{ytx}}{J}\left(1 + \frac{1}{Jn - 1}\right) = \frac{4\Delta_{ytx}}{J - \frac{1}{n}} \tag{A14}$$

and minimizes Equation A14, one obtains Equation 24, an approximation that is accurate to $O(J^{-4})$.

Received March 22, 1996
Revision received September 18, 1996
Accepted December 5, 1996 ■