Psychological Methods
2000, Vol. 5, No. 2, 199–213

1082-989X/00/$5.00 DOI: 10.1037//1082-989X.5.2.199

Statistical Power and Optimal Design for Multisite Randomized Trials

Stephen W. Raudenbush
University of Michigan

Xiaofeng Liu
Michigan State University

The multisite trial, widely used in mental health research and education, enables experimenters to assess the average impact of a treatment across sites, the variance of treatment impact across sites, and the moderating effect of site characteristics on treatment efficacy. Key design decisions include the sample size per site and the number of sites. To consider power implications, this article proposes a standardized hierarchical linear model and uses rules of thumb similar to those proposed by J. Cohen (1988) for small, medium, and large effect sizes and for small, medium, and large treatment-by-site variance. Optimal allocation of resources within and between sites as a function of variance components and costs at each level are also considered. The approach generalizes to quasiexperiments with a similar structure. These ideas are illustrated with newly developed software.

In multisite experiments, persons within a site are randomly assigned to one of two or more treatments, and this process is replicated at each of many sites. Examples in mental health research include assertive case management (Bond, Miller, Krumweid, & Ward, 1988), assertive community treatment (Burns & Santos, 1995), and the national evaluation of the Robert Wood Johnson Foundation's Program on Chronic Mental Illness. A prominent example in education is the Tennessee Class Size experiment (Finn & Achilles, 1990; Mosteller, 1995), in which students within each of many schools were randomly assigned to attend large or small classes. Such experiments, known as multisite clinical trials, are common in medicine. For example, Haddow (1991) investigated the effect of cotinine-assisted intervention in pregnancy on smoking and low birthweight delivery among 2,848 women at 139 clinical sites.

The popularity of the multisite trial may be attributed partly to logistical advantages. First, it is often much easier to recruit a large sample in a short period of time for a multisite trial than for a single-site trial. Second, the design is easy to manage: The administration at each site follows the same principles as in a small trial. Third, it is typically cheaper to sample participants who are geographically clustered than to recruit participants who are widely dispersed geographically (Fuller et al., 1994).

However, multisite trials differ crucially from single-site trials in allowing estimation of site-by-treatment interaction effects. The possibility that treatment effects will vary across sites can be viewed as a bane or a blessing. For example, differences in therapist skill, knowledge, or commitment may produce site differences in therapy effect, creating extra uncertainty about the nature and magnitude of the effect of the intended treatment. Yet the variation in treatment impact can be a critical dimension of generalizability. The multisite trial enables a formal test of the generalizability of the treatment impact over the varied settings in which the treatment may ultimately be implemented if its early results prove promising. In effect, a multisite trial lays the basis for a planned "meta-analysis," to which each site contributes an independent study of treatment efficacy.

Key design decisions crucial to the planning of multisite trials include the number of participants

Stephen W. Raudenbush, School of Education, University of Michigan; Xiaofeng Liu, Department of Education Studies and Survey Research Center, Michigan State University.

Xiaofeng Liu is now at the Department of Educational Psychology, University of South Carolina.

This research was supported by a National Institute of Mental Health grant entitled "Optimal Design for Multilevel and Longitudinal Research."

Correspondence concerning this article should be addressed to Stephen W. Raudenbush, School of Education, University of Michigan, 610 East University, Ann Arbor, Michigan 48109. Electronic mail may be sent to rauden@umich.edu.

sampled per site, the number of sites, and the possibility of incorporating site-level covariates to account for site variation in treatment effects. Sampling a large number of persons per site will increase the precision of the treatment effect estimate at each site. However, if the treatment effect varies substantially over sites, having a large number of sites will be important for inferences about the average impact of the treatment. Yet the more the treatment impact varies from site to site, the less interesting the average treatment effect becomes. It therefore is important to estimate both the mean and the variance of the treatment effect across sites. It also may be important to study moderator effects: The treatment may be especially effective at certain kinds of sites. A choice of sample size of participants per site, say *n,* and of the sample size of sites, say *J,* might be adequate for estimating certain parameters (e.g., the mean and variance of the treatment effect) and inadequate for estimating other parameters (e.g., the association between type of site and expected treatment effect). In contemplating such design choices, one cannot ignore costs. It may, for example, be far more expensive to sample a new site than to sample an additional participant within a site. The problem of research design is thus considerably more complicated for the multisite experiment than for the single-site experiment.

Below, we consider the determinants of power for detecting the main effect of treatment, the treatment-by-site variance, and the moderating effect of a site characteristic on the treatment effect. We then consider the problem of design. In the case of balanced designs, choosing a design involves two sample sizes: the number of participants per site and the number of sites. But these choices are constrained by the relative cost of sampling at each level. We address this problem by adopting the strategy of optimal allocation of resources.

Optimal allocation has a venerable tradition in survey research in which the design problem involves choosing, for example, optimal cluster sizes and the optimal number of clusters in a multistage cluster design (Cochran, 1977; Kish, 1965). Psychologists have used the same principles in constructing measurement instruments for which tradeoffs arise, for example, between the number of items and number of occasions of measurement in maximizing the reliability of the test (Cleary & Linn, 1969; Marcoulides, 1997). Methodologists have recently advocated the optimal allocation strategy for all features of experimental design in psychology, including the number of participants per treatment in an experimental design, the number of replicate observations per participant, and the choice of covariates (McClelland, 1997; Allison, Allison, Faith, Paultre, & Pi-Sunyer, 1997). Our approach closely parallels optimal allocation as applied to the cluster randomized trial, in which key tradeoffs involve the number of participants per cluster and the number of clusters per treatment (Jeanpretre & Kraftsik, 1989; Overall & Dalal, 1965; Raudenbush, 1997). In the case of the multisite trial, the interplay between the variation in the treatment effect across sites and the cost of sampling at each level drives optimal allocation.

To achieve these goals and to make the results applicable over many possible applications, we construct a standardized model for the data produced in a multisite trial. The model includes a standardized effect size measure, as is now common (e.g., Cohen, 1988), but it also includes standardized measures of site-by-treatment variance and of site-level moderating effects. We propose rules of thumb for deciding whether site-by-treatment variation and moderating effects are "small," "medium," or "large," and illustrate by example how assumptions about these parameters and about cost affect optimal allocation of resources and power.

To elucidate key concepts, we restrict our attention in this article to balanced designs, continuously measured dependent variables, and the case of two treatment groups, which we label the experimental group and the control group. However, the general model and approach can readily be extended to more complex settings.

Statistical Model and Tests

A Hierarchical Linear Model

We find it convenient and illuminating to formulate the linear model for the multisite trial as a hierarchical linear model (HLM). The formulation facilitates construction of a standardized model that is useful for planning studies and extends easily to the case of unbalanced designs and continuous or discrete covariates measured on participants or sites. Following the procedure of Raudenbush (1993), we conceive the Level 1 units as participants nested within the Level 2 units, the sites. Treatment contrasts are Level 1 explanatory variables with random effects that vary over sites. For simplicity, we consider the case of a contrast between the experimental and control groups, within each site.

Level 1 model. Within site j, the outcome Y_{ij} for participant i depends on a site mean and a treatment effect according to the simple regression model

$$Y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + r_{ij}, \quad r_{ij} \sim i.i.d.N(0,\sigma^2), \quad (1)$$

where β_{0j} is the mean outcome for site j; X_{ij} is a treatment contrast, with a value of 0.5 for members of the experimental group and −0.5 for members of the control group, $i = 1, \ldots, n$; β_{1j} is thus the mean difference between outcomes of experimental and control groups within site j; and r_{ij} is a person-specific residual assumed independently and normally distributed within sites, with constant variance σ^2.

Level 2 model. Within the framework of the HLM, the coefficients at Level 1 become outcome variables at Level 2. Thus, the site mean and the site-specific treatment effect vary randomly across sites according to the model

$$\beta_{0j} = \gamma_{00} + u_{0j} \quad (2)$$

$$\beta_{1j} = \gamma_{10} + u_{1j}.$$

Here, γ_{00} is the grand mean outcome and γ_{10} is the average treatment effect; u_{0j} and u_{1j} are site-specific random effects that are independent of r_{ij} and are assumed to have a bivariate normal distribution over sites, that is,

$$\begin{pmatrix} u_{0j} \\ u_{1j} \end{pmatrix} \sim i.i.d.N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} \tau_{00} & \tau_{01} \\ \tau_{01} & \tau_{11} \end{bmatrix}\right). \quad (3)$$

Here, $\tau_{00} = Var(u_{0j})$ is the variance of the site means, $\tau_{11} = Var(u_{1j})$ is the variance of the site-specific treatment effects, and $\tau_{01} = Cov(u_{0j}, u_{1j})$ is the covariance between the site mean and the treatment effect.

Substituting the Level 2 model (Equation 2) into the Level 1 model (Equation 1) yields the combined model

$$Y_{ij} = \gamma_{00} + \gamma_{10}X_{ij} + u_{0j} + u_{1j}X_{ij} + r_{ij}. \quad (4)$$

The combined model is widely termed a "mixed model," with fixed effects γ_{00}, γ_{10}, random effects u_{0j}, u_{ij} and a within-cell residual r_{ij}. Some software packages encourage users to specify the model in its hierarchical formulation, as in Equations 1 and 2 (e.g., HLM [Raudenbush, Bryk, Cheong, & Congdon, 2000] or MLN [Rasbash, Yang, Woodhouse, & Goldstein, 1995]). Other packages require specification of the "combined" or "mixed" version of the model, as in Equation 4 (SAS Proc Mixed [Littell, Milliken, Stroup, & Wolfinger, 1996] and Mixed Reg [Hedeker & Gibbons, 1996]). Singer (1998) provides a lucid discussion of the relationship between hierarchical and mixed formulations.

Estimation

Estimation of site-specific coefficients. To clarify the logic of estimation, we begin with the data from a single site. Given a balanced design (equal sample sizes in the experimental [E] and control [C] groups), the minimum-variance, unbiased estimators of the site mean and treatment effect, respectively, are

$$\hat{\beta}_{0j} = \bar{Y}_j = \sum_{i=1}^{n} Y_{ij}/n$$

and

$$\hat{\beta}_{1j} = \bar{Y}_{Ej} - \bar{Y}_{Cj} = \sum_{i=1}^{n/2} Y_{ij}/(n/2) - \sum_{i=n/2+1}^{n} Y_{ij}/(n/2), \quad (5)$$

where the data are arranged so that the first $n/2$ participants are those in the experimental group. The sampling variances of these estimates are equally straightforward:

$$Var(\hat{\beta}_{0j}|\beta_{0j}) = \sigma^2/n, \; Var(\hat{\beta}_{1j}|\beta_{1j}) = 4\sigma^2/n. \quad (6)$$

The notation $Var(\hat{\beta}_{0j}|\beta_{0j})$ can be read as the "conditional variance of $\hat{\beta}_{0j}$ given β_{0j}," that is, the variance of the estimator of β_{0j} with its true value held constant. Of course, β_{0j} and β_{1j} do indeed vary across sites. Thus, the unconditional variances (D_{00}, D_{11}) are

$$Var(\hat{\beta}_{0j}) = D_{00} = \tau_{00} + \sigma^2/n$$

and

$$Var(\hat{\beta}_{1j}) = D_{11} = \tau_{11} + 4\sigma^2/n. \quad (7)$$

In words, the unconditional variance D_{00} of the site mean estimate is the sum of two components: the variance of the true mean and the variance of the estimate given the true mean. The conditional variance D_{11} of the treatment effect across sites has a similar structure.

Estimation of fixed effects. In an unbalanced design, with varying sample sizes across sites, the estimators of fixed effects would be precision-weighted averages (cf. Bryk & Raudenbush, 1992, chap. 3). In a balanced design, the grand mean and average treatment effects are estimated by simple averages:

$$\hat{\gamma}_{00} = \sum_{j=1}^{j} \hat{\beta}_{0j}/J,\ \hat{\gamma}_{10} = \sum_{j=1}^{j} \hat{\beta}_{1j}/J. \tag{8}$$

The sampling variances of these estimates are

$$Var(\hat{\gamma}_{00}) = D_{00}/J$$

and

$$Var(\hat{\gamma}_{10}) = D_{11}/J. \tag{9}$$

Estimation of variance components. The sample variances of the $\hat{\beta}_{0j}$ and $\hat{\beta}_{1j}$ estimate their unconditional variances (Equation 7):

$$\hat{D}_{00} = \frac{\sum_{j=1}^{j} (\hat{\beta}_{0j} - \hat{\gamma}_{00})^2}{J-1},\ \hat{D}_{11} = \frac{\sum_{j=1}^{j} (\hat{\beta}_{1j} - \hat{\gamma}_{11})^2}{J-1}, \tag{10}$$

where

$$\hat{\sigma}^2 = \frac{\sum_{j=1}^{j} \sum_{i=1}^{n} (Y_{ij} - \hat{\beta}_{0j} - \hat{\beta}_{1j}X_{ij})^2}{J(n-2)}. \tag{11}$$

Estimates of the variance components are as follows:[1]

$$\hat{\tau}_{00} = \hat{D}_{00} - \hat{\sigma}^2/n,\ \hat{\tau}_{11} = \hat{D}_{11} - 4\hat{\sigma}^2/n. \tag{12}$$

The sampling variances of these fixed effects estimated are

$$Var(\hat{\gamma}_{00}) = \frac{\tau_{00} + \sigma^2/n}{J}$$

and

$$Var(\hat{\gamma}_{10}) = \frac{\tau_{11} + 4\sigma^2/n}{J}. \tag{13}$$

Hypothesis Tests

Of key interest in the multisite trial are the average treatment effect, γ_{01}, otherwise known as the main effect of treatment, and the variance of the treatment effect, τ_{11}, otherwise known as the treatment-by-site variance.

Average treatment effect. Under the null hypothesis H_0: $\gamma_{10} = 0$, an F test can readily be constructed:

$$F(1, J-1; \lambda) = J\hat{\gamma}_{10}^2/\hat{D}_{11}, \tag{14}$$

where $F(1, J-1; \lambda)$ follows the noncentral F distribution with numerator and denominator degrees of freedom of 1 and $J - 1$, respectively, and the noncentrality parameter

$$\lambda = \frac{nJ\gamma_{10}^2}{n\tau_{11} + 4\sigma^2}. \tag{15}$$

The noncentrality parameter is closely related to the ratio of expected mean squares

$$\frac{JE(\hat{\gamma}_{10}^2)}{E(\hat{D}_{11})} = \frac{nJ\gamma_{10}^2 + n\tau_{11} + 4\sigma^2}{n\tau_{11} + 4\sigma^2} = 1 + \lambda. \tag{16}$$

Note that, under the null hypothesis $\gamma_{10} = 0$, the noncentrality parameter, λ, is 0 and the ratio of the expected mean squares is 1.0. However, under the alternative hypothesis, the noncentrality parameter exceeds 0 and the ratio of expected mean squares exceeds 1.0. The larger the value of the noncentrality parameter, the greater is the power of the test. Inspection of the noncentrality parameter suggests that unless τ_{11} is null, increasing J, the number of sites, is more crucial than increasing n, the number of participants per site. Thus, the larger the variation of the treatment impact across sites, the more essential it becomes to include a large number of sites in the experiment to obtain high power in detecting a main effect of treatment. However, if the variation across sites in the impact of the treatment is truly large, the main effect of treatment becomes a poor indicator of the importance of the treatment in any particular site. For example, reporting only a small positive main effect of treatment in the context of large treatment-by-site variation would mask the fact that the treatment, although beneficial in some sites, is harmful in others.

Variance of the treatment effect. The variance of the treatment effect can also be tested by an F test. To test the null hypothesis H_0: $\tau_{11} = 0$, we compute

[1] The estimates $\hat{D}_{00}$ and $\hat{D}_{11}$ use denominators $J - 1$ under restricted maximum likelihood. Under full maximum likelihood, these denominators are J. If $\hat{D}_{00} \leq \hat{\sigma}^2/n$, we set $\hat{\tau}_{00} = 0$. Similarly, if $\hat{D}_{11} \leq 4\ \hat{\sigma}^2/n$, we set $\hat{\tau}_{11} = 0$. The covariance τ_{01} is similarly estimated by subtraction. On occasion, this covariance estimate, in combination with the variance estimates, will produce a correlation with an absolute value exceeding 1.0. Then the τ_{01} estimate may be set to the value that corresponds to the correlation with an absolute value 1.0. Consideration of these boundary value cases, though important in data analysis, has no special bearing on planning a study, the topic of interest in this article.

$$F = \frac{n\hat{D}_{11}}{4\hat{\sigma}^2}, \qquad (17)$$

where F follows the central F distribution with $df = J - 1, J(n - 2)$. The ratio of the expectation of the numerator to the expectation of the denominator is, in this case,

$$\omega = \frac{nE(\hat{D}_{11})}{4E(\hat{\sigma}^2)} = \frac{n\tau_{11} + 4\sigma^2}{4\sigma^2} = 1 + \frac{n\tau_{11}}{4\sigma^2}. \qquad (18)$$

Once again, under the null hypothesis $\tau_{11} = 0$, that is, no treatment-by-site variance, the ratio of expected mean squares is unity. Under the alternative hypothesis H_a: $\tau_{11} > 0$, increasing n raises the power of the test somewhat more efficiently than does increasing J. This is because the power is greatly determined by the extra piece $n\tau_{11}/(4\sigma^2)$ in the ratio of the expected mean squares. The power function is related to the critical F value times the inverse of ω, the ratio of expected mean squares. The more the ratio exceeds unity, the larger the power becomes.

Standardized Model for Two Groups

There are good reasons to translate our general model for the multisite trial into a model that includes standardized effect sizes (Cohen, 1988; Glass, 1976) and their variance. First, the logic of power analysis, sample size determination, and optimal allocation of resources become clearer in the context of a standardized model because key results do not depend on the scale of the outcome variable. Second, it is straightforward to translate specific examples into the standardized framework, putting many seemingly disparate cases on a common footing. Finally, there are many cases in which investigators have little prior knowledge of what effect sizes to expect; nevertheless, within the framework of the standardized model, such investigators can examine power and sample sizes in instances of what social scientists generally have come to regard as small, medium, and large effects. The broad utility and appeal of Cohen's (1988) book on power determination is partly explainable by his creation of a standardized framework for power analysis.

In the case of the multisite trial, however, it is not enough to specify a standardized effect size for the main effect of treatment. As we have seen above, the variance of the treatment impact across sites is also of interest, not only in itself, but also in determining power and sample sizes for inferences regarding the main effect. We therefore need to extend Cohen's (1988) approach by introducing a standardized metric for treatment-by-site variance.

In the context of our two-group model for site J, let us standardize the within-treatment, within-site variance to $\sigma^2 = 1.0$. Then the treatment effect for site j becomes a standardized effect size δ_j, that is, the standardized mean difference between experimental and control groups according to the Level 1 model

$$Y_{ij} = \mu_j + \delta_j X_{ij} + r_{ij},\ r_{ij} \sim N(0, 1), \qquad (19)$$

where Y_{ij} is the outcome, standardized to have unit variance; μ_j is the standardized mean at the jth site; δ_j is the standardized treatment effect at the jth site; and r_{ij} is the standardized within-cell error with unit variance.

At Level 2, we model the variability of the means and effect sizes across sites. Thus, our Level 2 model becomes

$$\mu_j = \mu + u_{0j}, \qquad (20)$$
$$\delta_j = \delta + u_{1j},$$

where μ is the standardized grand mean; δ is the standardized main effect of treatment; $Var(u_{0j}) = \sigma^2_\mu$ is the variance of the site means; $Var(u_{1j}) = \sigma^2_\delta$ is the variance of standardized treatment effects across sites; and $Cov(u_{0j}, u_{1j}) = \sigma_{\mu\delta}$ is the covariance between the standardized site mean and the treatment effect.

Sample Sizes and Power

We now consider how the number of sites and the sample size per site relate to power in the context of the multisite trial. Power analysis, of course, requires specification of the average effect size and the variance of the effect sizes across sites. Following procedures set forth by Cohen (1988), many social scientists have adopted rules of thumb for what constitutes a small, medium, or large effect. Cohen viewed standardized effect sizes of 0.20, 0.50, and 0.80 as small, medium, and large, respectively, though such interpretations are somewhat subjective and will inevitably be study-specific. We illustrate power analysis below, using software we have developed (available on request without charge)[2] that allows specification of a range of effect sizes. For simplicity, we adhere to Cohen's rule of thumb for small, medium, and large main effects of treatment.

[2] To obtain this software, please contact Stephen W. Raudenbush by e-mail at *rauden@umich.edu.*

What, then, might constitute reasonable rules of thumb for small, medium, and large variances of the treatment effect? Again, the answer to this question must be somewhat arbitrary. However, we have tentatively settled on 0.05, 0.10, and 0.15 as small, medium, and large variances. A variance of 0.05 implies that treatment effect sizes have a standard deviation slightly in excess of 0.20; a variance of 0.10 implies a standard deviation just over 0.30; and a variance of 0.15 implies a standard deviation just under 0.40. Thus, if we viewed most effect sizes to lie between about –0.10 and 0.30, we would view the variability as small; if most lie between, say, –0.20 and 0.40, the variability would be medium; and if most lie between –0.30 and 0.50, the variability would be large. In each case, the specified range is roughly two standard deviations, implying a probability in excess of 0.68 that a site-specific standardized effect size would fall in the specified range. It is, of course, a trivial matter to redefine these rules of thumb, but the current definitions will serve our purposes of illustration in the present article.

Main Effect of Treatment

The computation of power for the average treatment effect is straightforward. As mentioned, the test statistic $F = F(1, J - 1; \lambda)$ follows a noncentral F distribution with numerator and denominator degrees of freedom equal to 1 and $J - 1$, respectively, and with the noncentrality parameter, λ (see Equation 15). Let F_0 represent the critical value of F for a chosen significance level. Then power is

$$Prob[F(1, J-1; \lambda) > F_0] = 1 - Prob[F(1, J-1; \lambda) < F_0]. \quad (21)$$

Computation of power is easy to program once the degrees of freedom and noncentrality parameter are given because the distribution function for F, that is $Prob[F(df_{\text{numerator}}, df_{\text{denominator}}; \lambda) < F_0]$ is available as a subroutine on widely used packages such as SAS or S-plus (see Appendix for the SAS code).

To illustrate the logic of power, Figure 1A graphs power for a two-tailed test at $\alpha = .05$ as a function of n (sample size per site), holding constant the number of sites at $J = 4$. The maximum n is 400 for a total sample size of 1,600. In contrast, Figure 1B holds n constant at 20 and allows J to increase to 80. The maximum sample size in Figure 1B is again 1,600. In each case, the average effect size is either small (0.20), or medium (0.50), and the evaluation of power is displayed across small, medium, and large effect size variance, according to the definitions of the previous paragraph. Several principles come clearly into view. First, power increases as the variance of the treatment effect decreases. Second, although both n and J contribute to power, J is more crucial than n. In particular, allowing J to grow without bound pushes power inexorably toward 1.0 (Figure 1B). In contrast, as n increases without bound, holding J constant, power approaches a bound less than 1.0 (Figure 1A). This will always be the case unless the variance of the treatment effect is null. Third, the importance of n in increasing power depends strongly on the variance of the treatment effect: the larger this variance component, the less important is n for increasing power.

Variance of the Treatment Effect

Computation of power for the variance of the treatment effect is again straightforward, although the test statistic, F divided by ω (Equation 18), now follows a central F distribution $F[J - 1, J(n - 2)]$. Let F_0 represent the critical value of F for a chosen significance level. Then power is given by

$$\begin{aligned} &Prob\{F > F_0\} \\ &\quad = 1 - Prob\{F/\omega < F_0/\omega\} \\ &\quad = 1 - Prob\{F[J-1, J(n-2)] < F_0/\omega\}. \quad (22) \end{aligned}$$

The key to programming the computation of power is to evaluate the cumulative distribution function for the central F distribution at the value F_0/ω.

To illustrate the determinants of power, Figure 2 parallels Figure 1 in showing the consequences of increasing n and J, but this time for the variance of the treatment effect held constant at small, medium, or large values (0.05, 0.10, and 0.15, respectively). From Figure 2A, we see that increasing n is highly consequential: As n increases without bound (with $J = 4$), power for detecting this variance approaches 1.0. In contrast, Figure 2B shows that, for fixed $n = 20$, increasing J has a somewhat more modest impact on power. Thus, the effects of n and J on power for detecting site-by-treatment variance are, roughly, the reverse of what we found in the case of the average effect of treatment.

Multisite Trials With Site Characteristics as Moderators

The results of the previous section have somewhat ironic implications for design. Given nonzero variability in treatment effects across sites, increasing the

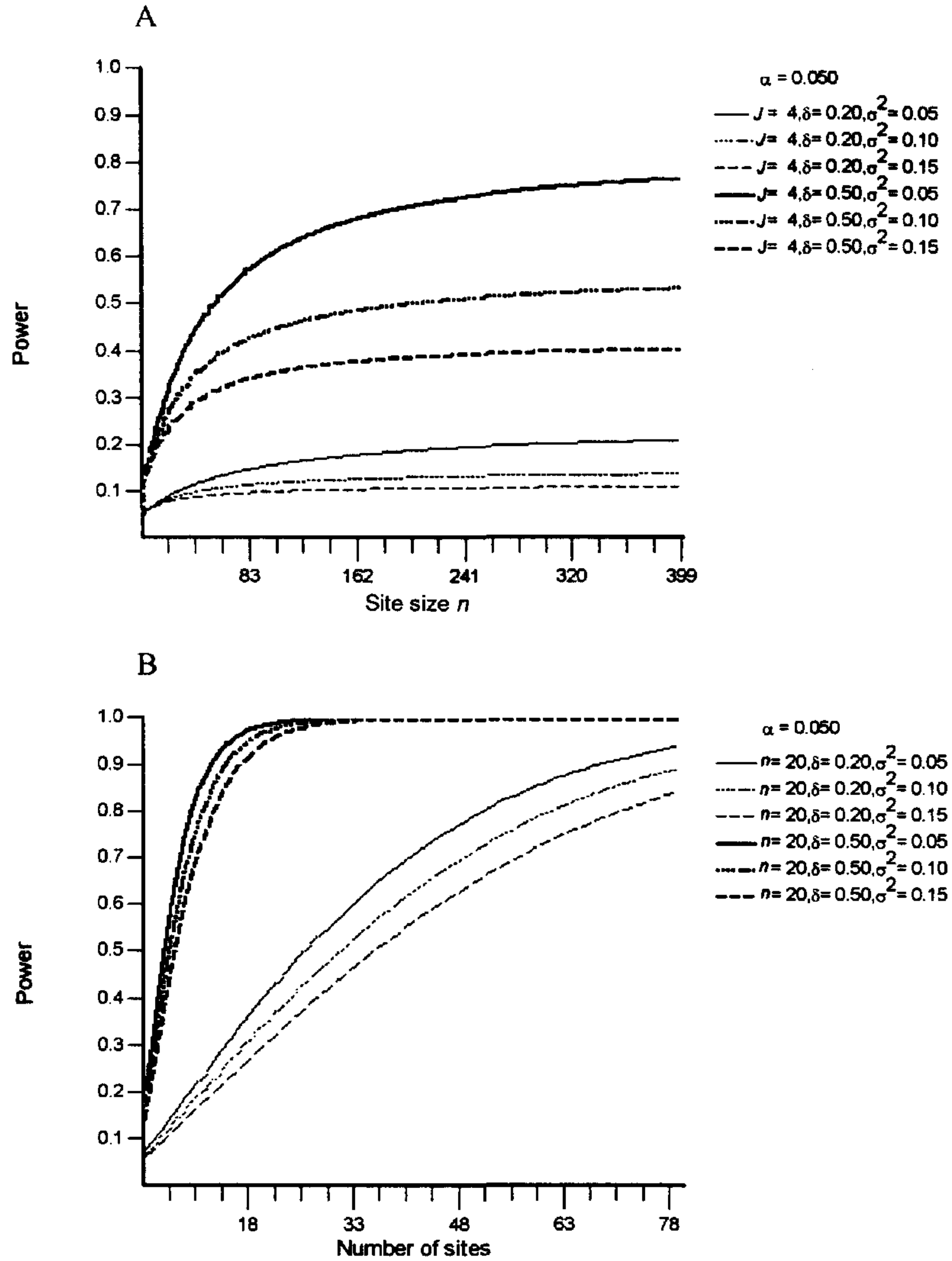

Figure 1. Power for detecting the average treatment effect as a function of effect size. A: increasing the number of participants per site (*n*), holding constant the number of sites (*J*). B: variance of the treatment effect for increasing *J*, holding constant *n*.

number of sites is essential to increase power in detecting the average effect of treatment. The larger the treatment-by-site variance, the more important it is to sample an adequate number of sites to achieve power in detecting the main effect of treatment. However, when treatment-by-site variance is large, the main effect of treatment becomes less interesting. A large treatment-by-site variance component cries out for understanding of moderating effects: characteristics of sites that can account for the large variation in the

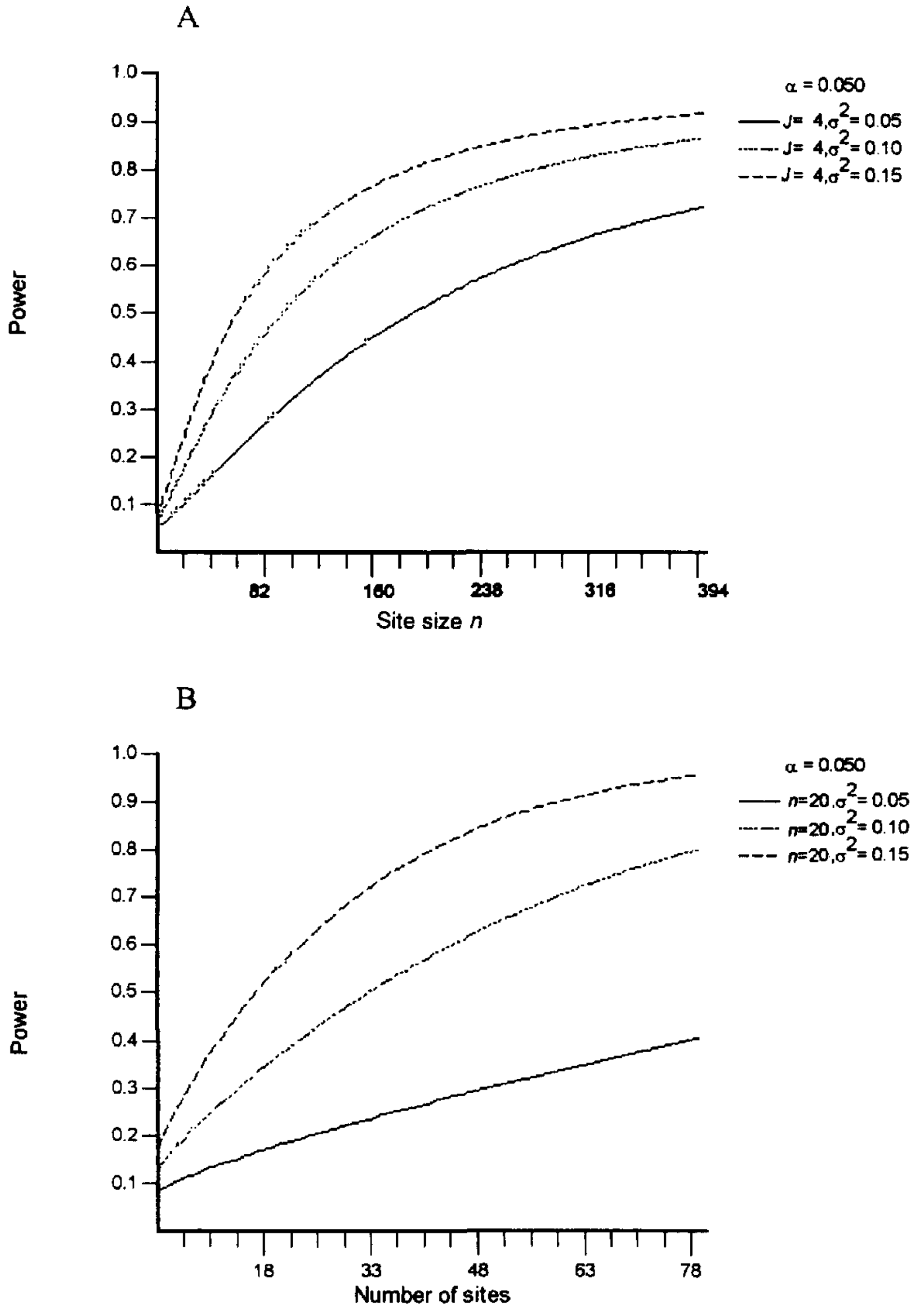

Figure 2. Power for detecting variance of the treatment effect as a function of the magnitude of the variance. A: for increasing the number of participants per site (*n*), holding constant the number of sites (*J*). B: for increasing *J*, holding constant *n*.

impact of the treatment effect. Thus, to the extent treatment effects are context-specific, it becomes scientifically important to understand the characteristics of context that account for such variation. The search for such moderators is equivalent to the search in meta-analysis for study characteristics that account for variation in findings between studies (cf. Hedges, 1994; Raudenbush, 1994).

HLM. We can readily elaborate the HLM to include site characteristics that might account for effect-size variation over sites. Because such characteristics vary at Level 2, that is, over sites, they become explanatory variables in the Level 2 model. The Level 1 model remains as before (Equation 19), whereas the Level 2 model is elaborated (in the case of a single study characteristic) to be

$$\mu_j = \gamma_{00} + \gamma_{01} W_j + u_{0j} \qquad (23)$$
$$\delta_j = \gamma_{10} + \gamma_{11} W_j + u_{1j},$$

where W_j is a site covariate with a mean of 0.0; γ_{10} is the average standardized treatment effect across sites; γ_{11} is association between the site covariate, W_j, and the effect at site *j*; $Var(u_{0j}) = \sigma_u^2$ is the residual variance of the standardized site means; $Var(u_{1j}) = \sigma_\delta^2$ is the residual variance of standardized treatment effect across sites; and $Cov(u_{0j}, u_{1j}) = \sigma_{u\delta}$ is the covariance between the standardized site mean and treatment effect, holding constant W_j.

Thus, W_j is a measured characteristic of site *j* that is hypothesized to account for variation in study effect sizes. We also include W_j in the model for site means. A failure to specify the effect of W_j on the site mean might cause a misspecification of the model for the contribution of W_j to the treatment effect (Bryk & Raudenbush, 1992, chap. 9) because the random effects (u_{0j}, u_{1j}) of the two Level 2 equations are correlated. Of interest are power and sample size requirements for inferences about γ_{11}, the moderating effect of W_j on the effect size.

Power and sample sizes. Basic principles for testing the moderating effect of the site characteristic closely parallel those for detecting the main effects of treatment. For example, consider the case in which sites are of two types, with each type consisting of *J*/2 sites. The estimate of γ_{11} is thus the difference in average treatment effect between two types of sites, namely,

$$\hat{\gamma}_{11} = \frac{\sum_{1}^{J/2} \hat{\delta}_j}{J/2} - \frac{\sum_{J/2+1}^{J} \hat{\delta}_j}{J/2}, \qquad (24)$$

where $\hat{\delta}_j$ is the estimate of standardized treatment effect at the *j*th site. Statistical inference closely parallels that for the main effect of treatment, so we omit the details. The null hypothesis H_0: $\gamma_{11} = 0$ can be tested by computing

$$F(1, J-2; \lambda) = \frac{J\hat{\gamma}_{11}^2}{4\hat{\sigma}_\delta^2}, \qquad (25)$$

where *F* follows the noncentral *F*-distribution with *df* = 1, *J* − 2, and the noncentrality parameter

$$\lambda = \frac{J\gamma_{11}^2}{4\left(\sigma_\delta^2 + \frac{4}{n}\right)}. \qquad (26)$$

Power for detecting the moderating effect of a site characteristic depends on the magnitude of the residual site-by-treatment variance. Although increasing *n* increases power, increasing *J* is more important; and this relative importance of *J* is greatest when the residual site-by-treatment variance is large. Our hope, of course, is that W_j will be a strong predictor of the effect size, such that the residual site-by-treatment variance will be small. We can, of course, test the hypothesis that this residual variance is null.

Optimal Allocation of Resources

Figures 1 and 2 display important trade-offs in designing multisite trials, trade-offs between the number of participants per site and the number of sites. We see from Figure 1 that for estimating the main effect of treatment, maximizing *J*, the number of sites, has a greater impact on power than does maximizing *n*, the number of participants per site. Testing moderating effects of site characteristics has similar implications: *J* is more important than *n* in maximizing power for detecting these moderating effects. Although these results seem to favor designs with many sites and few participants per site, such a design may be infeasible. It may be very expensive to add each additional site, whereas adding participants per site may be comparatively inexpensive. Thus, given the total resources available for the research, the number of sites that can be recruited is sharply constrained. Moreover, including a large *n* helps substantially in estimating the treatment-in-site variance. How, then, should resources be allocated within and between sites to optimize power? The answer to this question would depend on whether the aim is to make inferences about (a) the main effect of treatment, (b) the magnitude of the treatment-by-site variance, or (c) the moderating effect of site characteristics.

In each case, we consider a simple linear cost function:

$$T \geq (C_1 n + C)J, \qquad (27)$$

where *T* is the total variable cost of the study; *C* is the cost of sampling a site; and C_1 is the cost of sampling a participant within a site. Thus, the number of sites,

J, is a function of the total resources available for the study and of n, the number of participants per site:[3]

$$J \le \frac{T}{(C_1\, n + C)}. \qquad (28)$$

Our strategy is to choose "optimal n," that is, the number of participants per site that maximizes the noncentrality parameter in the power function given T and the hypothesized model parameters. J is then determined by inequality 28. Of course, in each case, we investigate optimal n, J, and power over a range of possible parameter estimates. We shall see that optimizing n for one purpose (maximizing power for the test of the main effect of treatment) will not, in general, optimize n for another purpose (e.g., estimating the variance of the treatment effect). Thus, in practice, it is necessary to weigh the relative importance of the various parameters that might be estimated and to ensure that power is at least adequate for all moderately important purposes. Let us consider each key parameter in turn.

Average Effect of Treatment

Our discussion of Figure 1 suggested that having a large J is particularly important when the site-by-treatment variance is large. However, the temptation to maximize J must be tempered by the relative cost ratio C/C_1, that is, the incremental cost of sampling a new site relative to the incremental cost of sampling a person within an already-sampled site. When we maximize power for the main effect of treatment, subject to the cost constraint (Equation 27), we see precisely how this logic plays out. The optimal n is then[4]

$$n_{opt} = 2\sqrt{\frac{C}{C_1\, \sigma_\delta^2}}. \qquad (29)$$

Equation 29 parallels the well-known formula for the optimal cluster size in a two-stage cluster sample (Cochran, 1977) and the optimal sample size per cluster in a cluster randomized trial with no covariates (cf. Allison et al., 1997; Overall & Dalal, 1965; Waters & Chester, 1987; Raudenbush, 1997). We see that optimal n is directly proportional to the square root of the cost ratio C/C_1 and inversely proportional to the square root of the treatment-by-site variance. Given optimal n, the number of sites, J, is then determined by T, the total resources for the study, that is,

$$J \le T/(C_1 n_{opt} + C). \qquad (30)$$

Consider a hypothetical example with $T = 500$ and C_1, implying that if the study were a single-site study, the investigators could afford a sample size of 500. Table 1 gives the optimal n, J, and power for varying values of the cost ratio and the variance of the treatment effect. We see from Table 1 that (a) the greater the cost of sampling sites relative to sampling participants within sites, the larger the optimal n per site, yielding fewer sites; (b) the more variable the treatment effect across sites, the smaller the optimal n, allowing more sites; and (c) a large main effect size, a small cost ratio, and a small treatment-by-site variance contribute to enhanced power for detecting the main effect of treatment. However, optimizing the study for the power of detecting the main effect of treatment does not optimize it for detecting treatment-by-site variance (compare trends in power between the last two columns of Table 1). Note that, from the medium (0.5) to the large (0.8) effect size, power is usually very high. It is less sensitive to the cost ratio and effect size variability. A design with moderate n and J is close to the optimal design in terms of power. The designs in Table 1 report higher power for the main effect of treatment than for the variance of treatment effect; powers for the variance are uniformly poor. Adequate power to detect treatment-by-site variance generally requires larger nJ than is required to detect the main effect of treatment.

Moderating Effect of a Site Characteristic

Now the question is whether a specific, measured characteristic of a site predicts the magnitude of the treatment effect at that site. In the language of the HLM, we are interested in the relationship between a Level 1 "slope" (the treatment effect at site j) and a Level 2 predictor, as described by the Level 2 model (see Equation 23). In this setting, J is more influential than n whenever the residual site-by-treatment variance is non-null. Again, however, the temptation simply to maximize J in designing the study must be tempered by C/C_1, the cost of sampling sites relative to the cost of sampling participants within sites. We therefore choose the optimal n per site that will maxi-

[3] The computed optimal n is rounded to the nearest even integer to maintain a balanced design. Also, J is the largest possible number of sites, given n, such that the total cost of the study does not exceed T.

[4] We derive Equation 29 by substituting $J = T/(C_1 n + C)$ into the formula $\lambda = nJ\delta^2/(n\sigma_\delta^2 + 4)$ and maximizing λ with respect to n.

Table 1
Optimal Number of Participants per Site (n), Number of Sites (J), and Power, as a Function of Cost Ratio (C/C_1), Treatment-by-Site Variance, and Effect Size

C/C_1	Treatment-by-site variance	Treatment main effect (δ)	Optimal *n*[a]	*J*[b]	Power for treatment main effect	Power for treatment-by-site variance
2	0.15	0.2	8	50	.405	.350
2	0.15	0.3	8	50	.732	.350
2	0.15	0.4	8	50	.930	.350
2	0.10	0.2	8	50	.433	.223
2	0.10	0.3	8	50	.766	.223
2	0.10	0.4	8	50	.947	.223
2	0.05	0.2	12	36	.470	.149
2	0.05	0.3	12	36	.807	.149
2	0.05	0.4	12	36	.965	.149
5	0.15	0.2	12	29	.322	.407
5	0.15	0.3	12	29	.612	.407
5	0.15	0.4	12	29	.849	.407
5	0.10	0.2	14	26	.352	.294
5	0.10	0.3	14	26	.658	.294
5	0.10	0.4	14	26	.884	.294
5	0.05	0.2	20	20	.397	.185
5	0.05	0.3	20	20	.721	.185
5	0.05	0.4	20	20	.924	.185
10	0.15	0.2	16	19	.257	.430
10	0.15	0.3	16	19	.499	.430
10	0.15	0.4	16	19	.741	.430
10	0.10	0.2	20	17	.294	.337
10	0.10	0.3	20	17	.564	.337
10	0.10	0.4	20	17	.807	.337
10	0.05	0.2	28	13	.327	.205
10	0.05	0.3	28	13	.619	.205
10	0.05	0.4	28	13	.854	.205
20	0.15	0.2	24	11	.187	.458
20	0.15	0.3	24	11	.359	.458
20	0.15	0.4	24	11	.567	.458
20	0.10	0.2	28	10	.210	.344
20	0.10	0.3	28	10	.405	.344
20	0.10	0.4	28	10	.629	.344
20	0.05	0.2	40	8	.244	.222
20	0.05	0.3	40	8	.472	.222
20	0.05	0.4	40	8	.708	.222

[a] The computed optimal *n* is rounded to its nearest even integer.

[b] The computed *J* is rounded. The total cost may slightly exceed the budget. For example, the seventh row has $n = 12$ and $J = 36$. The total cost will be 504. If we round the computed *J* down to 35, then the total cost will be 490. To meet the budget exactly, a researcher might add an additional site with only 8 people or use 32 sites with 12 people at each site, plus 4 sites with 10 people at each site. We therefore use rounding of *J* for computing consistency and simplicity. The provided power values should be close to the real power in those cases and can therefore be used as reference.

mize power given the cost ratio and the magnitude of the residual site-by-treatment variance. The resulting formula is identical to that given by Equation 29, keeping in mind that the variance of the treatment effects is now a residual variance, that is, the variance not explained by the moderating effect of the site characteristic. Given optimal *n*, *J* is again determined by the available resources (Equation 28).

Again let us consider a hypothetical example with $T = 500$. Suppose that sites are classified into two groups (e.g., urban sites vs. rural sites) on the basis of the hypothesis that the magnitude of the treatment effect depends on this site characteristic (e.g., urban sites are hypothesized to have smaller treatment effects than are rural sites). Table 2 gives the optimal *n*, *J*, and power for varying values of the cost ratio and the variance of the treatment effect.

We see from Table 2 that optimal *n* and *J* depend on the cost ratio and the variance of the treatment effect, just as in the average treatment effect (Table 1). Similarly, power increases as the cost ratio decreases, the variance of the treatment effect decreases, and the effect size increases. Generally, however, more data are needed to detect the moderating effect of a site characteristic than to detect the average effect of treatment, with the other factors held constant (compare power for site covariate effect in Table 2 to power for treatment-by-site variance in Table 1).

Discussion

The multisite trial enables experimenters to assess the average impact of a treatment across varied settings and the variability of the treatment impact across

Table 2
Optimal Number of Participants per Site (n), Number of Sites (J), and Power for Detecting Moderating Effect of a Site Covariance (Standardized Model)

Cost ratio (C/C_1)	Treatment-by-site variance	Optimal *n*	*J*	Effect of site covariate	Power for site covariate effect
2	0.15	8	50	0.2	.138
2	0.15	8	50	0.4	.405
2	0.15	8	50	0.6	.732
2	0.10	8	50	0.2	.146
2	0.10	8	50	0.4	.432
2	0.10	8	50	0.6	.765
2	0.05	12	36	0.2	.156
2	0.05	12	36	0.4	.470
2	0.05	12	36	0.6	.806
5	0.15	12	29	0.2	.116
5	0.15	12	29	0.4	.321
5	0.15	12	29	0.6	.611
5	0.10	14	26	0.2	.124
5	0.10	14	26	0.4	.351
5	0.10	14	26	0.6	.657
5	0.05	20	20	0.2	.135
5	0.05	20	20	0.4	.395
5	0.05	20	20	0.6	.718
10	0.15	16	19	0.2	.100
10	0.15	16	19	0.4	.256
10	0.15	16	19	0.6	.496
10	0.10	20	17	0.2	.109
10	0.10	20	17	0.4	.292
10	0.10	20	17	0.6	.561
10	0.05	28	13	0.2	.117
10	0.05	28	13	0.4	.323
10	0.05	28	13	0.6	.612
20	0.15	24	11	0.2	.083
20	0.15	24	11	0.4	.184
20	0.15	24	11	0.6	.353
20	0.10	28	10	0.2	.088
20	0.10	28	10	0.4	.205
20	0.10	28	10	0.6	.396
20	0.05	40	8	0.2	.095
20	0.05	40	8	0.4	.235
20	0.05	40	8	0.6	.453

those settings. If the treatment impact is indeed found to vary from site to site, it is typically useful to examine site characteristics that moderate the treatment effect. In this way, the multisite trial is a kind of planned meta-analysis, with each site contributing a "study" of the treatment impact, and the synthesis of findings across sites allowing for a study of the conditions under which the treatment appears most promising. It must be kept in mind that studies of the moderating effect of site characteristics on treatment effects are nonexperimenal (unless sites can be randomly assigned to characteristics). Nevertheless, the multisite trial can go well beyond the single-site trial in facilitating a study of generalizability of the treatment effect.

As this article has shown, costs and variance components drive the trade-off between maximizing the number of participants per site and maximizing the number of sites. If the goal is to maximize power in testing the average effect of treatment (or to minimize the length of the confidence interval for treatment impact), the logic is clear. The larger the variation in the treatment impact across sites, the more sites are needed to attain adequate power. However, the wish to maximize the number of sites will typically be constrained by cost: the larger the cost of sampling sites (relative to sampling participants within sites), the larger the optimal sample size per site needed to maximize power. A similar logic holds in maximizing power of tests of the moderating effect of a site characteristic. That is, adding sites is more consequential for power than is adding participants per site, and this advantage is greatest when the residual variation in the treatment impact across sites is large. Again, however, cost considerations cannot be ignored; and the desire to include many sites is tempered by the relative cost of sampling sites. This article has illustrated how these considerations can facilitate optimal design by using appropriate software.

However, the multiple purposes of a multisite trial create potential dilemmas in allocating resources. Optimizing the design to detect the main effect of treatment or the moderating effect of a site characteristic does not typically optimize the design for estimating the magnitude of the variance of the treatment effect. To estimate this variance component precisely generally requires a larger sample size per site than is optimal for the other purposes. Again, this article has illustrated software useful in determining power for this variance component.

The principles and the computations provided in this article generally extend to nonexperimental studies assessing the association between a person-level predictor and an outcome in each of many sites. For example, Raudenbush, Kidchanapanish, and Kang (1991) examined the association between preschool attendance and academic achievement of children in a nationally representative sample of elementary schools in Thailand. Of course, students were not assigned at random to attend preschool. However, it was possible to assess the mean difference between those attending and those not attending preschool in each of many schools, conceived as sites. Though causal inference is tentative in such nonexperimental studies, the power considerations are similar to those in true experiments.

In the current article we have limited our study to continuous outcomes, balanced designs, equal costs at each site (and for each treatment within each site), and the case of two treatments per site. Extensions to more general cases, including discrete outcomes, unbalanced designs, multiple treatments, and unequal costs are important. However, we can anticipate that the logic of optimal design and power determination will extend quite naturally to this broader class of cases.

References

Allison, D. B., Allison R. L., Faith, M. S., Paultre, F., & Pi-Sunyer, F. X. (1997). Power and money: Designing statistically powerful studies while minimizing financial costs. *Psychological Methods, 2,* 20–33.

Bond, G., Miller, L., Krumweid, R., & Ward, R. (1988). Assertive cases management in three CMHs: A controlled study. *Hospital and Community Psychiatry, 9,* 411–418.

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

Burns, B., & Santos, A. (1995). Assertive community treatment: An update of randomized trials. *Psychiatric Services, 47,* 669–675.

Cleary, T. A., & Linn, R. L. (1969). Error of measurement and the power of a statistical test. *British Journal of Mathematical and Statistical Psychology, 22,* 49–55.

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Finn, J. D., & Achilles, C. M. (1990). Answers and questions about class size: A statewide experiment. *American Educational Research Journal, 27,* 557–577.

Fuller, R. K., Mattson, M. E., Allen, J. P., Randall, C. L., Anton, R. F., & Babor, T. F. (1994). Multisite clinical trials in alcoholism treatment research: Organizational, methodological and management issues. *Journal of Studies on Alcohol* (Suppl. 12), 30–37.

Glass, G. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher, 5,* 3–8.

Haddow, J. (1991). Cotinine-assisted intervention in pregnancy to reduce smoking and low birthweight delivery. *British Journal of Obstetrics and Gynecology, 98,* 859–865.

Hedeker, D., & Gibbons, R. D. (1996). MIXOR: A computer program for mixed-effects ordinal regression analysis. *Computer Methods and Programs in Biomedicine, 49,* 157–176.

Hedges, L. V. (1994). Fixed effects models. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 285–300). New York: Russell Sage Foundations.

Jeanpretre, N., & Kraftsik, R. (1989). A program for the computation of power and determination of sample size in hierarchical experimental designs. *Computer Methods and Programs in Biomedicine, 29,* 179–190.

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

Littell, R. C., Milliken, G. A., Stroup, W. W. & Wolfinger, R. D. (1996). *SAS system for mixed models.* Cary, NC: SAS Institute.

Marcoulides, G. A. (1997). Optimizing measurement designs with budget constraints: The variable cost case. *Educational and Pyschological Measurement, 57,* 808–812.

McClelland, G. H. (1997). Optimal design in psychological research. *Psychological Methods, 2,* 3–19.

Mosteller, F. (1995). The Tennessee study of class size in the early school grades. *The Future of Children: Critical Issues for Children and Youth, 5,* 113–125.

Overall, J. E., & Dalal, S. N. (1965). Design of experiments to maximize power relative to cost. *Psychological Bulletin, 64,* 339–350.

Rasbash, J., Yang, M., Woodhouse, G., & Goldstein, H. (1995). *MLN: Command reference guide.* London: Institute of Education.

Raudenbush, S. (1993). Hierarchical linear models as generalizations of certain common experimental design models. In L. Edwards (Ed.), *Applied analysis of variance in behavioral science* (pp. 459–496). New York: Marcel Dekker.

Raudenbush, S. W. (1994). Random effects models. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 301–321). New York: Russell Sage Foundations.

Raudenbush, S. W. (1997). Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods, 2,* 173–185.

Raudenbush, S., Bryk, A. S., Cheong, Y., & Congdon, R. T. (2000). *HLM5: Hierarchical linear and nonlinear modeling.* Chicago: Scientific Software International.

Raudenbush, S., Kidchanapanish, S., & Kang, S. (1991). The effects of pre-primary access and quality on educational achievement in Thailand. *Comparative Education Review, 35,* 255–273.

Singer, J. D. (1998). Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models. *Journal of Educational and Behavioral Statistics, 23,* 323–355.

Waters, J. R., & Chester, A. J. (1987). Optimal allocation in multivariate, two-stage sampling design. *American Statistician, 41,* 46–50.

Appendix

SAS Program

```
/*=======================================================================*/
The sas program will produce table 1 and table 2 in text format in the files 'c:\table1.txt' and 'c:\table2.txt'
=========================================================================*/

options nodate;
data_null_;
 filename table1 'c:\table1.txt';  *power for treatment in mst;
 filename table2 'c:\table2.txt';  *power for site characteristic;

%let T=500;

array r{4} (2, 5, 10, 20);            * r is the cost ratio;
array sigma{3} (0.15, 0.10, 0.05);    * effect size variability;
array ses{3} (0.2, 0.3, 0.4);         * standardized effect size for treatment;
array ses_cov{3} (0.2, 0.4, 0.6);     * ses for site characteristic;

alpha=0.05;
do i=1 to 4;
 do m=1 to 3;
  do k=1 to 3;
   n1 =sqrt (r{i}/sigma{m});
   n1=round(n1,1);
   n=2*n1; * n now is the site size;
   J=&T/(n+r{i});
   J=round (J,1);
   lambda1= n*J*ses{k}**2/(sigma{m}*n+4);                         * equation 15;
   p1=1-probf(finv(1-alpha,1, J-1),1, J-1, lambda1);              * equation 21;
  *power for treatment main effect;
   omega=1+n*sigma{m}/4;                                           * equation 18;
   p2=1-probf(finv(1-alpha, J-1, J*(n-2))/omega, J-1, J*(n-2));   * equation 22;
  *power for treatment*site in mst;
   lambda3= n*J*ses_cov{k}**2/(4*(sigma{m}*n+4));                 * equation 26;
   p3=1-probf(finv(1-alpha,1, J-2),1, J-2, lambda3);
  *power for site characteristic;

 file table1;
  put @1 r{i} 2.0 @8 sigma{m} 3.2 @15 ses{k} 2.1 @20 n 3.0 @30 J 3.0 @40 p1 5.3 @48 p2 5.3;
 file table2;
  put @1 r{i} 2.0 @8 sigma{m} 3.2 @20 n 3.0 @30 J 3.0 @40 ses_cov{k} 2.1 @48 p3 5.3;

  end;
 end;
end;

run;
```

Received December 18, 1998
Revision received August 25, 1999
Accepted February 16, 2000 ■