Optimal Sample Allocation Under Unequal Costs in Cluster-Randomized Trials

Abstract

Conventional optimal design frameworks consider a narrow range of sampling cost structures that thereby constrict their capacity to identify the most powerful and efficient designs. We relax several constraints of previous optimal design frameworks by allowing for variable sampling costs in cluster-randomized trials. The proposed framework introduces additional design considerations and has the potential to identify designs with more statistical power, even when some parameters are constrained due to immutable practical concerns. The results also suggest that the gains in efficiency introduced through the expanded framework are fairly robust to misspecifications of the expanded cost structure and concomitant design parameters (e.g., intraclass correlation coefficient). The proposed framework is implemented in the *R* package XXX.

Key words: optimal design, design precision, cost efficiency, power analysis, cost structure

The statistical power to detect treatment effects in cluster-randomized trials is, in part, governed by how the total sample size is allocated across levels of the hierarchy and treatment conditions (Bloom, 2005; Hedges & Borenstein, 2014; Liu, 2003; Kelcey, Shen, & Spybrook, 2016; Spybrook, Shi, & Kelcey, 2016). For instance, holding constant the total sample size, designs can achieve vastly different levels of statistical power under different sampling plans (Hedges & Borenstein, 2014; Liu, 2003; Raudenbush, 1997). Equally, holding constant the total sample size, designs with different sampling plans may require dramatically different total costs because the costs of sampling a unit are not always equal across levels (Hedges & Borenstein, 2014; Raudenbush, 1997) and treatment conditions (Liu, 2003; Cochran, 1963; Nam, 1973).

As a result, an important first step in the design of such studies is to consider theoretical guidelines for sample allocation. Such guidelines have been typically derived from the conventional optimal design framework (e.g., Raudenbush, 1997). The conventional framework seeks to identify the sample allocation that produces the greatest statistical power to detect a treatment effect given a fixed budget by leveraging information regarding the marginal costs of sampling additional clusters and individuals (Hedges & Borenstein, 2014; Liu, 2003; Raudenbush, 1997). Implicit in this framework is the assumption that the costs of sampling additional control and treatment units are invariable.

However, prior theoretical and empirical work in the context of cluster-randomized trials suggests that the marginal costs potentially vary across treatment conditions and sampling levels. The potential for differences in costs of sampling a unit across levels in cluster-randomized trials has been recognized and modeled in previous literature (e.g., Hedges & Borenstein, 2014; Raudenbush, 1997). For example, in a classroom-randomized trial in which classrooms are the primary unit of randomization (e.g., Mosteller, 1995), recruiting one additional classroom is

much harder, also dramatically expensive, than sampling one additional student from dozens of students in an already sampled classroom.

The costs of sampling potentially vary between treatment conditions as well (Liu, 2003; Cochran, 1963; Nam, 1973). The marginal cost of sampling a unit in the control condition (C) includes the expenditures used to recruit and measure such a unit (e.g., business travels and work time of data collectors, incentive paid to the unit). The marginal cost of sampling a unit in the treatment condition (C^T) usually includes the same marginal cost of sampling a control unit (C) plus the marginal fees associated with the delivery and implementation of interventions to this unit (C^I; e.g., specialized training to become an intervention provider, work time of an intervention provider), or $C^T = C + C^I$. Thus, we have $C^T/C = 1 + C^I/C$. That is, the cost ratio of sampling between treatment condition (C^T/C) is potentially dependent on how expensive interventions are relative to the cost of sampling a control unit (C^I/C) or how cheap sampling a control unit is relative to the marginal cost of interventions.

There are notable examples of studies in which expenses varied across treatment conditions. Take for example the study reported by Springer et al. (2011) regarding a cluster-randomized evaluation of whether incentives in teacher performance improve student outcomes. In this study, teachers in the experimental group were eligible to receive a bonus payment of up to $15,000 per year based on their students' performance in tests. In contrast, teachers in the control condition carried on with business as usual. As a result, the costs of sampling each additional teacher in the experimental group typically exceed the cost associated with sampling an additional control teacher.

A similar example of differences in costs unfolded in the Tennessee class size experiment (Mosteller, 1995). This experiment evaluated the effects of student-teacher ratios on student

achievements (Mosteller, 1995). Students and teachers were randomly assigned to one of three treatment conditions: regular classrooms of 22 to 25 students (the control condition), small classrooms of 13 to 17 students, and regular classrooms of 22 to 25 students assisted by a paid and trained teacher aide. In this setting, classrooms staffed with an aide are likely to incur additional costs as are smaller classrooms.

Examples of differential sampling costs among treatment conditions are not limited to classrooms and schooling. This type of cost disparity often arises in health care. For example, many community health interventions include public education messaging and activities or the general promotion of novel policies (e.g., Glynn et al., 1995) and incorporate costly trainings for health care providers (e.g., Hiscock et al., 2008). In many of these instances, the nature of an intervention and its deployment incurs marginal costs above and beyond those realized in the control condition. A study of four-year smoking cessation community intervention includes activities of public education, training of health care providers, and promotion of policies to restrict the sale and use of tobacco (Glynn et al., 1995); another intervention is a three-session training program co-led by well child providers and a parenting expert (Hiscock et al., 2008). Some additional examples of costly interventions are ten days training, travels to professional development conferences (Greenleaf et al., 2011); ten, two-day on-site training sessions (Jacob, Goddard, Kim, Miller, & Goddard, 2015); four-day professional trainings with one day per month (Jayanthi, Gersten, Taylor, Smolkowski, & Dimino, 2017).

Although sampling costs plausibly vary across experimental and control conditions, empirical research suggests that such differences are predominantly found at the cluster level where the interventions are implemented (e.g., Liu, 2003; Mosteller, 1995; Springer et al., 2011).

The differences in sampling costs at the individual level, if there are any, will be relatively small comparing with the sampling costs at the cluster level.

Even the cost of sampling a unit potentially varies across levels of the design and treatment conditions, the budget functions in previous optimal design frameworks do not fully consider these variations in the cost structures of sampling, and the optimal design parameters chose to maximize the statistical power in these frameworks are also limited. For example, in the optimal design framework developed by Raudenbush (1997) for two-level cluster-randomized trials, the budget function only considers the cost variation across levels and assumes the cost of sampling one additional individual or cluster in the experimental group is equal to that in the control group. Along with the between-treatment equal cost assumption in the budget function, the Raudenbush (1997) framework optimizes the sampling ratio across levels but not between treatment conditions. Alternatively, Liu (2003) developed a framework that allows cost variation between treatment conditions. Yet, the Liu (2003) framework does not model cost variation across levels and thus optimizes the sampling ratio between treatment conditions but not across levels.

More generally, the perspectives presented in previous frameworks (e.g., Raudenbush, 1997; Liu, 2003; Connelly, 2003; Turner, Toby Prevost, & Thompson, 2004) only partially consider the potential sampling costs of a cluster-randomized trial and optimize the sample ratio either across levels or between treatment conditions. Each of these previous frameworks present a type of constrained optimization —that is, they optimize only one of the sampling ratios across levels and between treatment conditions and constrain the another one. As a result, each of these frameworks potentially return sub-optimal sampling schemes when sampling costs vary across levels of the design and treatment conditions.

In this study, we develop an optimal sampling framework that considers the potential for variation in costs across treatment conditions and levels of the hierarchy. We consider the design of two- and three-level cluster-randomized trials and organize our study as follows. We begin with a review of the literature regarding previous optimal design frameworks. We follow with the development of a more flexible optimal design framework that relaxes the typical parameter and cost constraints for two-level cluster-randomized designs and derives optimal sample allocation across levels and treatment conditions from multiple perspectives. We then extend this framework to three-level cluster-randomized trials. We follow by detailing the relative design precision and efficiency between different sample allocations, and subsequently use it to compare the results between the proposed and previous frameworks. In turn, we investigate the robustness of the proposed optimal sample scheme to the misspecification of design parameter values and cost structures. We end with a discussion.

Background

For single-level experiments in which individuals are assigned at random to experimental and control groups, prior literature has developed strategies to maximize statistical power under a fixed budget by minimizing the variance of a treatment effect (Cochran, 1963; Nam, 1973). The historical framework begins with a sample size for the experimental group (n^T) and the control group (n^C) and assumes that the costs of sampling an individual in the experimental and control groups are n^T and n^C. In turn, the total cost or budget function of the study can be described as $m = c^T n^T + cn$. Under this conventional framework, sampling is optimized in terms of power when the sampling ratio between treatment conditions under the budget function is

$$n^T/n^C = \sqrt{c/c^T}\,. \tag{1}$$

Once the optimal ratio is identified, the total sample size is a straightforward function of the available budget (through the budget function) or power (through a power formula). Equation 1 shows that the more expensive sampling an individual in the treatment condition is, the smaller the proportion of individuals that should be assigned to the treatment condition. If there is no difference in the cost of sampling between treatment conditions ($c = c^T$), the best sampling strategy is to assign an equal number of individuals to each treatment condition. Thus, a balanced design is the best one in terms of statistical power under a fixed budget if, and only if, there is no difference in the costs of sampling an additional individual between treatment conditions.

Compared to single-level experiments that only need to identify the optimal sampling ratio between treatment conditions, cluster-randomized trials need to additionally identify optimal sampling ratio across levels. Literature on the optimal sample size allocation for two-level cluster-randomized trials has separately addressed these two facets of optimal ratio in different frameworks but has not developed expressions to optimize them simultaneously in a single framework.

For example, Raudenbush (1997) developed an optimal design framework for two-level cluster-randomized trials in which there are a total number of J clusters and n individuals in each cluster. The budget function in the framework is $m = J(C_1 n + C_2)$, where C_1 and C_2 are the respective costs of sampling an additional individual and cluster regardless of which treatment condition the unit is assigned to. Rearranging the budget function, we have

$$J = \frac{m}{(C_1 n + C_2)}. \quad (2)$$

Given this budget function, the optimal sampling ratio across levels that produces the maximum power under the fixed budget by minimizing the variance of the treatment effect can be identified as

$$n = \sqrt{\frac{(1-\rho)(1-R_1^2)}{\rho(1-R_2^2)}}\sqrt{\frac{C_2}{C_1}}, \quad (3)$$

where ρ is the unconditional intraclass correlation coefficient in a population, R_1^2 and R_2^2 are the proportions of outcome variance explained by covariates at the individual and cluster levels, respectively. The cluster-level sample size J is then identified under a budget m using Equation 2 or a power formula once the optimal n in Equation 3 is given.

Equation 3 proposes two primary pathways for improving statistical power under a fixed budget. First, as the conditional variance at the cluster level is relatively large, it is more beneficial to sample fewer individuals per cluster in exchange for more clusters under a fixed budget. Second, when the sampling cost of a cluster is relatively large, the statistical power of a design can be improved by sampling more individuals per cluster in exchange for fewer clusters under a fixed budget (Raudenbush, 1997).

An implicit assumption of the conventional optimal design framework (Raudenbush, 1997) is that the cost of sampling a unit in the treatment condition is equal to that of a unit in the control condition, and only balanced designs with an equal number of clusters in each treatment condition are considered. As a result, the Raudenbush (1997) framework presents a type of constrained optimal design framework in which sample allocations are constrained to designs with an equal sample size and equal sampling costs between treatment conditions. However, such constraint are potentially incongruous with previous optimal design frameworks that recognize the potential for unequal sampling costs between treatment conditions (Cochran, 1963; Nam, 1973; Liu, 2003) and potentially restrictive in practice (e.g.,Mosteller, 1995; Greenleaf et al., 2011; Jacob et al., 2015; Springer et al., 2011) because they limit researchers abilities to

identify the sample size allocation that produces the greatest statistical power under a fixed budget.

Liu (2003) relaxed the between-treatment equal cost assumption and the constraint of balanced designs in the Raudenbush (1997) framework and shifted the optimal design in multilevel experiments back to the optimal sample size ratio between treatment conditions in single level experiments. However, by allowing sampling costs to vary between treatment conditions and considering unbalanced designs, the Liu (2003) framework omitted the optimization of sample size ratio across levels. More specifically, under this framework, a single unitary cost for sampling an additional cluster and its individuals is considered but that cost is allowed to differ by treatment condition.

For instance, presume that the combined cost of sampling an additional cluster together with its individuals in the treatment and control groups are C^T and C, respectively. The budget function is $m = (1-p)JC + pJC^T$ with p as the proportion of clusters to be assigned to the treatment condition and J as the number of total clusters. We can rearrange the budget function as

$$J = \frac{m}{(1-p)C+pC^T}. \tag{4}$$

Under this scenario Liu (2003) derived the optimal sampling ratio between treatment conditions as $\sqrt{C/C^T}$ (i.e., $(pJ)/[(1-p)J] = \sqrt{C/C^T}$), which has the same expression of Equation 1 for the single-level experiments. Thus, the optimal proportion of clusters to be assigned to the treatment condition is

$$p = \frac{\sqrt{C/C^T}}{1+\sqrt{C/C^T}}. \tag{5}$$

Although the work by Liu (2003) widened the scope and flexibility of cost structures and is consistent with earlier literature (Cochran, 1963; Nam, 1973), it did not model the cost variation across levels and retained constraints on the sample allocation across levels. Thus, the resulting framework presents a type of constrained optimal design, which often results in sub-optimal sample allocation.

The Raudenbush (1997) framework has also been extended to three-level cluster-randomized trials with the same between-treatment equal cost assumption and the balanced-design constraint (Moerbeek, van Breukelen, & Berger, 2000; Konstantopoulos, 2009, 2011; Hedges & Borenstein, 2014). Suppose K is the total number of level-three clusters, n and J are the sample sizes per level-two and level-three unit, respectively. The budget function is $m = K(nJC_1 + JC_2 + C_3)$, where C_1, C_2, and C_3 are the respective costs of sampling one additional level-one, level-two, and level-three unit. Thus, we have

$$K = \frac{m}{nJC_1 + JC_2 + C_3}. \tag{6}$$

Given the above budget function, literature has solved the optimal sample allocation across levels in a three-level cluster-randomized trial (Moerbeek, van Breukelen, & Berger, 2000; Konstantopoulos, 2009, 2011; Hedges & Borenstein, 2014) as

$$n = \sqrt{\frac{(1-\rho_2-\rho_3)(1-R_1^2)}{\rho_2(1-R_2^2)}}\sqrt{\frac{C_2}{C_1}}, \tag{7}$$

and

$$J = \sqrt{\frac{\rho_2(1-R_2^2)}{\rho_3(1-R_3^2)}}\sqrt{\frac{C_3}{C_2}}, \tag{8}$$

where ρ_2 and ρ_3 are the respective unconditional intraclass correlation coefficient at the level two and level three, and R_1^2, R_2^2, R_3^2 are the respective proportions of variance at the level one,

level two, and level three explained by covariates. These solutions can be reached in such a way that a three-level cluster-randomized trial is viewed as two-level cluster-randomized trials by omitting level-one or level-three units and then repeating the solution reported under the Raudenbush (1997) framework.

More specifically, by omitting the top-level units, a three-level cluster-randomized trial conceptually reduces to a two-level cluster-randomized trial with an (pseudo) intraclass correlation coefficient of $\rho_2/(1-\rho_3)$. By substituting $\rho_2/(1-\rho_3)$ as the ρ value into Equation 3, we can have the optimal n expression in Equation 7. Likewise, we can have J expression in Equation 8 by omitting the level-one units in a three-level cluster-randomized trial.

In summary, previous frameworks model same aspects of the sampling cost variation across treatment conditions and levels of the design in cluster-randomized trials, but none of them fully models or accounts for the cost variation across both levels and treatment conditions. That is, although previous optimal design frameworks develop strategies that improve design precision and efficiency through balancing costs and sample allocation, these frameworks are incomplete in terms of the budget function and the parameters for optimization. In the next section, we develop a more flexible optimal design framework by modeling the full cost variation and optimizing the sampling ratios across both levels and treatment conditions.

Optimal Sample Allocation in Two-Level Cluster-Randomized Trials

We first develop our framework within the context of two-level cluster-randomized trials. We begin with an assumption that sets the individual-level sample sizes to be equal between treatment conditions (i.e., $n = n^C = n^T$). Under the random assignment of clusters, such an assumption simplifies presentation, calculations, and implementation with sacrificing nugatory

gains in efficiency. However, we provide the optimal sample allocation solutions without such an assumption in the supplemental materials.

Models

Assuming a cluster-randomized design, we let the number of sampled individuals in each cluster be n, the number of total sampled clusters be J, and the proportion of clusters to be assigned to the treatment condition be p with pJ as an integer. We can estimate the treatment effect through multilevel linear models or ordinary least squares (Raudenbush & Bryk, 2002). Multilevel linear models and ordinary least squares will provide identical treatment effect estimations when the individual-level sample size in the same treatment condition does not vary across clusters. See Hedges and Hedberg (2007) and Hoover (2002) for the method of pooling the variance between treatment conditions when sample sizes are not equal between treatment conditions at the cluster level.

We present the analytic models in the format of multilevel linear models, and the individual-level model is

$$Y_{ij} = \beta_{0j} + \boldsymbol{\beta_I'} \boldsymbol{X_{ij}} + \varepsilon_{ij} \qquad \varepsilon_{ij} \sim N(0, \sigma^2_{1|}), \tag{9}$$

where Y_{ij} is the continuous outcome of individual i ($i = 1, 2, \dots, n$) in cluster j ($j = 1, 2, \dots, J$), β_{0j} is the conditional mean score of cluster j, $\boldsymbol{\beta_I} = (\beta_{I1}, \dots, \beta_{Ir})'$ is an r-length vector of individual-level regression coefficients, $\boldsymbol{X_{ij}}$ is an r-length vector of individual-level covariate values for individual i in cluster j that may vary within and across groups or only within groups, and ε_{ij} is the individual-level error term with a conditional variance $\sigma^2_{1|}$.

Similarly, the cluster-level model is

$$\beta_{0j} = \gamma_{00} + \delta T_j + \boldsymbol{\gamma_G'} \boldsymbol{Z_j} + u_{0j} \qquad u_{0j} \sim N\left(0, \sigma^2_{2|}\right), \tag{10}$$

where γ_{00} is the conditional mean across all clusters and individuals, T_j is the treatment indicator with $T_j = 1$ for clusters in the treatment group, otherwise $T_j = 0$ with δ as the treatment effect. $\boldsymbol{\gamma_G} = (\gamma_{G1}, \ldots, \gamma_{Gq})'$ is a q-length vector of cluster-level regression coefficients, $\boldsymbol{Z_j}$ is a q-length vector of cluster-level covariate values for cluster j, which could include variables measured directly at the cluster-level and/or cluster means of individual-level covariates, and u_{0j} is the random effect of cluster j with a conditional variance $\sigma^2_{2|}$. With unconditional variances at the individual and cluster levels as σ_1^2 and σ_2^2 the intraclass correlation coefficient is

$$\rho = \frac{\sigma_2^2}{\sigma_1^2+\sigma_2^2}. \quad (11)$$

If we standardize the outcome to have a variance of one in a population, the treatment effect (δ) is placed on a standardized mean difference scale and has a variance of

$$\sigma_\delta^2 = \frac{\rho(1-R_2^2)+(1-\rho)(1-R_1^2)/n}{p(1-p)J}. \quad (12)$$

When the null hypothesis is false (i.e., $\delta \neq 0$), the statistical power follows a noncentral t-distribution (Hedges & Hedberg, 2007; Liu, 2003) with the noncentrality parameter as

$$\lambda = \frac{\delta}{\sqrt{\sigma_\delta^2}} = \frac{\delta\sqrt{p(1-p)nJ}}{\sqrt{\rho(1-R_2^2)n+(1-\rho)(1-R_1^2)}}. \quad (13)$$

The statistical power at the significance level α for the two-tailed test (Hedges & Hedberg, 2007; Hoover, 2002; Donner & Klar, 2000; Rutterford, Copas, & Eldridge, 2015) is

$$P = 1 - H[c(\alpha/2,\ J-q-2), J-q-2,\ \lambda] + H[-c(\alpha/2,\ J-q-2), J-q-2,\ \lambda\ ], \quad (14)$$

where $c(\alpha/2, v)$ is the two-tailed critical value in a t-distribution with v degrees of freedom and the significance level α, and $H(x, v, \lambda)$ is the cumulative distribution function of the noncentral t-distribution with v degrees of freedom and a noncentrality parameter λ. Similarly, the

statistical power at the significance level α for the one-tailed test (Hedges & Hedberg, 2007; Hoover, 2002; Donner & Klar, 2000; Rutterford, Copas, & Eldridge, 2015) is

$$P = 1 - H[c(\alpha, J - q - 2), J - q - 2, \lambda]. \quad (15)$$

Methods

The intersection of the optimal design frameworks presented by Raudenbush (1997) and others (Liu, 2003; Cochran, 1963; Nam, 1973), with the cost structures often observed in multilevel studies (e.g., Tennessee class size experiment; Mosteller, 1995) suggest another prospect—the budget function should let the cost of sampling vary across both levels of the hierarchy and treatment conditions. For this reason, we integrate these frameworks to develop a more flexible framework with potentially more realistic cost structure. In this extended framework, we first assign c_1 as the cost of enrolling each additional individual within a cluster in the control condition and c_1^T as the cost of enrolling each additional individual within a cluster in the treatment condition. Similarly, we use c_2 as the cost of sampling each additional cluster in the control condition and c_2^T for an experimental cluster.

Thus, the budget function is $m = (1 - p)J(c_1 n + c_2) + pJ(c_1^T n + c_2^T)$. Rearranging the budget function, we have

$$J = \frac{m}{(1-p)(c_1 n + c_2) + p(c_1^T n + c_2^T)}. \quad (16)$$

Substituting J in Equation 16 to Equation 12, we can rewrite the variance of the treatment effect as

$$\sigma_\delta^2 = \frac{[\rho(1-R_2^2)n + (1-\rho)(1-R_1^2)][(1-p)(c_1 n + c_2) + p(c_1^T n + c_2^T)]}{p(1-p)nm}. \quad (17)$$

Optimal Sample Allocation

We can derive optimal sample size allocation from several different but linked perspectives, including minimizing the variance of the treatment effect under a fixed budget, minimizing the budget requested to achieve a fixed variance of the treatment effect, and maximizing the noncentrality parameter λ under a fixed budget. We will have identical results from these different perspectives. Consistent with prior frameworks, we can identify an optimal design that achieves the greatest statistical power under a fixed budget by minimizing the error variance of the treatment effect. To minimize the error variance in Equation 17, we derive its first-order derivatives with respect to p and n and set these derivatives equal to zero, yielding

$$p = \frac{\sqrt{(c_1 n + c_2)/(c_1^T n + c_2^T)}}{1 + \sqrt{(c_1 n + c_2)/(c_1^T n + c_2^T)}}, \tag{18}$$

$$n = \frac{\sqrt{(1-\rho)(1-R_1^2)}}{\sqrt{\rho(1-R_2^2)}} \sqrt{\frac{(1-p)c_2 + pc_2^T}{(1-p)c_1 + pc_1^T}}. \tag{19}$$

The above expressions can be used to identify the optimal sampling ratio across levels and treatment conditions. There are no simple closed form solutions to the roots of p and n in Equations 1818 and 19. We can numerically solve the roots by (1) substituting Equation 19 for n in Equation 18; (2) using the *uniroot* function in the R Package Stats (R Core Team, 2019) to find a root in (0, 1) that makes the difference between the right-hand and left-hand sides of the updated Equation 18 equal to or smaller than, e.g., 10^{-10}; (3) using the solved p value to have the root of n in Equation 19. We implement these solutions in the *R* package XXX (Citation masked).

Similar to the results of prior frameworks, the results indicate that the optimal p and optimal n are not a function of total budget m but rather are driven by the relative cost structure of sampling. Only the total number of clusters J is impacted by the total budget through Equation

16. The optimal p is driven by the control/treatment cost ratio of sampling a cluster and its individuals (i.e., $(c_1 n + c_2)\ /(c_1^T n + c_2^T)$), which is also influenced by the number of individuals sampled in each cluster (n). From Equation 18, we can see that a balanced design with $p = .5$ is the optimal one if, and only if, the costs of sampling a cluster and its individuals in each treatment condition are equal (i.e., $c_1 n + c_2 = c_1^T n + c_2^T$). Otherwise, the more expensive sampling a cluster and its individuals in the treatment condition is, the smaller the optimal p. That is, investigators should assign a smaller proportion of clusters to the experimental group when the cost of sampling in the treatment condition is more expensive than that in control.

The optimal n in Equation 19 is driven by two factors. The first factor is the square root of conditional variance ratio between levels (i.e., $\sqrt{\sigma_{1|}^2}\ /\sqrt{\sigma_{2|}^2} = \sqrt{(1-\rho)(1-R_1^2)}/$ $\sqrt{\rho(1-R_2^2)}$). This indicates that the larger the conditional cluster/individual variance ratio is, the smaller the resulting optimal n. It is intuitive that researchers need more clusters to identify the treatment effect with a larger conditional intraclass correlation coefficient because a larger proportion of variation at the group level requires more clusters to achieve a same level of statistical power or design precision (Hedges & Hedberg, 2007). The terms $(1-p)c_2 + pc_2^T$ and $(1-p)c_1 + pc_1^T$ can be viewed as the weighted costs of sampling one additional cluster and individual, respectively.

The second factor is the square root of the weighted sampling cost ratio between levels, with the proportion of clusters assigned to the experimental group as the weight (i.e., $\sqrt{(1-p)c_2 + pc_2^T}/\sqrt{(1-p)c_1 + pc_1^T}$). The larger the weighted cluster/individual cost ratio is, the bigger the optimal n. Put differently, when the weighted costs of sampling a cluster is more

expensive than sampling an individual, researchers should sample fewer clusters in favor of more individuals per cluster.

Constrained Optimal Sample Allocation and Relations to Previous Frameworks

There are practical considerations that may limit the use of optimal sample allocation (Hedges & Borenstein, 2014). For example, many classrooms have an upper limit of about 20 to 30 students and this may constitute a common constraint in classroom-based designs. We probe several such constraints in p and n in order to (a) delineate the conditions under which the proposed framework reduces to previous frameworks and (b) outline the flexibility of the proposed framework.

Constrained $\boldsymbol{p}$. Suppose the constrained proportion of clusters to be assigned to the treatment condition is p_0 (i.e., $p = p_0$). If we minimize the variance of the treatment effect in Equation 17 with respect to n the constrained optimal individual-level sample size has the exact same expression with Equation 19. Thus, the constrained optimal individual-level sample size can be obtained from Equation 19 along with $p = p_0$. If we let $p = .5$, $C_1 = (1-p)c_1 + pc_1^T$, and $C_2 = (1-p)c_2 + pc_2^T$, the constrained optimal individual-level sample size in Equation 19 will reduce to Equation 3, the optimal sample size expression under the Raudenbush (1997) framework.

Constrained $\boldsymbol{n}$. Suppose the constrained individual-level sample size is n_0 (i.e., $n = n_0$), minimizing the variance of the treatment effect in Equation 17 with respect to p the constrained optimal proportion has the exact same expression with Equation 18. Thus, the constrained optimal proportion can be obtained from Equation 18 along with $n = n_0$. If we let $C = c_1 n_0 + c_2$ and $C^T = c_1^T n_0 + c_2^T$, the constrained optimal proportion in Equation 18 will reduce to Equation 5, the optimal p expression under the Liu (2003) framework.

Optimal Sample Allocation in Three-Level Cluster-Randomized Trials

Similar to those for two-level cluster-randomized trials, the potential gains in design efficiency and/or statistical power in three-level cluster-randomized trials can mostly be achieved by optimizing sampling ratios between treatment conditions and among levels. We subsequently present the optimal sample allocation with the constraint of equal sample sizes at the individual and sub-cluster levels (i.e., $n = n^C = n^T$ and $J = J^C = J^T$). We provide the optimal sample allocation solutions without such a constraint as supplemental materials.

Models

Suppose a three-level cluster sampling design has a total number of *K* clusters (level-three units) with pK clusters assigned to the treatment condition, each cluster has J sub-clusters (level-two units) of size n. Let Y_{ijk} be the continuous outcome of unit i in sub-cluster j in cluster k with $i = 1, \dots, n$, $j = 1, \dots, J$, and $k = 1, \dots, K$. Let $\boldsymbol{X_{ijk}}$, $\boldsymbol{Z_{jk}}$, $\boldsymbol{W_k}$ be the vectors of covariates at the level one, level two, level three with corresponding regression coefficient vectors of $\boldsymbol{\beta_I}$, $\boldsymbol{\beta_J}$, $\boldsymbol{\beta_K}$ and lengths of r, s, and q, respectively. Similar to models for two-level cluster-randomized trials, the covariates could be variables measured at the same level or aggregated values of variables measured at a lower level.

When the sample size per (sub-)cluster do not vary across (sub-)clusters within each treatment condition, we can estimate the treatment effect using ordinary least squares or multilevel linear models (Raudenbush & Bryk, 2002). Under the multilevel formulation, the level-one model is

$$Y_{ijk} = \beta_{0jk} + \boldsymbol{\beta_I'X_{ijk}} + \varepsilon_{ijk} \qquad \varepsilon_{ijk} \sim N(0, \sigma^2_{1|}), \tag{20}$$

where β_{0jk} is the conditional mean score of sub-cluster j in cluster k and ε_{ijk} is the individual-level error term with a conditional variance $\sigma^2_{1|}$. Similarly, the level-two or sub-cluster-level model is

$$\beta_{0jk} = \gamma_{00k} + \boldsymbol{\beta}'_{\boldsymbol{j}}\boldsymbol{Z}_{\boldsymbol{jk}} + u_{0jk} \qquad u_{0jk} \sim N\left(0, \sigma^2_{2|}\right), \tag{21}$$

where γ_{00k} is the conditional mean score of cluster k, and u_{0jk} is the random effect of sub-cluster j in cluster k with a conditional variance $\sigma^2_{2|}$. The level-three or cluster-level model is

$$\gamma_{00k} = \pi_{000} + \delta T_k + \boldsymbol{\beta}'_{\boldsymbol{K}}\boldsymbol{W}_{\boldsymbol{k}} + u_{00k} \qquad u_{00k} \sim N\left(0, \sigma^2_{3|}\right), \tag{22}$$

where π_{000} is the conditional mean across all clusters, sub-clusters, and individuals, T_k is the treatment indicator with $T_k = 1$ for clusters in the experimental group and otherwise $T_k = 0$ with δ as the treatment effect, u_{00k} is the random effect of cluster k with a conditional variance $\sigma^2_{3|}$.

Let the unconditional variances at the individual-, sub-cluster-, and cluster-level be σ^2_1, σ^2_2, and σ^2_3, respectively. The total unadjusted variance is $\sigma^2_T = \sigma^2_1 + \sigma^2_2 + \sigma^2_3$. The intraclass correlation coefficient at the level two is

$$\rho_2 = \frac{\sigma^2_2}{\sigma^2_1+\sigma^2_2+\sigma^2_3} = \frac{\sigma^2_2}{\sigma^2_T}. \tag{23}$$

The intraclass correlation coefficient at the level three is

$$\rho_3 = \frac{\sigma^2_3}{\sigma^2_1+\sigma^2_2+\sigma^2_3} = \frac{\sigma^2_3}{\sigma^2_T}. \tag{24}$$

If we standardize the outcome to have a variance of one, the treatment effect (δ) is placed on a standardized mean difference scale and has a variance of

$$\sigma^2_\delta = \frac{nJ\rho_3\left(1-R^2_3\right)+n\rho_2\left(1-R^2_2\right)+(1-\rho_2-\rho_3)\left(1-R^2_1\right)}{p(1-p)nJK}, \tag{25}$$

where R^2_3, R^2_2, and R^2_1 are the proportions of outcome variance explained by covariates at the cluster, sub-cluster, and individual levels, respectively.

When the null hypothesis is false (i.e., $\delta \neq 0$), the statistical power follows a noncentral t distribution with the noncentrality parameter as

$$\lambda = \frac{\delta}{\sqrt{\sigma_\delta^2}} = \frac{\delta\sqrt{p(1-p)nJK}}{\sqrt{nJ\rho_3(1-R_3^2)+n\rho_2(1-R_2^2)+(1-\rho_2-\rho_3)(1-R_1^2)}}. \quad (26)$$

Statistical power for three-level cluster-randomized trials can be obtained by inserting the above noncentrality parameter into Equation 14 for the two-tailed test or Equation 15 for the one-tailed test with substituting J as K in the degree of freedom expression.

Optimal Sample Allocation

Suppose the respective costs of enrolling each additional level-one, level-two, and level-three unit in the control condition are c_1, c_2, and c_3, and the costs of enrolling each additional level-one, level-two, and level-three unit in the treatment condition are c_1^T, c_2^T, and c_3^T, respectively. Thus, the budget function is $m = (1-p)K(c_1nJ + c_2J + c_3) + pK(c_1^TnJ + c_2^TJ + c_3^T)$. Rearranging the budget function, we have

$$K = \frac{m}{(1-p)(c_1nJ+c_2J+c_3)+p(c_1^TnJ+c_2^TJ+c_3^T)}. \quad (27)$$

Substituting K in Equation 27 to Equation 25, we have the variance of the treatment effect as

$$\sigma_\delta^2 = \frac{nJ\rho_3(1-R_3^2)+n\rho_2(1-R_2^2)+(1-\rho_2-\rho_3)(1-R_1^2)}{p(1-p)nJ} \times \frac{(1-p)(c_1nJ+c_2J+c_3)+p(c_1^TnJ+c_2^TJ+c_3^T)}{m}. \quad (28)$$

Following similar methods of minimizing the error variance of the treatment effect, the optimal sampling plan for each parameter can then be delineated as

$$p = \frac{\sqrt{(c_3+c_2J+c_1nJ)/(c_3^T+c_2^TJ+c_1^TnJ)}}{1+\sqrt{(c_3+c_2J+c_1nJ)/(c_3^T+c_2^TJ+c_1^TnJ)}}, \quad (29)$$

$$n = \sqrt{\frac{(1-\rho_2-\rho_3)(1-R_1^2)}{\rho_3(1-R_3^2)J+\rho_2(1-R_2^2)}}\sqrt{\frac{(1-p)(c_3+c_2J)+p(c_3^T+c_2^TJ)}{(1-p)c_1J+pc_1^TJ}}, \quad (30)$$

$$J = \sqrt{\frac{n\rho_2(1-R_2^2)+(1-\rho_2-\rho_3)(1-R_1^2)}{n\rho_3(1-R_3^2)}}\sqrt{\frac{(1-p)c_3+pc_3^T}{(1-p)(c_2+c_1n)+p(c_2^T+c_1^Tn)}}. \quad (31)$$

Each of the expressions in Equations 29 to 31 identifies the optimal sampling plan when one of the parameters is malleable. When all three of these parameters are freed, there are no simple closed-form solutions. However, we can solve the multivariate partial derivatives numerically. We implement this solution in the R package XXX (Citation masked) by (1) initiating random values for n and J (e.g., sample one integer of $n \in (2, 100)$ and one integer of $J \in (2, 100)$) and calculating an initial value of p using Equation 29; (2) Updating the value of n in Equation 30 using the updated p and J; (3) Updating the value of J in Equation 31 using the updated p and n; (4) Updating the value of p in Equation 29 using the updated n and J; (5) Steps 2 to 4 form one iteration. Repeat steps 2 to 4 until each parameter converges to a specified tolerance level (e.g., $1/10^{10}$). The resulting converged values of p, n, and J in the final iteration capture the sampling plan that jointly optimizes over these parameters.

Implications

The optimal design parameters in Equations 29 to 31 provide a more flexible framework for identifying optimal sample allocations across levels and treatment conditionds. These optimal design parameters are driven by cost structure and design parameters in a similar but extended fashion with those in two-level cluster-randomized trials. These equations can also be used to improve the precision of cluster-randomized trials with additional constraints. For any given constraint, one just needs to use the relevant constraint to substitute the corresponding optimal design parameter expressions and solve the remaining equations. For example, researchers may constrain the level-one sample size per level-two unit as 20 (i.e., $n = 20$), the constrained

optimal sample allocation would be solved by using $n = 20$ to substitute Equation 30 and solving the roots of p and J from Equations 29 and 31.

Again, we can see that a balanced design with $p = 0.5$ is the optimal one if and only if the costs of sampling a cluster and its subsequent subunits in each treatment condition are equal (i.e., $c_3 + c_2 J + c_1 nJ = c_3^T + c_2^T J + c_1^T nJ$). When we additionally let $p = .5$, $C_1 = (1-p)c_1 + pc_1^T$, $C_2 = (1-p)c_2 + pc_2^T$, and $C_3 = (1-p)c_3 + pc_3^T$, the above optimal sample allocation expressed in Equations 30 and 31 reduces to solutions in previous frameworks but with different formulations (Konstantopoulos, 2009, 2011; Hedges & Borenstein, 2014).

Relative Precision and Efficiency

There are many practical reasons that may constrain the use of the optimal sampling allocation guidelines derived above. From a practical standpoint, for instance, the number of clusters available to researchers in a particular study may be below the number suggested by the formulas. In response, researchers may intentionally expend resources by sampling additional individuals within clusters in an attempt to compensate for this constraint. Similarly, from a design standpoint, we may eventually find that the parameter values used to plan a study differ from the observed values. Here, we suffer from a type of design misspecification because the proposed optimal sampling plan (based on predicted values) may prove to be sub-optimal once data have been collected. When the optimal sample allocation is not a viable option or was incorrectly identified, we can identify the specific loss of statistical precision and efficiency an alternative design presents relative to the true optimal design (based on true values). Such statistical precision and efficiency analyses help provide a sense of what constitutes efficient designs and can assist researchers in identifying designs with the most statistical precision and efficiency among the many constrained designs that may be viable.

Our analysis of statistical precision and design efficiency considers two complementary planning perspectives. In the first perspective, we consider the statistical precision as measured through the relative variance of studies in which the sampling plan is malleable, but the budget and remaining parameters are constrained to preset values. In this setting, we compare the variances of the treatment effect estimator under a sub-optimal sampling plan with that of an optimal sampling plan. Conceptually, this assessment of relative statistical precision captures the increased sampling variance incurred by using sub-optimal sample allocations. To facilitate interpretations using a common metric, we subsequently frame this analysis in terms of the minimum detectable effect size (MDES; Bloom, 1995) because the MDES is a design parameter that researchers often use in planning studies.

In the second perspective, we consider the relative efficiency of designs in terms of study cost such that the total budget is now free, but the effect size, statistical power and other parameters are fixed. Under this approach, we detail the total additional cost a study under sub-optimal sampling would require to achieve an error variance comparable to a study that used optimal sampling. Conceptually, this evaluation quantifies the increased resources required to carry out sub-optimal designs.

For the first perspective, the relative precision (RP) is

$$RP = \frac{\sigma_{\delta o}^2}{\sigma_{\delta}^2}, \tag{32}$$

where $\sigma_{\delta o}^2$ is the smallest possible variance of the treatment effect a type of trial can achieve under a fixed budget and σ_{δ}^2 is the variance of the treatment effect an alternative and sub-optimal design can achieve under the same budget. The values of RP range from 0 to 1, with the RP approaching one when a sub-optimal design achieves a precision level near the optimal design benchmark.

For the second perspective, we can define the relative cost efficiency (RCE) as

$$RCE = \frac{m^o}{m}, \tag{33}$$

where m^o is the smallest budget to achieve a desired level of variance of the treatment effect (or statistical power) under the optimal sample allocation and m is the budget to achieve the same level of design precision under an alternative and sub-optimal design.

Using information from Equation 17, both perspectives share a more general relative precision and efficiency (RPE) expression for a sub-optimal design relative to the optimal design for two-level cluster-randomized trails as

$$RPE = \frac{[\rho(1-R_2^2)n^o+(1-\rho)(1-R_1^2)][(1-p^o)(c_1n^o+c_2)+p^o(c_1^Tn^o+c_2^T)]p(1-p)n}{[\rho(1-R_2^2)n+(1-\rho)(1-R_1^2)][(1-p)(c_1n+c_2)+p(c_1^Tn+c_2^T)]p^o(1-p^o)n^o}, \tag{34}$$

where p^o and n^o represent the optimal design parameter values or the roots of p and n in Equations 18 and 19, and n and p represent the alternative parameter values identified under a different framework or a study actually carried out. The RPE in Equation 34 can be used to measure (1) the relative variance increased in a study than that in the optimal design under a fixed budget, and (2) the relative budget requested by the optimal design to that by a sub-optimal study to achieve a same level of error variance. Comparing with the optimal design benchmark, the percentage of increased variance/budget by a study is $(1 - RE)/RE \times 100\%$. RPE values of at least .90 are generally considered good, and an RPE between .80 and .90 is considered acceptable (Korendijk, Moerbeek, & Maas, 2010; Hedges & Borenstein, 2014).

Unlike power, effect size, or sample sizes, the variance of the treatment effect is not the simplest design parameter researchers usually face. To systematically improve statistical precision for designs, it is important to transfer such a measure to the ultimate parameter researchers can directly consider. We can take a third perspective to further transfer the measure under the first perspective. Let the statistical power and the budget be fixed between the optimal

and sub-optimal designs and further compare the relative values of MDES between two designs. Under this perspective, the statistical power and thus the noncentrality parameter λ are equal between optimal and sub-optimal designs.

Thus, we have $\lambda = \lambda^o$ or $\delta/\sqrt{\sigma_\delta^2} = \delta^o/\sqrt{\sigma_{\delta^o}^2}$ with the additional subscripts to denote parameters in the optimal design. Rearranging this equation, we have

$$\delta^o = \delta\sqrt{RPE}, \tag{35}$$

where δ^o and δ are the respective MDES in the optimal and sub-optimal designs under a same budget to maintain the same level of statistical power. Equation 35 quantifies the relative statistical precision, measured by MDES, between an optimal and sub-optimal design, thus it can be used to improve statistical precision by carefully choosing the best available optimal sample allocation and MDES. A design with an RPE of .90 can detect about a 5% smaller effect if it uses the optimal design ($\sqrt{0.90} \approx 0.95$). A design with an RPE of .80 can detect about a 11% smaller effect if the optimal design is used ($\sqrt{0.80} \approx 0.89$). Additionally, given specific design parameters, researchers can directly compute the relative statistical power of a sub-optimal and optimal design by using statistical power formulas (Equation 14 or 15).

Similarly, the RPE for a sub-optimal design relative to the optimal design for three-level cluster-randomized trails is

$$RPE = \frac{n^o J^o \rho_3(1-R_3^2)+n^o\rho_2(1-R_2^2)+(1-\rho_2-\rho_3)(1-R_1^2)}{nJ\rho_3(1-R_3^2)+n\rho_2(1-R_2^2)+(1-\rho_2-\rho_3)(1-R_1^2)} \times \frac{[(1-p^o)(c_1 n^o J^o+c_2 J^o+c_3)+p^o(c_1^T n^o J^o+c_2^T J^o+c_3^T)]p(1-p)nJ}{[(1-p)(c_1 nJ+c_2 J+c_3)+p(c_1^T nJ+c_2^T J+c_3^T)]p^o(1-p^o)n^o J^o}, \tag{36}$$

where p^o, n^o, and J^o represent the solved values for optimal design parameters expressed in Equations 29 to 31, respectively. p, n and J represent the actual values a three-level design carried out or identified under a different framework.

A Comparison with Previous Frameworks

In the derivation section, we have shown that previous optimal design frameworks for two-level cluster-randomized trials (Raudenbush, 1997; Liu, 2003) are special cases of our proposed framework. The optimal design parameters for two-level cluster-randomized trials are n and p in our proposed framework. They are n and the constraint of $p = .5$ in the Raudenbush (1997) framework, p and a predetermined value of n in the Liu (2003) framework. Both previous frameworks can be viewed as constrained optimal designs in our proposed framework. Thus, we can directly assess the RPE values of designs identified by previous frameworks comparing with the benchmark designs identified under our proposed framework. Since the conclusion for three-level cluster-randomized trials is the same for two-level cluster-randomized trials, next we only present the results for two-level cluster-randomized trials.

For the cost structures, we considered both equal and unequal costs between treatment conditions and set the cost of sampling one additional individual in the control condition as one (i.e., $c_1 = 1$). For the equal costs between treatment conditions, we considered cluster/individual cost ratios as 3, 10, and 30 to reflect potential differences in the costs of sampling a cluster and an individual within a cluster (e.g., Raudenbush, 1997) and presented them in the first three rows of the left panel in Table 1. We considered two scenarios for the unequal costs between treatment conditions. The first scenario fixes the cluster/individual cost ratio in the control condition as 10 and considers a cluster-level treatment to control cost ratio of 3 (e.g., efficacy studies of interventions; Greenleaf et al., 2011; Jacob et al., 2015), 10 (e.g., teacher pay for performance;

Springer et al., 2011), and 30 (e.g., Tennessee class size experiment; Mosteller, 1995). These cost structures are presented in the rows 4 to 6 of the left panel in Table 1. The second scenario considers the cluster/individual cost ratio in the control condition as 3 or 10 and varies the treatment/control cost ratio as those of first scenario (3, 10, 30) but at both the cluster and individual levels. These cost structures are presented in the rows 7 to 12 of the left panel in Table 1.

For the intraclass correlation coefficient, we considered values of 0.15 and 0.25 (e.g., Hedges & Hedberg, 2007). For the R squared values or the proportions of outcome variance explained by covariates, we considered three types of design. The first type of design has no covariate adjustment (i.e., $R_1^2 = R_2^2 = 0$). The second type of design has a half of cluster-level outcome variance explained by a cluster-level covariate (i.e., $R_1^2 = 0$, $R_2^2 = 0.5$, and $q = 1$). The third type of design has covariates explained a half of outcome variances at both the cluster and individual levels (i.e., $R_1^2 = R_2^2 = 0.5$, and $q = 1$).

For simplicity, we used $n = 20$ as the predetermined individual-level sample size in the framework by Liu (2003). In the computation we rounded the values of n to integers and the values of p and RPE to two decimal places. The results for designs with a cluster-level covariate are presented in Table 1. For other two types of design (i.e., designs without a covariate and designs with covariates at both levels), the conclusions are similar with those in Table 1 and are not repeatedly presented.

Across all values of cost structures and design parameters, there are 11 out of 24 designs identified under the Raudenbush (1997) framework have RPE values below the good level of .90 (Table 1). From a relative perspective, designs identified under the Raudenbush (1997) framework achieve lower statistical power under the same budgets requested by the proposed framework. The statistical power drops to .70 when the treatment/control sampling cost ratio is 10, and .63 for a cost

ratio of 30 (Table 1). Half (12 of 24) of the designs identified under the Liu (2003) framework have RPE values below the good level of .90 (Table 1).

For designs identified under previous frameworks, the RPE values and the relative statistical power are directly influenced by how far the constrained values depart from the optimal values in our proposed framework. For example, when the costs of sampling are equal between treatment conditions (e.g., first three cost structures in Table 1), the constrained p under the Raudenbush (1997) framework is equal to the optimal $p = .5$ in our framework, thus the Raudenbush framework can identify identical designs with RPE values of one. When the constrained $p = .5$ departs far away from the optimal values, designs identified under the Raudenbush framework have much lower RPE values and statistical power (e.g., the last cost structure in Table 1).

We can see similar patterns for the Liu (2003) framework in Table 1, when the predetermined $n = 20$ is close to the optimal values under the proposed framework, the RPE values for designs under the Liu (2003) framework are close to one (e.g., the third to sixth cost structures in Table 1). When the predetermined individual-level sample sizes are far from the optimal values, we have much lower RPE values and statistical power (e.g., the first cost structures in Table 1). Collectively, the results comparing with previous frameworks show that the proposed framework can be used to significantly improve design precision and efficiency, especially when the cost of sampling a treatment unit is multiple times that for a control unit.

Table 1.

Comparison of Proposed Framework with Previous Frameworks for Two-Level Cluster-Randomized Trials.

Cost Structures	ρ	Proposed			Raudenbush				Liu			
		p	n	J	n	J	RPE	Pr	p	J	RPE	Pr
$c_1^T = 1, c_2 = c_2^T = 3$	.15	.50	6	172	6	172	1.0	.80	.50	94	**.72**	.65
	.25	.50	4	247	4	247	1.0	.80	.50	130	**.59**	.56
	.15	.50	11	121	11	121	1.0	.80	.50	94	.91	.76

$c_1^T = 1, c_2 = c_2^T = 10$	.25	.50	8	174	8	174	1.0	.80	.50	130	**.81**	.71
$c_1^T = 1, c_2 = c_2^T = 30$	.15	.50	18	98	18	98	1.0	.80	.50	94	1.0	.80
	.25	.50	13	145	13	145	1.0	.80	.50	130	.97	.79
$c_1^T = 1, c_2 = 10, c_2^T = 30$	.15	.43	14	111	15	105	.98	.79	.44	96	.98	.79
	.25	.42	10	163	11	154	.97	.79	.44	131	.91	.76
$c_1^T = 1, c_2 = 10, c_2^T = 100$	.15	.34	21	103	25	88	.91	.76	.33	106	1.0	.80
	.25	.32	15	160	18	133	**.89**	.75	.33	146	.99	.79
$c_1^T = 1, c_2 = 10, c_2^T = 300$	.15	.26	31	106	42	77	**.83**	.72	.23	132	.97	.79
	.25	.24	22	173	30	120	**.80**	.70	.23	182	1.0	.80
$c_1^T = 3, c_2 = 3, c_2^T = 9$	.15	.37	6	184	6	172	.93	.77	.37	101	**.72**	.66
	.25	.37	4	265	4	247	.93	.77	.37	139	**.59**	.57
$c_1^T = 10, c_2 = 3, c_2^T = 30$	.15	.24	6	235	6	172	**.79**	.70	.24	128	**.72**	.66
	.25	.24	4	338	4	247	**.79**	.70	.24	177	**.59**	.57
$c_1^T = 30, c_2 = 3, c_2^T = 90$	.15	.15	6	335	6	172	**.68**	.63	.15	183	**.72**	.66
	.25	.15	4	483	4	247	**.68**	.63	.15	252	**.59**	.57
$c_1^T = 3, c_2 = 10, c_2^T = 30$	.15	.37	11	130	11	121	.93	.77	.37	101	.91	.76
	.25	.37	8	186	8	174	.93	.77	.37	139	**.81**	.71
$c_1^T = 10, c_2 = 10, c_2^T =$	.15	.24	11	166	11	121	**.79**	.70	.24	128	.91	.76
100	.25	.24	8	237	8	174	**.79**	.70	.24	177	**.81**	.71
$c_1^T = 30, c_2 = 10, c_2^T =$	.15	.15	11	236	11	121	**.68**	.63	.15	183	.91	.76
300	.25	.15	8	339	8	174	**.68**	.63	.15	252	**.81**	.71

Note. Pr is the statistical power of designs identified by previous frameworks for the same budget that produces a power of .80 under the proposed framework. The Raudenbush (1997) framework assumes $p = .5$, the results for the Liu (2003) framework are based on a predetermined individual-level sample size of 20.

To illustrate the difference in the required total sample size under different optimal design framework, further suppose researchers plan to implement the cluster-randomized trials to detect a standardized effect of 0.2 (Spybrook, Shi, & Kelcey, 2016). We reported the total number of clusters (J) needed to have a power level of 0.8 for the effect size of 0.2 in Table 1. The results show that we can sample more clusters under the proposed framework than those under the Raudenbush (1997) framework but with less budget request to achieve a power of 0.8 (e.g., see J and RPE values for the forth to last cost structures in Table 1).

Comparing with the Raudenbush (1997) framework, the proposed framework gains efficiency mainly through sampling less clusters in the experimental group but much more clusters in the control group. This mechanism results in the opposite directions in the change of

the optimal proportion p and the number of total clusters J. For example, comparing results in the first and last three cost structures in Table 1, we can clearly see that the more expensive sampling in treatment is, the smaller the optimal p and the larger the number of total clusters J. This mechanism of opposite directions in the change of p and J ensures that we still have enough clusters (e.g., classrooms) in the treatment condition.

For example, in the last cost structure where sampling a treatment cluster (e.g., regular class assisted by a teacher aide; Mosteller, 1995) costs 30 times that of sampling a cluster in control (e.g., a regular class), with $\rho = 0.25$ we need to sample 87 clusters in each treatment condition under the Raudenbush (1997) framework. However, under the proposed framework we have an optimal p of .15 and J of 339. The number of total clusters to be sampled is about twice the number (174) in the balanced design. Under the proposed framework, there will be 51 cluster in the treatment condition, 36 clusters less than that in the balanced design, and 288 clusters in the control condition, 201 clusters more than that in the balanced design. Yet, the balanced design will require a 47% larger budget than that required under the proposed framework to achieve comparable power.

Given the same requested budget by previous framework to detect an effect of 0.20 with a power of 0.8, we can detect a smaller effect under the proposed framework, and the MDES under proposed framework can be calculated based on these RPE values. Taking the same example mentioned above with an RPE of .68, we can detect an effect of 0.16 under the proposed framework with the same budget, which is 20% smaller than 0.20. The optimal sample allocation can significantly improve design precision than that under the previous framework, and a smaller MDES can account for the overestimate of an effect size due to sampling error and other factors. In conclusion, we have shown that the proposed framework can be used to recover

more gains in statistical precision and efficiency that have gone unconsidered in previous frameworks.

Design Sensitivity

To further probe the loss of efficiency resulting from constrained designs and the sensitivity of optimal designs to misspecifications of parameter values at the planning stage, we examined the extent to which proposed designs are robust to incorrect initial values of the cost structure and the design parameter values. Similarly, we only present the results for two-level cluster-randomized trials, as the conclusion is the same for three-level experiments.

In our analyses, we first calculated the true optimal design parameter values (n^o and p^o) based on the true values and then computed the optimal design parameter values (n and p) under misspecified initial values. Using Equation 34 we then computed the RPE values designs achieved. For the comparison, we used the same cost structures and design parameter values that have been used in the previous section. We rounded the values of n to integers and the values of p and RPE to two decimal places in the computation. We presented the result for designs with a covariate at the cluster level ($R_1^2 = 0$ and $R_2^2 = .5$), results for other types of designs have similar conclusions and will be provided upon request.

Robustness to the Misspecification of Intraclass Correlation Coefficient

In terms of the range of misspecification on intraclass correlation coefficient, we considered multiplicative values of the true parameter—0.25, 0.5, 2, and 3 times the true values—mapping the range of 0.25 to 2.75 times the true values within which constrained optimal designs ($p = 0.5$) showed robustness in previous literature (Korendijk, Moerbeek, & Maas, 2010). Across cost structures, R squared values, and intraclass correlation coefficients, when the misspecification of intraclass correlation coefficients is 0.5 or 2 times the true values,

designs averaged an RPE of .96 or .94, respectively. Practically, the results suggest that planning studies under misspecifications of this type and magnitude will often require a budget that is only about 5% larger than the optimal design benchmark, or the optimal design can detect a less than 3% smaller effect.

When the misspecification of the intraclass correlation coefficient is even larger—for example 0.25 or 3 times the true values—the average RPE values are about .88 and .81, respectively. Our initial probe suggests that the optimal sample allocation identified under the proposed framework is fairly robust to the misspecification of the intraclass correlation coefficients.

Table 2.

Robustness of Optimal Sample Allocation to the Misspecification of Intraclass Correlation Coefficients.

Cost Structures	ρ	Misspecification of ρ			
		0.25	0.5	2	3
$c_1^T = 1, c_2 = c_2^T = 3$	.15	.89	.96	.97	.91
	.25	.88	.97	.88	.63
$c_1^T = 1, c_2 = c_2^T = 10$	.15	.87	.96	.96	.87
	.25	.86	.95	.90	.81
$c_1^T = 1, c_2 = c_2^T = 30$	.15	.88	.97	.96	.89
	.25	.87	.97	.95	.74
$c_1^T = 1, c_2 = 10, c_2^T = 30$	.15	.87	.96	.95	.88
	.25	.86	.96	.93	.71
$c_1^T = 1, c_2 = 10, c_2^T = 100$	.15	.87	.97	.95	.85
	.25	.87	.96	.95	.79
$c_1^T = 1, c_2 = 10, c_2^T = 300$	.15	.89	.97	.96	.87
	.25	.89	.97	.95	.81
$c_1^T = 3, c_2 = 3, c_2^T = 9$	.15	.89	.96	.97	.91
	.25	.88	.97	.88	.63
$c_1^T = 10, c_2 = 3, c_2^T = 30$	.15	.89	.96	.97	.91
	.25	.88	.97	.88	.63
$c_1^T = 30, c_2 = 3, c_2^T = 90$	.15	.89	.96	.97	.91
	.25	.88	.97	.88	.63
$c_1^T = 3, c_2 = 10, c_2^T = 30$	.15	.87	.96	.96	.87
	.25	.86	.95	.90	.81
	.15	.87	.96	.96	.87

$c_1^T = 10, c_2 = 10, c_2^T = 100$	.25	.86	.95	.90	.81
$c_1^T = 30, c_2 = 10, c_2^T = 300$	.15	.87	.96	.96	.87
	.25	.86	.95	.90	.81
Average		.88	.96	.94	.81

Robustness to the Misspecification of Cost Structures

As for the misspecification of initial cost structure, we investigated the robustness of optimal design to the misspecification on initial cluster/individual cost ratio (CICR) and treatment/control cost ratio (TCCR). The range of the misspecification was set as 0.25, 0.5, 2, and 4 times the true values. The results are presented in Table 3. When the misspecification is 0.5 or 2 times the true CICR, designs have an average RPE value of .97. Even when the misspecification is 0.25 or 4 times the true CICR, designs have average RPE values of .89 or .90, respectively. As for the misspecification of initial TCCR values, the results are similar. Even when the misspecification is 0.25 or 4 times the true TCCRs, designs have an average RPE value of .90. The results suggest that designs optimized under moderate misspecifications of cost ratios largely retain their RPE values.

Table 3.

Robustness of Optimal Sample Allocation to Misspecification of Cost Structures Measured by Relative Precision and Efficiency.

Cost Structures	ρ	Misspecification of CICR				Misspecification of TCCR			
		0.25	0.5	2	4	0.25	0.5	2	4
$c_1^T = 1, c_2 = c_2^T = 3$	.15	.91	.97	.98	.89	.88	.97	.97	.88
	.25	.88	.97	.97	.91	.88	.97	.97	.88
$c_1^T = 1, c_2 = c_2^T = 10$	.15	.87	.98	.97	.89	.88	.97	.97	.88
	.25	.90	.95	.97	.90	.88	.97	.97	.88
$c_1^T = 1, c_2 = c_2^T = 30$	.15	.89	.97	.97	.89	.88	.97	.97	.88
	.25	.91	.97	.97	.90	.88	.97	.97	.88
$c_1^T = 1, c_2 = 10, c_2^T = 30$	.15	.87	.97	.97	.88	.89	.97	.97	.88
	.25	.88	.96	.97	.89	.90	.97	.97	.89
$c_1^T = 1, c_2 = 10, c_2^T = 100$	.15	.89	.97	.97	.89	.90	.97	.97	.90
	.25	.90	.97	.97	.91	.90	.98	.97	.90
$c_1^T = 1, c_2 = 10, c_2^T = 300$	.15	.90	.97	.98	.90	.91	.98	.98	.90
	.25	.92	.98	.98	.91	.91	.98	.98	.89

$c_1^T = 3, c_2 = 3, c_2^T = 9$	.15	.91	.97	.98	.89	.89	.97	.97	.89
	.25	.88	.97	.97	.91	.89	.97	.97	.89
$c_1^T = 10, c_2 = 3, c_2^T = 30$	.15	.91	.97	.98	.89	.91	.98	.98	.92
	.25	.88	.97	.97	.91	.91	.98	.98	.92
$c_1^T = 30, c_2 = 3, c_2^T = 90$	.15	.91	.97	.98	.89	.94	.98	.98	.93
	.25	.88	.97	.97	.91	.94	.98	.98	.93
$c_1^T = 3, c_2 = 10, c_2^T = 30$	.15	.87	.98	.97	.89	.89	.97	.97	.89
	.25	.90	.95	.97	.90	.89	.97	.97	.89
$c_1^T = 10, c_2 = 10, c_2^T = 100$	.15	.87	.98	.97	.89	.91	.98	.98	.92
	.25	.90	.95	.97	.90	.91	.98	.98	.92
$c_1^T = 30, c_2 = 10, c_2^T = 300$	.15	.87	.98	.97	.89	.94	.98	.98	.93
	.25	.90	.95	.97	.90	.94	.98	.98	.93
Average		.89	.97	.97	.90	.90	.97	.97	.90

Note. CICR is the cluster/individual cost ratio. TCCR is the treatment/control cost ratio.

Discussion

Prior literature has developed a host of strategies and tools to improve the efficiency with which designs can estimate effects (e.g., Bloom, Richburg-Hayes, & Black, 2007; Raudenbush, Martinez, & Spybrook, 2007; Kelcey, B., & Phelps, 2013; Kelcey, Shen, & Spybrook, 2016; Schochet, 2008; Borenstein, Hedges, & Rothstein, 2012; Dong & Maynard, 2013). Previous optimal design frameworks have been limited in their modeling the cost structures of sampling and optimizing the sampling ratios across levels and treatment conditions. In this paper, our proposed framework addresses this need by developing a flexible cost framework that more naturally maps onto practical design settings. The results of the extended framework identify potentially important gains in statistical precision and efficiency that have previously gone unconsidered.

Even when some of the parameters are constrained by practical considerations, our results suggest that within a broad range of applied settings the proposed framework can identify sampling strategies with more precision and efficiency than those detailed in previous literature. In this way, the introduction of a treatment-condition specific cost framework and the

optimization of sampling ratios across levels and treatment conditions can be useful for adjudicating among several potential designs with varying constraints. Additionally, the proposed framework performed better than previous frameworks even when the parameter values were misspecified.

To design cluster-randomized trials with adequate statistical power and efficiency under an optimal design framework, researchers additionally need the cost information about sampling. The information about the cost of sampling a unit can usually be estimated through pilot studies, budget planning, similar studies, or cost centers (e.g., CostOut at https://www.cbcse.org/costout). Even when cost estimation may not be strictly accurate, our initial probe of the proposed optimal design framework suggested that the results are fairly robust to the misspecification on initial values of intraclass correlation coefficient and cost structures. In this way, our results suggest that even when some parameters are constrained, and some are misspecified, there are still advantages to probing more flexible sampling plans.

In the presence of unequal sampling costs between treatment conditions, we have illustrated that unbalanced designs can be more efficient than balanced ones. Put another way, unbalanced designs can return more statistical power than balanced designs under unequal sampling costs between treatment conditions. It is generally assumed that the treatment or intervention itself does not change the standardized variance of an outcome. For designs with unequal number of clusters between treatment conditions, the assumption of homogeneity of variance between treatment conditions (controlling for the treatment effect) can still be tested the same way with balanced designs as the variance formulas adjust for the number of clusters.

We illustrated the opposite directions in the change of the optimal p and the number of total clusters needed for a certain level of statistical power. This mechanism ensures that

unbalanced designs still result in enough clusters to be sampled in a treatment condition. However, when the number of total clusters is small and the proportion of clusters to be assigned to the treatment condition is also small, there may be an issue whether the treatment arm can correctly reflect the population variance, and thus there may be a homoskedasticity issue between treatment conditions. Further studies address small number of clusters in unbalanced design is needed.

Despite the utility of our framework and the potential gains in statistical precision and efficiency it offers, we caution readers that the resulting optimal sampling plans are intended to serve as a starting point for planning a cluster-randomized trial rather than a rigid tool. For example, an analysis of optimal design may suggest a small value of optimal proportion (p) if sampling costs vastly different between treatment conditions. In power analysis, a small value of p may suggest a large number of total clusters that exceeds the clusters researchers could practically reach. In this case, researchers should constraint the optimal proportion to a larger number than that the analysis gives so that a feasible design can be achieved. In practice, the optimal sampling plan operates as a type of initial strategy or benchmark that is subsequently moderated by practical design considerations and constraints to reach a final sampling plan.

To facilitate end-user calculations, we have developed a freely available *R* package XXX (citation masked) that implements the proposed framework. The package also can perform power analysis accommodating costs by default (e.g., required budget/sample size calculation, power calculation under a given budget, minimum detectible effect size calculation under a given budget) and conventional power analysis (e.g., sample size, power, and MDES calculation).

References

Bloom, H. S. (1995). Minimum detectable effects: A simple way to report the statistical power of experimental designs. *Evaluation Review, 19*(5), 547-556.

Bloom, H. S. (2005). Randomizing groups to evaluate place-based programs. In H. S. Bloom (Ed.), Learning more from social experiments: Evolving analytic approaches. New York, NY: Russell Sage Foundation.

Bloom, H. S., Richburg-Hayes, L., & Black, A. R. (2007). Using covariates to improve precision for studies that randomize schools to evaluate educational interventions. *Educational Evaluation and Policy Analysis, 29*(1), 30-59.

Borenstein, M., Hedges, L. V., & Rothstein, H. (2012). CRT-power [computer software]. Teaneck, NJ: Biostat. Available from http://www.crt-power.com.

Cochran, W. (1963). *Sampling techniques*. (2nd ed.). New York: Wiley.

Connelly, L. B. (2003). Balancing the number and size of sites: an economic approach to the optimal design of cluster samples. *Controlled Clinical Trials, 24*(5), 544-559.

Dong, N., & Maynard, R. A. (2013). PowerUp!: A tool for calculating minimum detectable effect sizes and sample size requirements for experimental and quasi-experimental designs. *Journal of Research on Educational Effectiveness, 6*(1), 24-67.

Donner, A. & Klar, N. (2000). *Design and analysis of cluster randomization trials in health research.* London: Arnold

Glynn, T. J., Shopland, D. R., Manley, M., Lynn, W. R., Freedman, L. S., Green, S. B., ... & Chapelsky, D. A. (1995). Community intervention trial for smoking cessation (COMMIT): I. Cohort results from a four-year community intervention. *American Journal of Public Health, 85*(2), 183-192.

Greenleaf, C. L., Litman, C., Hanson, T. L., Rosen, R., Boscardin, C. K., Herman, J., ... & Jones, B. (2011). Integrating literacy and science in biology: Teaching and learning impacts of reading apprenticeship professional development. *American Educational Research Journal, 48*(3), 647-717.

Hedges, L. V., & Borenstein, M. (2014). Conditional optimal design in three-and four-level experiments. *Journal of Educational and Behavioral Statistics, 39*(4), 257-281.

Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis, 29*(1), 60-87.

Hiscock, H., Bayer, J. K., Price, A., Ukoumunne, O. C., Rogers, S., & Wake, M. (2008). Universal parenting programme to prevent early childhood behavioural problems: cluster randomised trial. *British Medical Journal, 336*(7639), 318-321.

Hoover, D. R. (2002). Power for T-test comparisons of unbalanced cluster exposure studies. *Journal of Urban Health, 79*(2), 278-294.

Jacob, R., Goddard, R., Kim, M., Miller, R., & Goddard, Y. (2015). Exploring the causal impact of the McREL Balanced Leadership Program on leadership, principal efficacy, instructional climate, educator turnover, and student achievement. *Educational Evaluation and Policy Analysis, 37*(3), 314-332.

Jayanthi, M., Gersten, R., Taylor, M. J., Smolkowski, K., & Dimino, J. (2017). Impact of the developing mathematical ideas professional development program on grade 4 students' and teachers' understanding of fractions (REL 2017–256). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education

Evaluation and Regional Assistance, Regional Educational Laboratory Southeast. Retrieved from http://ies.ed.gov/ncee/edlabs

Kelcey, B., & Phelps, G. (2013). Strategies for improving power in school-randomized studies of professional development. *Evaluation Review, 37*(6), 520-554.

Konstantopoulos, S. (2009). Incorporating cost in power analysis for three-level cluster-randomized designs. *Evaluation Review, 33*(4), 335-357.

Konstantopoulos, S. (2011). Optimal sampling of units in three-level cluster randomized designs: An ANCOVA framework. *Educational and Psychological Measurement, 71*(5), 798-813.

Korendijk, E. J., Moerbeek, M., & Maas, C. J. (2010). The robustness of designs for trials with nested data against incorrect initial intracluster correlation coefficient estimates. *Journal of Educational and Behavioral Statistics, 35*(5), 566-585.

Liu, X. (2003). Statistical power and optimum sample allocation ratio for treatment and control having unequal costs per unit of randomization. *Journal of Educational and Behavioral Statistics, 28*(3), 231-248.

Moerbeek, M., van Breukelen, G. J., & Berger, M. P. (2000). Design issues for experiments in multilevel populations. *Journal of Educational and Behavioral Statistics, 25*(3), 271-284.

Mosteller, F. (1995). The Tennessee study of class size in the early school grades. *The Future of Children, 5*(2): 113-127.

Nam, J. M. (1973). Optimum sample sizes for the comparison of the control and treatment. *Biometrics,* 29, 101-108.

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

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

R Core Team. (2019). *R: A Language and Environment for Statistical Computing.* [Software]. Available at https://www.R-project.org/.

Rutterford, C., Copas, A., & Eldridge, S. (2015). Methods for sample size determination in cluster randomized trials. *International Journal of Epidemiology, 44*(3), 1051-1067.

Schochet, P. Z. (2008). Statistical power for randomized assignment evaluation of education programs. *Journal of Educational and Behavioral Statistics, 33*(1), 62–87.

Springer, M. G., Ballou, D., Hamilton, L., Le, V. N., Lockwood, J. R., McCaffrey, D. F., ... & Stecher, B. M. (2011). *Teacher pay for performance: Experimental evidence from the Project on Incentives in Teaching (POINT).* Society for Research on Educational Effectiveness. Available at https://files.eric.ed.gov/fulltext/ED518378.pdf.

Spybrook, J., Shi, R., & Kelcey, B. (2016). Progress in the past decade: an examination of the precision of cluster randomized trials funded by the US Institute of Education Sciences. *International Journal of Research & Method in Education, 39*(3), 255-267.

Turner, R. M., Toby Prevost, A., & Thompson, S. G. (2004). Allowing for imprecision of the intracluster correlation coefficient in the design of cluster randomized trials. *Statistics in Medicine, 23*(8), 1195-1214.

Acknowledgements

We thank the editor, Dr. Daniel McCaffrey, three anonymous reviewers, and Dr. XXX (masked for blind review) at Harvard University for their insightful comments and suggestions on earlier drafts of the manuscript.