More Powerful Cluster Randomized Control Trials

Brendon McConnell and Marcos Vera-Hernández*

August 29, 2025

Abstract

Balanced experimental designs, in which the number of treatment and control units are the same, do not maximize power subject to a cost constraint when treatment units are more expensive than control ones. Despite this, such balanced designs are the norm in economics. This paper describes methods to optimally choose the number of treatment and control clusters, and the number of units within treatment and control clusters, allowing for full flexibility. We use three archetypal examples from the development literature to illustrate the magnitude of the power gains, which lie between 7.9 and 19.0 percentage points.

Keywords: Power analysis, Sample size calculations, Randomized Control Trials, Cluster Randomized Control Trials
JEL Codes: C8, C9

*We gratefully acknowledge the ESRC-NCRM Node Programme Evaluation for Policy Analysis (ESRC grant reference ES/I02574X/1) and the ESRC Centre for the Microeconomic Analysis of Public Policy (ES/T014334/1) for funding this project. We thank Bet Caeyers for valuable comments. We also benefited from comments received at a workshop of The Centre for the Evaluation of Development Policies at The Institute for Fiscal Studies. We thank Nancy Daza Baez and Jiacheng Fan for excellent research assistance. All errors are our own responsibility. Author affiliations and contacts: McConnell (City St George's, University of London: brendon.mcconnell@gmail.com); Vera-Hernández (UCL, IFS, and CEPR: m.vera@ucl.ac.uk)

1 Introduction

One of the key challenges in economics is to estimate causal relationships between economic variables and policy instruments. Randomized Controlled Trials (RCT) have become one of the main tools that researchers use to accomplish this objective (Hausman and Wise, 1985; Burtless, 1995; Heckman and Smith, 1995; Duflo et al., 2007; Hamermesh, 2013; Olken, 2020). More simple RCTs are usually set up with the objective of estimating the impact of a certain policy or intervention, while more complex RCTs can be implemented to test between competing hypotheses that explain a phenomenon (also known as field experiments, see Duflo (2006); Levitt and List (2009); Bandiera et al. (2011); List (2011); List and Rasul (2011); Karlan and Appel (2016); Duflo (2020)).

The focus of this paper is on maximizing the statistical power – the probability that the null hypothesis of zero effect is correctly rejected – of a cluster RCT given a budget constraint. This is important because not only do underpowered RCTs have a smaller probability of detecting a true effect, but they also have a smaller probability that a statistically significant result reflects a true effect (Wacholder et al., 2004; Ioannidis, 2005; Button et al., 2013). Ioannidis et al. (2017) find that the median statistical power in Economics (in general, not specifically in RCT studies) is 18%. Moreover, low-powered RCTs are more likely to lead to estimates whose sign is the opposite to the true one, and estimates whose size is much larger than the true effect size (Gelman and Carlin, 2014).

It is well known that in the case of *individual* level randomization, a given power can be achieved at a smaller cost if more control and fewer treatment units are sampled than in the balanced design (Cochran, 1963; Nam, 1973; Duflo et al., 2007).[1] However, this problem is considerably more complex in cluster RCTs, not only because there are no closed form solutions, but because the pattern of the optimal solution might not be monotonic. Indeed, depending on the cost structure, it might be optimal to have more treatment clusters but fewer units within treatment clusters than control ones, or fewer treatment clusters but more units within treatment than control clusters.

In the Statistics literature, there is a long tradition of solving the dual problem: minimizing costs subject to achieving a given level of power. Instead, the primal problem of maximizing power subject to a cost constraint is more relevant in economics research, in which funding bodies specify a maximum funding amount per project. Although we focus on the primal problem, the

[1] Although, for a fixed total number of units, having an unbalanced number of treatment and control units decreases the power of the RCT, this can be compensated, at a lower cost, by increasing the number of control units (as they are cheaper than the treatment ones).

Appendix includes the methods to solve the dual one.

This paper makes two key contributions. First, we derive methods to calculate the optimal sample for a cluster RCT with an endline outcome measure, allowing for different number of treatment and control clusters, as well as different number of units within treatment and control clusters. Prior to our work, no existing method provided a solution for optimal sample allocation in cluster RCTs that allowed full flexibility in both the number of treatment and control clusters and the number of units within clusters in each arm.[2] Our method can be applied to solve the primal problem of maximizing power subject to a cost constraint, or to minimize costs subject to a given level of power.

We model the cost function of the RCT as having a fixed cost per cluster as well as a cost per sampled unit (unit cost). We consider two pure cases and a hybrid one: (i) the fixed cost per cluster is different between treatment and control but the unit cost is the same, (ii) the fixed cost per cluster is the same between treatment and control but the unit cost is the different, as well as, (iii) the hybrid case in which both the fixed and unit costs are different. It is important to highlight that even in a pure case, there are gains in a fully flexible solution that allows for a different number of clusters and units within clusters for treatment and control arms.

Our second contribution is to show that the gains in power are very significant in typical cluster RCTs from economics. The first example is a cluster RCT in which headteachers are given an unconditional grant to improve the school, and the experiment measures the effect of the grant on children's hemoglobin levels (a biomarker for nutritional status, and in particular, anaemia) as in Luo et al. (2019). This is an example in which the fixed cost per school is much larger in treatment than control clusters (because of the grant) but the unit cost of sampling a child (hemoglobin test and questionnaire time) is the same in treatment and control schools.

The second example is the case of an unconditional cash transfer program, as is analysed by Haushofer and Shapiro (2016), in which treated households receive a large unconditional cash transfer, and in which a cluster design is used to take into account of spillovers and general equilibrium effects. Unlike the previous example, the fixed cost per cluster is the same independently of whether it is a treatment or control one, as the only fixed cost per cluster is the transportation one. However, the cost of a treatment unit is much larger than a control one, as the cost of the treatment unit includes the unconditional cash transfer and the interviewing time, but only the latter for control units.

[2]Cochran (1963), Nam (1973), and Duflo et al. (2007) all provide an optimal solution in the individual randomization setting. Shen and Kelcey (2020) provides a solution for the the cluster randomization case but restricts the number of units per cluster to be the same in the treatment and control arms.

Our example for the hybrid case refers to the so called "graduation model" in which households are given large productive assets (i.e., a large animal), time-limited cash transfers, access to health services, as well as training and life skills coaching as in Banerjee et al. (2015) and Bandiera et al. (2017). Because these programs provide training, coaching and access to health services, they need certain infrastructure in the treatment clusters to deliver these services and hence the fixed cost per treatment cluster is higher. In addition, the cost of each treated unit is higher because of the productive asset and cash transfer. Hence, this example synthesizes the previous two cases, by having both larger fixed cluster costs as well as larger unit treatment costs.

Our results indicate that, compared to a balanced design, optimally allocating the number of clusters and the number of treatment units can increase power between 7.9 and 19.0 percentage points. To obtain these results we use realistic cost estimates based predominantly on the previous studies and reasonable assumptions on parameters which are unknown to us. We then compare the cost of the balanced design —in which the number of clusters and the number of units per cluster are identical across treatment and control arms— with that of the optimal allocation derived from our method. It should be noted that we do not replicate all the features of the previous studies, and hence our results should not be understood as what the previous studies could have gained. Instead, one should view our results as benchmark power gains that can be obtained in a typical cluster RCT using our proposed method.

We further consider the benefits of our approach, by attaching a monetary value to the power improvements. To do so, we ask: how much larger a budget would be required to achieve *the same power* attained using our approach if instead one implemented a balanced design? In answering this question, we document sizeable values associated with the power improvement based on our approach. Expressed in terms of the original budget, these are respectively 22%, 20% and 53% for the three case studies. Put differently, the value of our approach is akin to using the standard balanced design but being granted a budget of between 20% to 53% larger. This valuation exercise underscores the advantage of our approach. By moving away from a balanced design in a manner that accounts for differential costs, one can make sizeable power gains for a given budget.

A general feature of the results is that, in all three cases, both the number of clusters and the number of units within clusters are different between the treatment and control arms, in a compensating manner. For instance, when the fixed cost per cluster is larger in treatment than control clusters but the unit costs are the same, not only it is optimal to have fewer treatment

than control clusters, but also to sample more units per treatment than control clusters (in the margin, it is more efficient to increase the units per treatment cluster than paying the cost of an additional treatment cluster). In the hybrid case, depending on the differences in fixed and unit costs, the optimal solution could even involve not only more clusters but also more units per cluster in the control than in the treatment arm.

This paper contributes to a growing literature on methods to improve the design of RCTs. Hahn et al. (2011) consider using the propensity score to reduce the variance of the treatment effect in a setting in which an experiment is run in multiple waves or replicate previous experiments. McKenzie (2012) studies the problem of how many waves of post-treatment data to collect to maximize power given a budget constraint, noting that the standard choice of one baseline and one follow-up wave is unlikely to be optimal in many cases. Carneiro et al. (2019) focus on the choice of what covariates to collect to maximize power subject to a cost constraint. Chassang et al. (2012) show how to modify RCTs to improve external validity in a context in which the outcomes are significantly affected by unobserved effort decisions taken by experimental subjects, and Banerjee et al. (2020) study experimental design issues by an ambiguity-averse decision-maker who is concerned with both subjective expected performance and robust performance guarantees. Baird et al. (2018) studies the optimal design of experiments in which an individual's outcome depends on the outcomes of others in her group.Burlig et al. (2020) advise against using sample size formulae for the ANCOVA estimator (in which the post-treatment outcome variable is regressed over its baseline value and the treatment indicator), and hence we focus our paper on the case in which only the post-treatment values of the dependent variable are used in the estimation of the treatment effect. McKenzie (2025) offers practical guidance on enhancing statistical power in RCTs by outlining decisions and strategies that researchers can adopt at the design, implementation, and analysis stages

With respect to the literature that considers unequal costs of treatment and control units, the comprehensive reviews by Duflo et al. (2007), List et al. (2011) and Glennerster and Takavarasha (2013) all consider the case of unequal costs, but for *individual* RCTs instead of cluster ones. In the statistics literature, Liu (2003) is a pioneer in considering different costs in a cluster RCT, but the scenarios considered are relatively constrained, allowing only either fixed or unit heterogeneous costs, and requiring that either the number of clusters or the number of units per cluster be equal across the treatment and control arms. The closest to our work is Shen and Kelcey (2020) which has recently relaxed some of Liu (2003)'s constraints. In addition to allowing full flexibility in the number of treatment and control clusters, as well as in the number

of units per cluster in each arm, our approach offers two further advantages.[3] First, we explicitly maximize statistical power, while their method minimizes the variance of the treatment effect estimator. The latter ignores the degrees of freedom adjustment that depends on the total number of clusters —a factor that can affect inference when the number of clusters is relatively small (Leyrat et al., 2017). Second, our solution yields integer values for the number of treatment and control clusters as well as the number of units per cluster, thereby avoiding the need for post hoc rounding, which can affect power and is non-trivial under a cost constraint.

The rest of the paper is organized as follows: the next section describes the data generating process and defines the estimator. Section 3 describes our proposed method to determine the optimal sample size to maximize power given a cost constraint. Section 4 presents three examples from the literature to which we apply our method, and show the power gains that our proposed method achieves with respect to a balanced design. Section 5 describes some of the challenges that researchers can face when putting our solution into practice, and suggests solutions when possible, and Section 6 concludes. In the Appendix, we show the results of several useful simulations, show how we estimate the costs of our three empirical examples, and finally describe the dual approach of minimizing costs subject to achieving a given level of power.

2 Data Generating Process and Estimators

In this paper, we derive sample size calculations for a cluster RCT in which $j = 1, ..., K$ clusters have been randomized into treatment (denoted by $T_j = 1$) or control ($T_j = 0$). For each cluster j, we observe the outcome variable Y_{ij} for unit i (which may represent an individual, firm, etc.) at the time point when the treatment effect is to be estimated (i.e., at endline). The data generating process is:

$$Y_{ij} = \alpha + \delta T_j + v_j + \epsilon_{ij}, \tag{1}$$

where α represents the population mean of the outcome variable in the control group, δ is the treatment effect, v_j is a cluster-level error term distributed $N(0, \sigma_v^2)$ and ϵ_{ij} is an independent and $N(0, \sigma_\epsilon^2)$ distributed individual level error term. The intra-cluster correlation (ICC), which

[3]Shen and Kelcey (2020) already derived the first order conditions for the optimization problem and noted that it did not have closed form solutions (see their Appendix). Although their Odr package constraints the number of units per cluster to be the same in treatment and control, it has other advantageous features such as three and four level cluster experiments, multisite experiments, and includes power calculations for mediators and moderator effects (Shen and Kelcey, 2025).

is a key parameter in determining the required sample size in cluster RCTs is given by:

$$\rho = \frac{\sigma_v^2}{\sigma_v^2 + \sigma_\epsilon^2}$$

Sample size calculations are particular of the estimator that will be used to estimate the treatment effect, δ. In this case, the most standard is the Ordinary Least Squares (OLS) estimator of δ, $\hat{\delta}_{OLS}$, in the regression:

$$Y_{ij} = \alpha + \delta T_c + u_{ij}, \tag{2}$$

where u_{ij} is a zero mean error term, with $\text{var}(u_{ij})=\sigma_v^2+\sigma_\epsilon^2 = \sigma^2$, $\text{cov}(u_{ij}, u_{hj})=\sigma_v^2$ for $i \neq h$, and $\text{cov}(u_{ij}, u_{hl})=0$ if $j \neq l$. The above discussion makes explicit that the outcome variable is measured at the individual level.[4]

3 Optimal Sample Size Determination - Maximizing Power

3.1 General Problem

We now focus on optimal sample size determination for the case where the researcher wants to maximize power subject to a fixed budget.[5] The power, κ, of the two-tailed test at α significance for the null hypothesis that $H_0 : \delta = 0$ when using estimator $\hat{\delta}_{OLS}$ is given by Teerenstra et al. (2012) as:

$$\kappa = T_{k_0+k_1-2}\left(\frac{\delta}{\sqrt{var(\hat{\delta})}} - t_{\frac{\alpha}{2},k_0+k_1-2}\right) \tag{3}$$

where T is the cumulative distribution function of the t-distribution with (k_0+k_1-2) degrees of freedom, $t_{\frac{\alpha}{2},k_0+k_1-2}$ is the $1-\frac{\alpha}{2}$ percentile of the t-distribution with (k_0+k_1-2) degrees of freedom, and the variance of $\hat{\delta}$ is given by

$$var(\hat{\delta}) = \sigma^2\left[\frac{1+(m_0-1)\rho}{m_0 k_0} + \frac{1+(m_1-1)\rho}{m_1 k_1}\right], \tag{4}$$

where k_0 and k_1 are the respective numbers of control and treatment clusters, and m_0 and m_1 are

[4]Cluster RCTs can also be analyzed with cluster level outcomes, such as prices, in which only one observation per cluster of the outcome variable is available. In these cases, one would use the standard formulae thought for individual level randomization (Cochran, 1963; Nam, 1973), but where the cost parameters reflect those of the cluster.

[5]For the reader interested in the dual problem, whereby one minimizes the total cost of the RCT subject to achieving a given level of power, please see Appendix B.

the number of sample units per control and treatment clusters respectively.[6,7] Our formulation implicitly assumes that the variance of the two error components, σ_v^2 and σ_ϵ^2 do not change with the number of clusters or the number of units per cluster.

A researcher will want to optimize the design of the cluster RCT by determining the sample that maximizes its power, subject to a budget constraint. We assume that the costs of the RCT are given by:

$$C = (f_0 + v_0 m_0)k_0 + (f_1 + v_1 m_1)k_1, \tag{5}$$

where f_0 and f_1 represent the fixed costs per control and treatment cluster respectively, and v_0 and v_1 represent the costs of control and treatment units respectively (unit costs). In cases in which all units in a treatment cluster are treated, the difference in costs are better reflected in the fixed costs, as one would expect treatment and control clusters to be of the same size in average. This would be the case, for instance, of our first example in which school principals are given a grant to improve the school.

3.2 Numerical Solution

We write the constrained optimization problem that the researcher faces as:

$$\max_{\{m_0,m_1,k_0,k_1\}} \quad T_{k_0+k_1-2}\left(\frac{\delta}{\sqrt{var(\hat{\delta})}} - t_{\frac{\alpha}{2},k_0+k_1-2}\right)$$

$$\text{s.t.}$$

$$B \geq [(f_0 + v_0 m_0)k_0 + (f_1 + v_1 m_1)k_1], \tag{6}$$

where B is the available budget. In its general form, the constrained optimization problem above does not have closed-form solutions. However, it can be solved numerically using two different approaches.

Our preferred approach is an integer grid search with a step of 1 on the four variables m_0, m_1, k_0, k_1, which takes advantage of the structure of the maximization problem to set the lower and upper bounds for the grid. Conservatively, the lower bounds for each variable is 1, and the upper bound for each variable is the largest integer value which satisfies the budget constraint in (6) when all other variables are set to 1. The solution is the set of values of the grid for m_0,

[6] Shen and Kelcey (2020) express the variance differently, but we show in the Appendix C that their formulation is equivalent to formula (4).

[7] It is straightforward to adapt (4) in order to allow the variance of the outcome to differ across treatment and control units: $var(\hat{\delta}) = \left[\sigma_0^2 \frac{1+(m_0-1)\rho}{m_0 k_0} + \sigma_1^2 \frac{1+(m_1-1)\rho}{m_1 k_1}\right]$.

m_1, k_0, k_1 which fulfills the budget constraint and delivers the highest value of the objective function (power).

An alternative to the integer grid search solution is to use Simulated Annealing as optimizer (Corana et al., 1987; Goffe et al., 1994; Goffe, 1996; Xiang et al., 2013). To incorporate the budget constraint, the algorithm searches through values of m_0, m_1, k_1 and finds the corresponding value of k_0 by assuming that the budget constraint holds with equality.

The main advantage of using Simulated Annealing over gradient-based methods is that the algorithm can be prevented from failing when the combination of m_0, m_1, k_1 is such that the implied k_0 is negative, and hence the variance would take negative values. When that is the case, the objective function can be set to give a very low number, which also prevents the algorithm from searching in areas where the implied k_0 is negative. This would be problematic with a gradient-based algorithm which assumes the function to be differentiable. An additional advantage over gradient-based methods is that we can set lower and upper bounds in each of the variables it optimizes over: m_0, m_1, k_1. However, the Simulated Annealing algorithm presents two limitations in this context. First, it does not guarantee convergence to a global optimum, although it is generally more robust to local optima than gradient-based methods. Second, the algorithm typically returns non-integer solutions, whereas sample sizes must be specified in integer values. While one would expect the integer-valued optimum to lie near the real-valued solution, there is no guarantee that it will simply be the nearest integer, as rounding can interact non-trivially with the budget constraint and potentially lead to suboptimal or infeasible allocations.

Although the integer grid search algorithm is slower, it has three main advantages over Simulated Annealing or gradient-based optimizers: First, the solution is a four-tuple of integers, and hence it is implementable without any rounding. Second, the procedure will always provide the global maxima, and third, additional feasibility constraints such as the total number of clusters being less than a certain value, or that the number of units per cluster must be smaller than another given value can easily be incorporated. Because of these advantages, below we report results using the integer grid search algorithm.[8]

[8]We provide a Stata ado file `clusterpowermax` which implements the integer grid search algorithm to find the sample allocation that maximizes power subject to a cost constraint. We also provide `clustercostmin` which minimizes costs subject to a level of power (see Appendix Section B). Alternatively, an R package, `Optimal.sample` (https://brendonmcconnell.github.io/Optimal.sample/) fulfills both tasks using Simulated Annealing as an optimizer. To avoid directly optimizing over the degrees of freedom in the t distribution, the R package uses the Normal distribution to obtain an initial estimate of the total number of clusters, which is then used with the t-distribution in an iterative process, until convergence. The `Optimal.sample` routine is faster than either `clusterpowermax` or `clustercostmin` but we recommend the latter two because of the reasons given in the text. If `Optimal.sample` was used, we recommend trying different initial conditions.

3.3 Pure Cases with Limited Flexibility

Given that the general problem above does not have a closed-form solution, we present below the solutions to simpler cases to gain intuition about the solution to the general problem. In these simpler cases, we abstract from the effect of the degrees of freedom on power and aim at minimizing the variance of the estimator, $\hat{\delta}$. Moreover, we allow the solutions to take non-integer values, which simplifies the optimization by ensuring that the constraint holds with equality.

3.3.1 Homogeneous Unit Costs Within Cluster

Here we consider the case in which the unit cost is the same in treatment and control ($v_0 = v_1 = v$), and we simplify the optimization by using the restriction that the number of units per cluster is also the same in treatment and control ($m_0 = m_1 = m$).[9] We allow for the fixed costs per cluster to be different between treatment and control ($f_0 \neq f_1$), and we solve for the number of treatment and control clusters (k_0, k_1), conditional on m. In this more restricted scenario, we substitute ($m_0 = m_1 = m$) and ($v_0 = v_1 = v$) in the general formula of the variance (4), and rewrite the cost function as $C = (f_0 + vm)k_0 + (f_1 + vm)k_1$, giving the optimization problem as:

$$\begin{aligned} \min_{\{k_0,k_1\}} \quad & \sigma^2 \frac{1+(m-1)\rho}{m}\left[\frac{1}{k_0}+\frac{1}{k_1}\right] \\ & \text{s.t.} \\ & B = (f_0 + vm)k_0 + (f_1 + vm)k_1 \end{aligned} \tag{7}$$

where the only unknowns are k_0 and k_1 because the number of units to be sampled per each cluster is exogenously given by m.

The solution to the optimization problem yields the following optimality condition:

$$\frac{k_1}{k_0} = \sqrt{\frac{(f_0 + vm)}{(f_1 + vm)}}, \tag{8}$$

which clarifies that cheaper clusters will be over-sampled, but that the difference between the number of treatment and control clusters will be less than proportional to the difference in costs.

Using the cost function formula, we can write the optimal values of k_0 and k_1 as functions

[9] It should be noted that even if ($v_0 = v_1$), we would not expect ($m_0 = m_1$) to hold in the unconstrained optima.

of the model parameters:

$$k_0^* = \frac{B}{(f_0 + vm) + \sqrt{(f_0 + vm)}\sqrt{(f_1 + vm)}}, \qquad \text{and} \tag{9}$$

$$k_1^* = \frac{B}{(f_1 + vm) + \sqrt{(f_0 + vm)}\sqrt{(f_1 + vm)}} \tag{10}$$

The level of power can be obtained by substituting (9) and (10) in (3). Note that the above closed-form solutions were obtained using the assumption that the number of units to be sampled within each cluster, m, was exogenously given. In practice, it is straightforward to circumvent this assumption by performing a grid search on m – compute the optimal values of k_0 and k_1 for different values of m, and choose the one that maximizes power. Hence, the key assumptions for this special case to be useful are $m_0 = m_1$ and $v_0 = v_1$.

3.3.2 Homogeneous Fixed Costs Per Cluster

In this subsection, we consider the case in which the fixed cost per cluster is the same in treatment and control ($f_0 = f_1 = f$), and we simplify the optimization by using the restriction that the number of clusters is also the same in treatment and control ($(k_0 = k_1 = k)$). We allow for the unit costs within cluster to be different between treatment and control ($v_0 \neq v_1$), and we solve for the number of treatment and control units per cluster (m_0, m_1), conditional on k. These simplifications allow us to re-write the cost function as $C = (f + v_0 m_0)k_0 + (f + v_1 m_1)k_1 = 2fk + v_0 m_0 k + v_1 m_1 k$.

In this case, we write the constrained optimization problem as:

$$\min_{\{m_0, m_1\}} \quad \sigma^2 \frac{1}{k}\left[\frac{1 + (m_0 - 1)\rho}{m_0} + \frac{1 + (m_1 - 1)\rho}{m_1}\right]$$

$$\text{s.t.}$$

$$B = 2fk + v_0 m_0 k + v_1 m_1 k \tag{11}$$

The solution to the optimization problem yields the following optimality condition:

$$\frac{m_1}{m_0} = \sqrt{\frac{v_0}{v_1}}, \tag{12}$$

which clarifies that the over-sample of the cheaper units is less than proportional to the difference in costs. Using the cost function formula, we can write the optimal values of m_0 and m_1 as

functions of the model parameters:

$$m_0^* = \frac{B - 2fk}{(v_0 + \sqrt{v_0}\sqrt{v_1})k} \qquad \text{and} \qquad m_1^* = \frac{B - 2fk}{(v_1 + \sqrt{v_0}\sqrt{v_1})k} \tag{13}$$

The level of power can be obtained by substituting the relations in equations (13) into (3). Similar to the case above, one can circumvent the assumption of a fixed, exogenously given k, by running a grid search over different values of k, – compute the optimal values of m_0 and m_1 for different values of k, and choose the one that maximizes power. Hence, the actual important assumptions for this special case to be useful are $k_0 = k_1$ and $f_0 = f_1 = f$.

4 Empirical Examples

This section applies the integer grid search method described in Section 3.2 to prominent archetypes of cluster RCTs to obtain realistic estimates of the power gains that can be achieved when choosing the sample to maximize power when the costs differ between treatment and control.[10] Whenever possible, we use actual costs from the experiments, but make realistic assumptions when they are not available. It should be noted that we do not replicate all the features of the previous studies, and hence our estimates should not be understood as the power that previous studies could have gained, but more like benchmark gains that can be obtained in a typical cluster RCT. See Appendix Section A.2 for a detailed explanation of the parameter values used in the computations below.

The sample size estimates that we will report are for a double-sided test of means at 5% significance. We set an effect size δ of 0.25, and a standard deviation, σ, of 1, which is just above what Cohen denoted as a small effect size (Cohen, 1988).

4.1 Heterogeneous Fixed Costs per Cluster

In many cluster RCTs, the treatment costs are divorced from the sampling costs. The sampling costs involve the time and material costs of recruiting, testing, and interviewing subjects, while the treatment costs are fixed per cluster and do not depend on the number of sampled subjects. An example of such an RCT is a school grant program that aims at increasing school resources and improves students' outcomes.[11] The sampling costs will be the same in treatment and control clusters ($v_0 = v_1 = v$) , while the fixed cost of including a treatment cluster, f_1, are larger than

[10] In Appendix A.1, we present the results of a simulation exercise that confirms the validity of the power calculations given for each of the empirical examples reported in Tables 1 to 3.

[11] The amount of the grant might depend on the number of children in the school but not on the number of children sampled, hence the cost of the grant is fixed per cluster.

the control cluster fixed costs, f_0, because the fixed cost treatment cluster includes the school grant. The cost function that represents this scenario is given by $C = (f_0+vm_0)k_0+(f_1+vm_1)k_1$, which is obtained from substituting $v_0 = v_1 = v$ in (5).

We build our illustrative example based on Luo et al. (2019) in which one of the treatment arms considered is a school grant provided for rural primary schools in five prefectures of western China. We conduct sample size calculations for students' blood hemoglobin concentration, which is their primary outcome. Based on Luo et al. (2019)'s data, we estimate the intra-cluster correlation coefficient, ρ, to be 0.27. This relatively large value is consistent with findings from other school-level studies in Sub-Saharan Africa, where intra-cluster correlation (ICC) coefficients have been reported to range between 0.15 and 0.60 (Kelcey et al., 2016).

Panel B of Table 1 reports our sample size estimates for this school grant program taking into account that the fixed cost is much larger in treatment than in control clusters ($f_1 > f_0$) but the within-cluster unit cost is the same ($v_1 = v_0 = v = \$9.36$). Column 2 of Table 1 reports the estimates for our benchmark scenario based on their cost figures ($f_0 = \$\,189, f_1 = \$\,1776.4$). We calibrate the available budget to be $\$148,841$ so that the solution to the power maximization problem (Panel B) provides 80% power in this benchmark case (column 2).

Table 1: Heterogenous Fixed Costs per Cluster – School Grant Program

	(1)	(2)	(3)
Fixed Cost Control (f_0)	189	189	189
Fixed Cost Treatment (f_1)	1,000	1,776	3,000
Unit Cost (v)	9.36	9.36	9.36
Available Budget ($)	148,841	148,841	148,841
A.) Balanced Allocation			
$k_0 = k_1 = k$	105	65	41
$m_0 = m_1 = m$	12	17	23
Power	0.880	0.715	0.528
B.) Optimal Allocation			
k_0	199	170	144
k_1	84	53	34
m_0	7	7	7
m_1	18	23	32
Power	0.915	0.800	0.651
Power Improvement vs Approach A	0.036	0.085	0.123
Value of Improvement vs Approach A ($)	19,421	33,059	48,182
Value of Improvement as Percent of Budget	13.0%	22.2%	32.4%

Notes: The optimization algorithm used in both panel A and panel B is based on a grid-search, integer-optimization approach. This yields a directly implementable solution without the need for rounding. The values for number of individuals per cluster (m) and number of clusters (k) are those that achieve 80% power at 5% significance for the cost parameters specified in the top 3 rows. Other assumed parameters: effect size 0.25, standard deviation 1, intra-cluster correlation (ρ) = 0.27. We calculate the *Value of the improvement vs Approach A* by solving a cost minimization problem for panel A in order to achieve the same power as calculated in Panel B, and then calculating the budget required to pay for this new optimal sample allocation.

The optimal number of treatment clusters is much smaller than the number of control clusters ($k_0^* = 170, k_1^* = 53$), which reflects the fact that treatment clusters are much more expensive because their fixed cost includes the school grant. Interestingly, to partially compensate for this, the number of students per cluster is much larger in treatment than control clusters, $m_1^* = 23 > m_0^* = 7$. Hence, we find that $k_0^* > k_1^*$ but $m_0^* < m_1^*$. The same insights hold for columns 1 and 3, which assume smaller and larger values respectively for the treatment cluster fixed cost, f_1.

To quantify the power gains achieved by using the optimal sample allocation —as opposed to a balanced design with an equal number of clusters and units per cluster in the treatment and control arms— we apply the same integer grid search optimization method described in Section 3.2, but impose the constraints $k_0 = k_1$ and $m_0 = m_1$. The results, which are reported in Panel A, show that in the benchmark case (column 2), the optimal allocation yields an increase in power of 8.5 percentage points compared to the balanced design. This gain is higher, 12.2 percentage points, in column (3) where the difference in costs between treatment and control clusters is larger than in column (2), but smaller, 3.6 percentage points, in column (1) where the difference in costs is smaller.

4.2 Heterogeneous Unit Costs per Cluster

In this example, we consider the case where fixed costs per cluster are equal in treatment and control, but unit costs are different. This leads to a cost function of the form $C = (f + v_0 m_0)k_0 + (f + v_1 m_1)k_1$. A real life example is one of an unconditional cash transfer in which only some households in the treatment clusters are given the cash transfer (see for instance, Haushofer and Shapiro (2016)). In this type of RCT, treatment and control sampled households will have very different costs because the cost of any sampled household includes identification, enrollment, and interviewing costs, whilst the costs of treatment households also include the cash transfer. There is a fixed cost per cluster, representing the costs of transporting the interviewing field team between clusters, which is the same in treatment and control clusters.

We conduct sample size calculations for household consumption per capita, a key outcome in the evaluation of cash transfer programs. We assume an ICC coefficient of 0.05, corresponding to the average value reported by Seidenfeld et al. (2023) for household consumption per capita across a sample of four Sub-Saharan African countries.

Panel B of Table 2 reports our sample size estimates of the cash transfer program, taking into account that the cost per household in the treatment arm is much higher than in the

control arm. We have calibrated the available budget to be $\$260,855$ so that the power in our benchmark scenario is 0.80 (column 2, Panel B). As expected, the number of households in the control arm is larger than in the treatment arm. For instance, the benchmark case (column 2) involves 6 households per control cluster while only 2 per treatment cluster. Interestingly, to partially compensate for this, the number of treatment clusters is slightly larger than the number of control clusters (in the benchmark scenario: 95 vs. 88).

Table 2: Heterogenous Unit Costs per Cluster – Unconditional Cash Transfer

	(1)	(2)	(3)
Fixed Cost (f)	250	250	250
Unit Cost Control (v_0)	100	100	100
Unit Cost Treatment (v_1)	500	854	1,200
Available Budget ($)	260,855	260,855	260,855
A.) Balanced Allocation			
$k_0 = k_1 = k$	89	77	84
$m_0 = m_1 = m$	4	3	2
Power	0.871	0.721	0.603
B.) Optimal Allocation			
k_0	94	88	79
k_1	98	95	73
m_0	7	6	6
m_1	3	2	2
Power	0.908	0.799	0.707
Power Improvement vs Approach A	0.037	0.079	0.104
Value of Improvement vs Approach A ($)	32,100	51,856	69,400
Value of Improvement as Percent of Budget	12.3%	19.9%	26.6%

Notes: The optimization algorithm used in both panel A and panel B is based on a grid-search, integer-optimization approach. This yields a directly implementable solution without the need for rounding. The values for number of individuals per cluster (m) and number of clusters (k) are those that achieve 80% power at 5% significance for the cost parameters specified in the top 3 rows. Other assumed parameters: effect size 0.25, standard deviation 1, intra-cluster correlation (ρ) = 0.05. We calculate the *Value of the improvement vs Approach A* by solving a cost minimization problem for panel A in order to achieve the same power as calculated in Panel B, and then calculating the budget required to pay for this new optimal sample allocation.

As we did in the example with heterogeneous fixed costs per cluster, we report in Panel A the level of power that is feasible to obtain with the same budget while using a balanced design. In the benchmark case, column 2, the optimal allocation leads to 7.9 percentage points higher power than the balanced allocation. The equivalent power gain in column 3, where the treatment unit costs are 12 times higher than the control ones, is 10.4 percentage points, while only 3.7 percentage points higher in column 1 where the treatment unit costs are 5 times higher than the control ones.

4.3 Heterogeneous Unit and Fixed Costs per Cluster

Another prominent example of cluster RCTs in which treatment observations are much more expensive than control ones are graduation programs, in which extremely poor households are given a very large transfer, typically including a productive asset, training, and temporary income support, combined with access to financial services.[12] As in the previous example, we conduct sample size calculations for household consumption per capita, and assume an ICC coefficient of 0.05.

Table 3 reports the results for the graduation example, in which both the fixed cost per cluster, and the unit cost per household within cluster are larger in the treatment than the control arm. As in previous examples, we calibrate the budget to be $\$526,300$ so that in the benchmark case (column 2), the optimal sample size allocation delivers a power of 0.80.

As expected, the number of control clusters is much larger than the number of treatment clusters (190 vs. 19), but interestingly, in columns (1) to (3) there are more individuals per treatment cluster than per control cluster, despite the unit cost being higher in treatment than control clusters. Intuitively, there are so many more control than treatment clusters, that in order to partially offset this, it is optimal to sample more households per treatment cluster despite each being more expensive than their control counterparts. In column (4), in which f_0 is four times that of column (2), the difference between the number of control and treatment clusters is smaller than in the other columns (although still $k_0 > k_1$), and in that case we find, contrary to the other columns, that the number of individuals per control cluster is slightly larger than per treatment cluster (13 vs. 12).

As in the previous two examples, we compare the power achieved with the optimal allocation compared to the balanced one. Across the four columns, the gains in power are between 17 and 19 percentage points. It must be noted though, that these gains in power might come at the expense of a rejection rate higher than the nominal rate of 0.05, and additional constraints in the optimal allocation might be needed to restore the nominal rejection rate of 0.05 (see Section 5 for a detailed explanation of this issue).

[12] See, for instance, Banerjee et al. (2015) and Bandiera et al. (2017).

Table 3: Heterogenous Fixed and Unit Costs per Cluster – Graduation Program

	(1)	(2)	(3)	(4)
Fixed Cost Control (f_0)	125	250	500	1,000
Fixed Cost Treatment (f_1)	18,000	18,000	18,000	18,000
Unit Cost Control (v_0)	100	100	100	100
Unit Cost Treatment (v_1)	2,150	2,150	2,150	2,150
Available Budget ($)	994,017	994,017	994,017	994,017
A.) Balanced Allocation				
$k_0 = k_1 = k$	22	23	24	18
$m_0 = m_1 = m$	12	11	10	16
Power	0.613	0.610	0.603	0.592
B.) Optimal Allocation				
k_0	308	190	119	89
k_1	19	19	18	18
m_0	4	6	9	13
m_1	12	12	13	12
Power	0.809	0.800	0.785	0.763
Power Improvement vs Approach A	0.196	0.190	0.183	0.171
Value of Improvement vs Approach A ($)	569,475	526,300	486,550	432,900
Value of Improvement as Percent of Budget	57.3%	52.9%	48.9%	43.6%

Notes: The optimization algorithm used in both panel A and panel B is based on a grid-search, integer-optimization approach. This yields a directly implementable solution without the need for rounding. The values for number of individuals per cluster (m) and number of clusters (k) are those that achieve 80% power at 5% significance for the cost parameters specified in the top 3 rows. Other assumed parameters: effect size 0.25, standard deviation 1, intra-cluster correlation (ρ) = 0.05. We calculate the *Value of the improvement vs Approach A* by solving a cost minimization problem for panel A in order to achieve the same power as calculated in Panel B, and then calculating the budget required to pay for this new optimal sample allocation.

4.4 Robustness

The estimates of the gains in power due to the optimal allocation as compared to the balanced one have been obtained with particular values of the ICC, 0.27 for Table 1, and 0.05 for Tables 2, and 3. To show that our estimates of power gain are valid more broadly, we present respective power curves for both equal and optimal allocations at different values of the ICC.

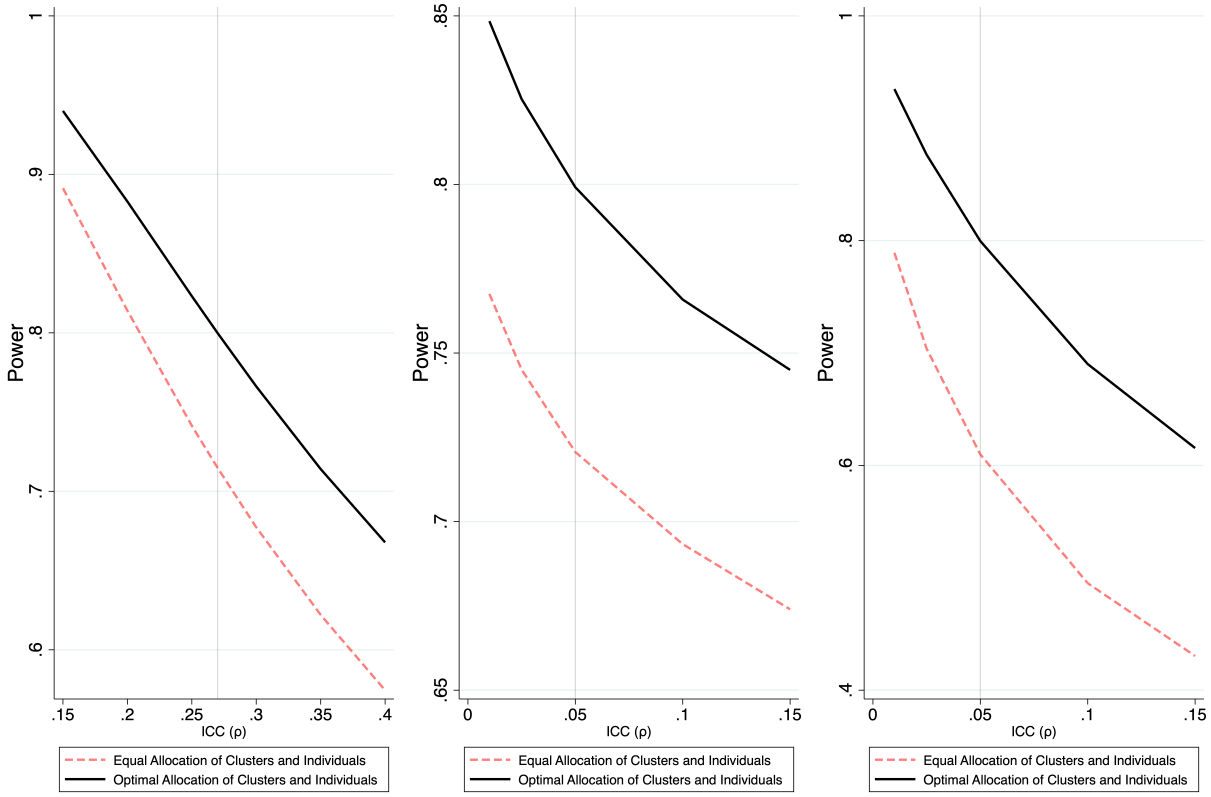
Figure 1 confirms that the power gains we document from our optimal allocation approach for specific ICCs are valid for a wide range of ICCs, and at least for these three examples, the gains in power increase with the value of the ICC.

4.5 Valuing the Improvement in Power

An alternative way to conceptualize the improvement in power using the approach we develop in this paper compared to the simple, balanced design is to ask the following: How much larger a budget would be required to achieve the power attained using our approach (Panel B) if one implemented a balanced design (Panel A)? This is a useful alternative approach as it enables us to better grasp the value of our approach.

To answer this question, we take the following approach. We start with the power achieved under the optimal allocation (Panel B). We then return to the equal allocation approach. Here

Figure 1: Calculating Power for Various Values of the ICC



(a) Heterogeneous Fixed Costs

(b) Heterogeneous Unit Costs

(c) Heterogeneous Fixed and Unit Costs

Notes: In these figures, we present power curves for our baseline specifications from each table – Column 2 from Table 1, Column 2 from Table 2 and Column 2 from Table 3. With the exception of the ICC, we fix all parameters as presented in the respective tables. In each figure, we mark the baseline value of the ICC with a thin, gray, vertical line.

we solve the dual problem to power maximization – cost minimization. Specifically, we solve the cost minimization problem for the equal allocation case subject to achieving the power obtained in the Panel B setting. This will always be weakly more expensive to do, as in the equal allocation setting we constrain $k_0 = k_1 = k$ and $m_0 = m_1 = m$. The resulting solution to this constrained cost minimization problem yields the panel B power at the lowest budget. Subtracting this new budget from the original yields the penultimate row in Tables 1-3, the value of the power improvement in dollars. In all three of our archetypal examples, the value of the power improvement is sizable. Expressed in terms of the original budget, these are respectively 22%, 20% and 53% for the three case studies. Put differently, the value of our approach is akin to using the standard, balanced approach but being granted a budget of between 20% to 53% larger.

5 Empirical Challenges

In this section, we discuss several empirical challenges that arise from the unbalanced design proposed in this paper and provide guidance on how to address them. First, we examine feasibility constraints related to the total number of clusters available for the experimental design. Second, we consider inferential issues that may emerge when the number of treatment clusters is small. Finally, we briefly review a set of additional complications that may arise in the context of unbalanced designs.

5.1 Feasibility Constraints

Optimizing the sample design to take into account that treatment clusters are more expensive than control ones might bring some challenges in practice. One challenge is that the total number of clusters will need to be larger than in a balanced design (to compensate for the loss of power due to having a different number of treatment and control clusters). For instance, the optimal design requires 223 clusters in column (2) of Table 1, while only 158 in a balanced design that achieves 80% power.[13] Hence, the optimal design might not be feasible because the required total number of clusters might be larger than available in the population, or because of some logistical constraint. Fortunately, a constraint in the total number of clusters (as well as in the number of individuals per cluster) can easily be incorporated in the integer grid search algorithm that we describe in Section 3.2.

Table 4 provides the optimal design for the parameters in Column (2) from Tables 1 to 3, when we constrain the total number of clusters below several upper bounds. For the top and bottom panel, which refer to cases in which the fixed cost is higher for treatment than control clusters, the more binding the constraint is, the closer the number of treatment and control clusters are.[14,15] To partially compensate for this, the number of individuals per cluster increases. In the scenarios that we consider in Table 4, power does not seem to decrease too quickly with the constraint. For instance, in the top example, the unconstrained number of clusters is 223 but when we constrain the number of clusters to 150, the resulting power is 0.766. Note that the design with a constrained number of clusters will require a larger budget to achieve the same power than in the unconstrained case.[16]

[13] Note that an allocation with 158 clusters and 17 individuals per cluster will not be feasible given the budget constraint. The calculation assumes all the same parameters as in Table 1.

[14] Reducing the wedge between the number of treatment and control clusters can help to reduce bias when the number of units per cluster vary and outcomes are correlated with cluster size (Middleton, 2008).

[15] This is not necessarily the case for the middle panel, since fixed costs are constant across clusters, and the

Table 4: The Consequences of Imposing a Constraint on Total Number of Clusters

	(1)	(2)	(3)	(4)	(5)
Constraint, κ	k_0	k_1	m_0	m_1	Power
A.) Heterogenous Fixed Costs per Cluster (Table 1)					
No Constraint	170	53	7	23	0.800
200	145	55	8	25	0.798
175	118	57	11	24	0.788
150	93	57	16	30	0.766
125	70	55	28	38	0.721
100	52	48	55	60	0.642
75	38	37	107	109	0.526
50	25	25	213	213	0.375
B.) Heterogenous Variable Costs per Cluster (Table 2)					
No Constraint	88	95	6	2	0.799
175	75	93	8	2	0.799
150	82	65	7	3	0.798
125	59	66	10	3	0.791
100	50	50	13	4	0.780
75	33	42	19	5	0.751
50	26	24	24	9	0.690
C.) Heterogenous Fixed and Variable Costs per Cluster (Table 3)					
No Constraint	190	19	6	12	0.800
200	170	19	7	12	0.799
175	154	19	8	12	0.799
150	129	19	10	12	0.797
125	105	19	9	13	0.794
100	82	18	13	14	0.788
75	56	19	19	13	0.774
50	32	18	37	14	0.737

Notes: Using the baseline scenario from Tables 1, 2, and 3, we impose constraints of the form $k_0 + k_1 \leq \kappa$. As κ decreases, the constraint becomes increasingly severe. The optimization algorithm used here is based on a constrained grid-search, integer-optimization approach. This yields a directly implementable solution without the need for rounding.

5.2 Number of Treatment Clusters and Inference

In several instances, the optimal design will involve allocating more clusters to the treatment than to the control arm. Although such asymmetry is explicitly incorporated into the formula used to compute statistical power, standard inference procedures based on cluster-robust standard errors can yield misleading results when the number of treated clusters is too small. In such situations, standard errors tend to be underestimated, inflating test statistics and leading to over-rejection of the null hypothesis (MacKinnon et al., 2023).[17]

There is no clear threshold below which over-rejection arises, as the severity of the problem depends on factors such as the sizes of the treated and control clusters and the distribution of covariates. The problem typically diminishes rapidly as the number of treated clusters increases.

variation arises from differences in unit costs.

[16]The Stata package that we provide allows to simultaneously constrain the total number of clusters, and the number of individuals per cluster.

[17]An additional challenge is that a small number of treated clusters may make it more difficult to achieve balance —both on observed and unobserved characteristics— between the treatment and control groups, potentially undermining internal validity.

In contrast, increasing the total number of clusters while holding the number of treated clusters fixed does not resolve the issue and can, in fact, exacerbate over-rejection.

To shed light on this challenge, we employ simulation methods, as described in Appendix A.1.2, to estimate the size of the t-test (the probability of rejecting the null hypothesis of no treatment effect when it is true), when using cluster-robust variance estimators —reflecting typical empirical practice. The sizes are reported in the final column of Table A2, which should be compared to a nominal value of 0.05. For the example of heterogeneous fixed costs per cluster (top panel), the rejection probabilities are very close to 0.05 for the first two cases in which the numbers of treated clusters are 84 and 53, but there is mild over-rejection (0.058 vs. 0.05) when the number of treated clusters is 34 and the ratio treated to control clusters is 0.24. While the example with heterogeneous unit costs per cluster (middle panel) presents no evidence of over-rejection, the scenario with heterogeneity in both fixed and unit costs (bottom panel) —in which the number of treated clusters is below 20 and the ratios of treated to control clusters are between 0.06 and 0.2— exhibits elevated rejection probabilities between 0.064 and 0.071.

When there is concern that the number of treated clusters may be too small to use cluster-robust standard errors, the researcher can determine the optimal design subject to a lower bound in the number of treated clusters.[18] For instance, constraining the number of treated clusters to be at least 40 in the third line of Panel A in Table A2 would lead to an estimated size of 0.05 and power of 0.62, which is still higher than would have been delivered with the balanced design (0.53 according to column (3) of Table 1).[19] Similarly, constraining k_1 to be at least 32 in the second row of Panel C leads to an estimated size of 0.056 and power of 0.70 significantly higher than the power achieved with the balanced design (0.61 according to column 2 of Table 3).[20]

An alternative approach would be to employ at the analysis stage inference methods that are explicitly designed to be robust to the problem of few treated clusters. Several such methods exist, and the choice among them depends on the specific features of the empirical exercise (see MacKinnon and Webb (2018), MacKinnon et al. (2023), and Alvarez et al. (2025)). Importantly, power calculations should be based on the inference method that will ultimately be used in the analysis. However, for many of these robust inference procedures, there are currently no widely accepted methods for conducting power analyses, and simulation methods might be needed.[21]

[18]The accompanying Stata package provides a simulation-based estimate of the rejection rate of the cluster-robust t-test, as illustrated in Table A2. It also allows users to impose a lower bound in the number of treated clusters, as well as on the number of units per cluster, to compute a constrained-optima design.

[19]The value of the other parameters are $k_0 = 88$, $k_1 = 40$, $m_0 = 8$, and $m_1 = 15$.

[20]The size of the equal allocation design test is 0.052. When the lower bound for k_1 is 32, the value of the other parameters are $k_0 = 87$, $k_1 = 32$, $m_0 = 6$, $m_1 = 5$.

[21]See McConnell and Vera-Hernandez (2025) for a guide on computing power through simulation methods.

In any case, even if the inference problem was solved, it would not be advisable to run a cluster RCT with a very small number of treatment clusters as it might be difficult to argue that observable and unobservable variables are balanced between treatment and control clusters.

5.3 Additional Issues

As indicated in Section 3.1, our formulation implicitly assumes that the variances of the cluster level error term and of the individual level error term do not change when the number of clusters or units per cluster change. Although ideally the population is fixed a priori and independent of the design of the experiment, in practice, researchers may broaden their target population to reach the required number of clusters. This might lead to an increase in the variance of the cluster level error term, as well as variability in the number of units per cluster, both of which would decrease power. Alternatively, the researcher may impose a constraint on the total number of clusters, setting it equal to the number available in the population of interest, and then solve for the constrained optimum, as discussed above.

Practical challenges might also emerge when the optimal design delivers a relatively small number of units per cluster, which might lead to the complete loss of clusters in case of attrition. To prevent this, the researcher may prefer to optimize the design subject to a lower bound on the number of units per cluster.

An important factor that may influence the optimal trade-off between the number of clusters and the number of units per cluster is the availability of baseline data and covariates, which are often more readily available at the cluster level. For example, information on village or school characteristics may already exist in census records or administrative data from the Ministry of Education.

Researchers often wish to consider multiple outcomes rather than focusing solely on a single primary outcome. The ICCs and effect sizes are likely to vary across outcomes, and consequently, the optimal experimental design will also differ. This challenge is not unique to our approach; even in cluster RCTs with homogeneous costs, the optimal number of clusters and individuals per cluster typically varies across outcomes. Efficiently addressing multiple outcomes requires adopting a definition of power that explicitly accounts for outcome multiplicity, incorporating assumptions about the dependence structure between outcomes, and adjusting the level of significance; see Porter (2018) for a detailed elaboration of these considerations.[22]

[22] In some cases, a particular outcome may necessitate both the largest number of clusters per treatment arm and the largest number of individuals per cluster. Even in such cases, however, the level of significance must be adjusted downwards to maintain the desired power once corrections for multiple hypothesis testing are applied at the analysis stage.

6 Conclusion

In cluster RCTs, researchers commonly use a balanced design, in which the same number of treatment and control clusters and units within treatment and control clusters are sampled. However, in many cluster RCTs, treatment clusters and/or sampled units within treatment clusters are more expensive than control ones because the former incorporate the costs of implementing the intervention. Under these cost differences, the researcher can maximize the power subject to a cost constraint (or minimize the costs subject to achieving a pre-determined level of power) by allowing the number of clusters and number of sampled units within clusters to be different in treatment and control. We develop methods to optimally compute these four sample parameters, contributing to the existing literature by allowing for full flexibility of the solution. We focus the paper on the primal problem of maximizing power subject a cost constraint, but our method can also be applied to the dual problem of minimizing costs subject to a level of power, as we do in the Appendix.

Due to practical or statistical reasons, the researcher might want to deviate from the optimal design that maximizes power, and impose a lower bound on the number of treated clusters, the number of units per cluster, or an upper bound on the total number of clusters. An advantage of the integer optimization method that we develop is that it can easily handle such bounds.

To illustrate the relevance of our methods, we apply them to three prominent examples from the development economics literature, each characterised by a distinct cost structure: one in which fixed costs per cluster differ between treatment and control groups but unit costs are the same; another in which unit costs differ but fixed costs are equal; and a third in which both fixed and unit costs vary between treatment and control.

Using realistic cost estimates, we find substantial power gains with respect to the balanced design, of between 7.9 and 19.0 percentage points. As expected, we observe some compensation between clusters and units per cluster. For instance, if it is optimal to have more control than treatment clusters, then the number of units per treatment cluster may be larger than that of controls. However, this is not necessarily the case when both the fixed cost per cluster and the unit cost are higher in treatment than control. In such cases, depending on the specific cost parameters, it might be optimal to have more control clusters, as well as more units sampled per control cluster.

We obtain our results using realistic cost estimates based on the specific examples that we study and reasonable assumptions on parameters which are unknown to us, and comparing the power of our method with that of the balanced design. It should be noted that we do not replicate

all the features of the studies from which we derive our examples, and hence our results should not be understood as what the previous studies could have gained, but more like benchmark power gains that can be obtained in a typical cluster RCT. We further consider the benefits of our approach by attaching a monetary value to the power improvements. We compute that to obtain the same power as we do with our method, a balanced design with the same number of treatment and control clusters would need to increase the budget by between 20% to 53%. These cost implications underscore the value of our approach.

References

Alvarez, L., B. Ferman, and K. Wüthrich (2025): "Inference with few treated units," .

Baird, S., J. A. Bohren, C. McIntosh, and B. Özler (2018): "Optimal Design of Experiments in the Presence of Interference," *The Review of Economics and Statistics*, 100, 844–860.

Bandiera, O., I. Barankay, and I. Rasul (2011): "Field Experiments with Firms," *Journal of Economic Perspectives*, 25, 63–82.

Bandiera, O., R. Burgess, N. Das, S. Gulesci, I. Rasul, and M. Sulaiman (2017): "Labor Markets and Poverty in Village Economies," *The Quarterly Journal of Economics*, 132, 811–870.

Banerjee, A., E. Duflo, N. Goldberg, D. Karlan, R. Osei, W. Parienté, J. Shapiro, B. Thuysbaert, and C. Udry (2015): "A multifaceted program causes lasting progress for the very poor: Evidence from six countries," *Science*, 348.

Banerjee, A. V., S. Chassang, S. Montero, and E. Snowberg (2020): "A Theory of Experimenters: Robustness, Randomization, and Balance," *American Economic Review*, 110, 1206–30.

Burlig, F., L. Preonas, and M. Woerman (2020): "Panel data and experimental design," *Journal of Development Economics*, 144, 102458.

Burtless, G. (1995): "The Case for Randomized Field Trials in Economic and Policy Research," *Journal of Economic Perspectives*, 9, 63–84.

Button, K. S., J. P. A. Ioannidis, C. Mokrysz, B. A. Nosek, J. Flint, E. S. J. Robinson, and M. R. Munafò (2013): "Power failure: why small sample size undermines the reliability of neuroscience," *Nature Reviews Neuroscience*, 14, 365–376.

Carneiro, P., S. Lee, and D. Wilhelm (2019): "Optimal data collection for randomized control trials," *The Econometrics Journal*, 23, 1–31.

Chassang, S., G. Padró I Miquel, and E. Snowberg (2012): "Selective Trials: A Principal-Agent Approach to Randomized Controlled Experiments," *American Economic Review*, 102, 1279–1309.

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

Cohen, J. (1988): *Statistical power analysis for the behavioral sciences*, Lawrence Erlbaum Associates.

Corana, A., M. Marchesi, C. Martini, and S. Ridella (1987): "Minimizing multimodal functions of continuous variables with the "simulated annealing" algorithm," *ACM Transactions on Mathematical Software (TOMS)*, 13, 262–280.

Duflo, E. (2006): "Field Experiments in Development Economics," Tech. rep., MIT.

——— (2020): "Field Experiments and the Practice of Policy," *American Economic Review*, 110, 1952–73.

Duflo, E., R. Glennerster, and M. Kremer (2007): "Using Randomization in Development Economics Research: A Toolkit," Elsevier, vol. 4 of *Handbook of Development Economics*, chap. 61, 3895 – 3962.

Gelman, A. and J. Carlin (2014): "Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors," *Perspectives on Psychological Science*, 9, 641–651, pMID: 26186114.

Glennerster, R. and K. Takavarasha (2013): *Running Randomized Evaluations: A Practical Guide*, Princeton University Press.

Goffe, W. L. (1996): "SIMANN: A Global Optimization Algorithm using Simulated Annealing," *Studies in Nonlinear Dynamics & Econometrics*, 1.

Goffe, W. L., G. D. Ferrier, and J. Rogers (1994): "Global optimization of statistical functions with simulated annealing," *Journal of Econometrics*, 60, 65–99.

Hahn, J., K. Hirano, and D. Karlan (2011): "Adaptive Experimental Design Using the Propensity Score," *Journal of Business & Economic Statistics*, 29, 96–108.

Hamermesh, D. S. (2013): "Six Decades of Top Economics Publishing: Who and How?" *Journal of Economic Literature*, 51, 162–72.

Haushofer, J. and J. Shapiro (2016): "The Short-term Impact of Unconditional Cash Transfers to the Poor: Experimental Evidence from Kenya," *The Quarterly Journal of Economics*, 131, 1973–2042.

Hausman, J. and D. Wise, eds. (1985): *Social Experimentation*, University of Chicago Press.

Heckman, J. J. and J. A. Smith (1995): "Assessing the Case for Social Experiments," *Journal of Economic Perspectives*, 9, 85–110.

Ioannidis, J. P. A. (2005): "Why Most Published Research Findings Are False," *PLoS Medicine*, 2.

Ioannidis, J. P. A., T. D. Stanley, and H. Doucouliagos (2017): "The Power of Bias in Economics Research," *The Economic Journal*, 127, F236–F265.

Karlan, D. and J. Appel (2016): *Failing in the Field: What We Can Learn When Field Research Goes Wrong*, Princeton University Press.

Kelcey, B., Z. Shen, and J. Spybrook (2016): "Intraclass Correlation Coefficients for Designing Cluster-Randomized Trials in Sub-Saharan Africa Education," *Evaluation Review*, 40, 500–525, pMID: 27784814.

Levitt, S. D. and J. A. List (2009): "Field experiments in economics: The past, the present, and the future," *European Economic Review*, 53, 1 – 18.

Leyrat, C., K. E. Morgan, B. Leurent, and B. C. Kahan (2017): "Cluster randomized trials with a small number of clusters: which analyses should be used?" *International Journal of Epidemiology*, 47, 321–331.

List, J. and I. Rasul (2011): "Field Experiments in Labor Economics," Elsevier, vol. 4A, chap. 2, 103–228, 1 ed.

List, J., S. Sadoff, and M. Wagner (2011): "So you want to run an experiment, now what? Some simple rules of thumb for optimal experimental design," *Experimental Economics*, 14, 439–457.

List, J. A. (2011): "Why Economists Should Conduct Field Experiments and 14 Tips for Pulling One Off," *Journal of Economic Perspectives*, 25, 3–16.

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, 231–248.

Luo, R., G. Miller, S. Rozelle, S. Sylvia, and M. Vera-Hernández (2019): "Can Bureaucrats Really Be Paid Like CEOs? Substitution Between Incentives and Resources Among School Administrators in China," *Journal of the European Economic Association*, 18, 165–201.

MacKinnon, J. G. and M. D. Webb (2018): "The wild bootstrap for few (treated) clusters," *The Econometrics Journal*, 21, 114–135.

MacKinnon, J. G., M. Ørregaard Nielsen, and M. D. Webb (2023): "Cluster-robust inference: A guide to empirical practice," *Journal of Econometrics*, 232, 272–299.

McConnell, B. and M. Vera-Hernandez (2025): "Going beyond simple sample size calculations: a practitioner's guide," *Fiscal Studies*, 46.

McKenzie, D. (2012): "Beyond baseline and follow-up: The case for more T in experiments," *Journal of Development Economics*, 99, 210 – 221.

——— (2025): "Designing and Analyzing Powerful Experiments: Practical Tips for Applied Researchers," *Fiscal Studies*, 46.

Middleton, J. A. (2008): "Bias of the regression estimator for experiments using clustered random assignment," *Statistics & Probability Letters*, 78, 2654–2659.

Nam, J.-M. (1973): "Optimum Sample Sizes for the Comparison of the Control and Treatment," *Biometrics*, 29, 101–108.

Olken, B. A. (2020): "Banerjee, Duflo, Kremer, and the Rise of Modern Development Economics," *The Scandinavian Journal of Economics*, 122, 853–878.

Porter, K. E. (2018): "Statistical Power in Evaluations That Investigate Effects on Multiple Outcomes: A Guide for Researchers," *Journal of Research on Educational Effectiveness*, 11, 267–295.

Seidenfeld, D., S. Handa, T. de Hoop, and M. Morey (2023): "Intraclass Correlations Values in International Development: Evidence Across Commonly Studied Domains in sub-Saharan Africa," *Evaluation Review*, 47, 786–819, pMID: 36729038.

Shen, Z. and B. Kelcey (2020): "Optimal Sample Allocation Under Unequal Costs in Cluster-Randomized Trials," *Journal of Educational and Behavioral Statistics*, 45, 446–474.

——— (2025): *odr: Optimal Design and Statistical Power for Experimental Studies Investigating Main, Mediation, and Moderation Effects*, r package version 1.5.0.

Teerenstra, S., S. Eldridge, M. Graff, E. de Hoop, and G. F. Borm (2012): "A simple sample size formula for analysis of covariance in cluster randomized trials," *Statistics in Medicine*, 31, 2169–2178.

Wacholder, S., S. Chanock, M. Garcia-Closas, L. E. Ghormli, and N. Rothman (2004): "Assessing the Probability That a Positive Report is False: An Approach for Molecular Epidemiology Studies," *JNCI: Journal of the National Cancer Institute*, 96, 434–442.

Xiang, Y., S. Gubian, B. Suomela, and J. Hoeng (2013): "Generalized Simulated Annealing for Global Optimization: The GenSA Package," *The R Journal*, 5, 13–28.

Appendix

A Additional Results

A.1 Simulation Results

A.1.1 Power Simulation Results

Table A1 compares the analytical power derived as the solution to (6) —reported in Panel B of Tables 1 to 3— with the simulated power obtained by running 10,000 replications of the data-generating process specified in equation (1), using the same sample allocations. For ease of reference, Table A1 reproduces the sample allocations and in Column 5 the analytical power results from the earlier tables, while Column 6 reports the corresponding simulated power estimates.

The simulated power, and the analytical power we calculate from our integer optimization approach, are very similar in all the cases we consider. This comparison provides reassurance about the validity of the methods that we have developed.

Table A1: Power Simulations

	(1)	(2)	(3)	(4)	(5)	(6)
Scenario	k_0	k_1	m_0	m_1	Power	Simulated Power
A.) Heterogenous Fixed Costs per Cluster (Table 1)						
1	199	84	7	18	0.915	0.912
2	170	53	7	23	0.800	0.800
3	144	34	7	32	0.651	0.652
B.) Heterogenous Variable Costs per Cluster (Table 2)						
1	94	98	7	3	0.908	0.908
2	88	95	6	2	0.799	0.809
3	79	73	6	2	0.707	0.703
C.) Heterogenous Fixed and Variable Costs per Cluster (Table 3)						
1	308	19	4	12	0.809	0.815
2	190	19	6	12	0.800	0.805
3	119	18	9	13	0.785	0.794
4	89	18	13	12	0.763	0.777

Notes: Columns 1-5 replicate the key values from panel B of Tables 1, 2, and 3. These are provided as reference for the simulation results. For the simulation results we simulate data to match the DGP presented in Equation (1). For every scenario, we run 10,000 simulations and report the mean power achieved from these runs.

A.1.2 Test Size Simulation Results

Table A2 presents the results of a second simulation exercise. Here, we set the standardized effect size, δ, to zero, and report the percentage of simulations (out of the $10,000$) in which the

null hypothesis that $\delta = 0$ is rejected (i.e. the size of the test). In all cases, the target test size, α, is 0.05. The scenarios where we see the greatest test size gap is when k_1 is small, and the $k_1 : k_0$ ratio is small too.

Table A2: Test Size Simulations

	(1)	(2)	(3)	(4)	(5)
Scenario	k_0	k_1	$k_1 : k_0$ Ratio	Target Size	Simulated Size
A.) Heterogenous Fixed Costs per Cluster (Table 1)					
1	199	84	0.422	0.050	0.052
2	170	53	0.312	0.050	0.052
3	144	34	0.236	0.050	0.058
B.) Heterogenous Variable Costs per Cluster (Table 2)					
1	94	98	1.043	0.050	0.052
2	88	95	1.080	0.050	0.052
3	79	73	0.924	0.050	0.051
C.) Heterogenous Fixed and Variable Costs per Cluster (Table 3)					
1	308	19	0.062	0.050	0.071
2	190	19	0.100	0.050	0.064
3	119	18	0.151	0.050	0.065
4	89	18	0.202	0.050	0.064

Notes: Columns 1-3 replicate the key values from panel B of Tables 1, 2, and 3. These are provided as reference for the simulation results. For the simulation results we simulate data to match the DGP presented in Equation (1), specifying a treatment effect, δ, equal to zero. For every scenario, we run 10,000 simulations and report the mean test size achieved from these runs.

A.2 Justification for Parameter Values Used in the Empirical Examples

A.2.1 Heterogeneous Fixed Costs per Cluster

For our example on heterogeneous fixed costs per cluster, based on the school grant programme based by Luo et al. (2019), we used their budget data to estimate the fixed cost per control school (f_0) to be $ 189. The fixed cost of a treatment school (f_1) includes the same transportation cost of $ 189 plus a school grant of $ 1,587 giving a total of $ 1,776.[23] The cost per each sampled student includes the interviewing costs (field team cost of administering the questionnaires, questionnaire printing costs, as well as costs of measuring students' blood hemoglobin concentration through finger-prick blood samples.) Using their budget data, we estimate the cost per sampled child, v, to be $ 9.36.

[23]The school grant was computed as 48 RMB per student in the school, and the average school has 210 students. Exchange rate $ 1 = 6.3 RMB.

A.2.2 Heterogeneous Unit Costs per Cluster

For our example on unconditional cash transfers in which unit costs are higher in treatment than control clusters, we use the average transfer amount ($ 709) and transfer fee of ($ 45) as per Haushofer and Shapiro (2016). We do not have data on the cost of interviewing households in this setting, but we will assume it is $ 100. Hence, the cost of a control household is $ 100, and the cost of a treatment household is $ 854 (=709+45+100).[24] The fixed costs per cluster are the same in treatment and control and equal to $ 250 (our own assumption on the transportation cost per cluster).

A.2.3 Heterogeneous Unit and Fixed Costs per Cluster

For our example on both heterogeneous unit and fixed costs, we use the costs of the graduation programme by Banerjee et al. (2015). We assume that the value of the transfer per household is $ 800. Banerjee et al. (2015) reports that the supervision costs associated with this type of programmes are very important. A share of these supervision costs will be fixed at the cluster level: office rental costs, IT equipment, etc. As we do not have information on what share of the total supervision costs is fixed and what is per unit, we make the assumption that half of cluster supervision costs are fixed ($ 17477), and half are per unit ($1250 per household). We also make the assumption that recruitment and interviewing costs are $100 per household, which are the same in treatment and control, and that the transportation cost of each interviewing team to a cluster amounts to $250. Hence, our assumptions are that $v_0 = 100$, $v_1 = 100+800+1250 = 2150$, $f_0 = 250$, $f_1 = 250 + 17,477 = 17,727 \approx 18,000$.

[24] We ignore here that some households in treatment clusters might be sampled but not given the cash transfer to estimate the size of the spillovers associated to the cash transfer.

B Optimal Sample Size Determination - Minimizing Costs

In this section, we repeat large swathes of Section 3, but here focus on minimizing total costs subject to achieving a given level of power. This approach may be of interest to the researcher wishing to place a competitive budget for a grant application/evaluation tender. In order to best present this approach, we provide all the necessary detail that one would need in order to work through the material independently of what is in the body of the text, hence the elements of repetition.

The power, κ, of the two-tailed test at α significance for the null hypothesis that $H_0 : \delta = 0$ when using the post estimator, δ is given by:

$$\kappa = T_{K-2}(\frac{\delta}{\sqrt{var(\hat{\delta})}} - t_{\frac{\alpha}{2},K-2}) \tag{14}$$

where T_{K-2} is the cumulative distribution function of the t -distribution with $K-2$ degrees of freedom (DoF), and the variance of $\hat{\delta}_A$ is given by:

$$var(\hat{\delta}) = \sigma^2 \left[\frac{1+(m_0-1)\rho}{m_0 k_0} + \frac{1+(m_1-1)\rho}{m_1 k_1} \right], \tag{15}$$

A researcher will want to optimize the design of the cluster RCT by determining the sample that minimizes the cost conditional on achieving a pre-specified level of power. We assume that the costs of the RCT are given by:

$$C = (f_0 + v_0 m_0)k_0 + (f_1 + v_1 m_1)k_1, \tag{16}$$

where k_0 and k_1 are the respective numbers of control and treatment clusters, f_0 and f_1 represent the fixed costs per control and treatment cluster respectively, m_0 and m_1 are the number of sample units per control and treatment cluster, and v_0 and v_1 represent the costs per control and treatment units respectively (unit costs).

The researcher who wants to minimize costs subject to attaining a level of statistical power,

κ, will want to solve:

$$\begin{aligned} \min_{\{m_0,m_1,k_0,k_1\}} \quad & [(f_0 + v_0 m_0)k_0 + (f_1 + v_1 m_1)k_1] \\ \text{s.t.} \quad & \\ \kappa \quad & \leq T_{K-2}\left(\frac{\delta}{\sqrt{var(\hat{\delta})}} - t_{\frac{\alpha}{2},K-2}\right) \end{aligned} \tag{17}$$

For mathematical convenience, it is useful to rewrite the constraint solving for δ^2, and hence the optimization problem will be:

$$\begin{aligned} \min_{\{m_0,m_1,k_0,k_1\}} \quad & [(f_0 + v_0 m_0)k_0 + (f_1 + v_1 m_1)k_1] \\ \text{s.t.} \quad & \\ \delta^2 \quad & \geq (t_{\alpha/2,K-2} + t_{\kappa,K-2})^2 var(\hat{\delta}) \end{aligned} \tag{18}$$

In its general form, the constrained optimization problem above does not have closed-form solutions. However, it can be solved numerically using robust numerical optimization methods described in Section 3.2.[25]

B.1 Pure Cases with Limited Flexibility

Given that the general problem above does not have a closed-form solution, we present below the solutions to simpler cases to gain intuition about the solution to the general problem. In these simpler cases, we assume the effect of the number of clusters in the degrees of freedom of the t-distribution to be negligible and we allow the solutions to take non-integer values, which simplifies the optimization by ensuring that the constraint holds with equality.

B.1.1 Heterogeneous Fixed Costs per Cluster

It is possible to obtain closed form solutions to the optimization problem in (18) under two additional assumptions: (1) the unit costs are homogeneous $v_0 = v_1 = v$, and (2) the number of units to sample within the clusters are equal in treatment and control clusters and exogenously given ($m_0 = m_1 = m$). In this more restricted scenario, we can rewrite the cost function as $C = (f_0 + vm)k_0 + (f_1 + vm)k_1 = (f_0 + vm)k_0 + (f_1 + vm)k_1$, giving the optimization problem

[25] We provide a Stata ado file Clustercostmin and a R package Optimal.sample to perform this optimization and obtain the optimal sample allocation. See footnote 8 for details.

as:

$$\min_{\{k_0,k_1\}} \quad (f_0 + vm)k_0 + (f_1 + vm)k_1 \tag{19}$$

s.t.

$$\delta^2 \quad = (t_{\alpha/2} + t_\kappa)^2\sigma^2(1 + (m-1)\rho)\frac{1}{m}\left(\frac{1}{k_0} + \frac{1}{k_1}\right) \tag{20}$$

where the only unknowns are k_0 and k_1 because the number of units to be sampled per each cluster is exogenously given by m. Note that the constraint is the same as the constraint in (18) but where the conditions ($m_0 = m_1 = m$) and ($v_0 = v_1 = v$) has been substituted in the formulae for $var(\hat{\delta})$ in (15), and the constraint holds with equality.

The solution to the optimization problem yields the following optimality condition,

$$\frac{k_1}{k_0} = \sqrt{\frac{(f_0 + vm)}{(f_1 + vm)}} \tag{21}$$

Using the squared MDE formula (20), we can write the optimal values of k_0 and k_1 as functions of the model parameters:

$$k_0^* = (t_{\alpha/2} + t_\kappa)^2\sigma^2(1 + (m-1)\rho)\frac{1}{m}\frac{1}{\delta^2}\left(\frac{\sqrt{(f_0 + vm)} + \sqrt{(f_1 + vm)}}{\sqrt{(f_0 + vm)}}\right), \quad \text{and} \tag{22}$$

$$k_1^* = (t_{\alpha/2} + t_\kappa)^2\sigma^2(1 + (m-1)\rho)\frac{1}{m}\frac{1}{\delta^2}\left(\frac{\sqrt{(f_0 + vm)} + \sqrt{(f_1 + vm)}}{\sqrt{(f_1 + vm)}}\right). \tag{23}$$

We can now present an expression for the minimum total cost, C^*, required in order to achieve a power of κ with a given value of δ, by substituting the relations in equations (22) and (23) into the cost function $C = (f_0 + vm)k_0 + (f_1 + vm)k_1$:

$$C^* = (t_{\alpha/2} + t_\kappa)^2\sigma^2(1 + (m-1)\rho)\frac{1}{m}\frac{1}{\delta^2}\left(\sqrt{(f_0 + vm)} + \sqrt{(f_1 + vm)}\right)^2 \tag{24}$$

Note that the above closed form solutions were obtained using the assumption that the number of units to be sampled within each cluster, m, was exogenously given. In practice, it is straightforward to circumvent this assumption by doing a grid search on m, that is, the optimal values of k_0 and k_1 can be computed for different values of m, and choose the one that minimizes the costs. Hence, the actual important assumption for this special case to be useful is that $m_0 = m_1$.

B.2 Heterogeneous Unit Costs per Cluster

It is also possible to obtain closed form solutions to the optimization problem in (18) under another alternative scenario: (1) fixed costs per cluster are equal in treatment and control $f_0 = f_1 = f$, (2) the number of clusters are equal across treatment and control arms and exogenously given ($k_0 = k_1 = k$). In this more restricted scenario, we can rewrite the cost function as $C = (f + v_0 m_0)k_0 + (f + v_1 m_1)k_1 = 2fk + v_0 m_0 k + v_1 m_1 k$, giving the optimization problem as:

In this case, we write the constrained optimization problem as:

$$\min_{\{m_0, m_1\}} \quad 2fk + v_0 m_0 k + v_1 m_1 k \tag{25}$$

s.t.

$$\delta^2 \quad = (t_{\alpha/2} + t_\kappa)^2 \sigma^2 \frac{1}{k} \left(\frac{1 + (m_0 - 1)\rho}{m_0} + \frac{1 + (m_1 - 1)\rho}{m_1} \right). \tag{26}$$

The solution to the optimization problem yields the following optimality condition,

$$\frac{m_1}{m_0} = \sqrt{\frac{v_0}{v_1}}. \tag{27}$$

Using the squared MDE formula (26), we can write the optimal values of k_0 and k_1 as functions of the model parameters:

$$m_0^* = \frac{(t_{\alpha/2} + t_\kappa)^2 \sigma^2 \left(\frac{1-\rho}{k}\right)}{\delta^2 - (t_{\alpha/2} + t_\kappa)^2 \sigma^2 \left(\frac{2\rho}{k}\right)} \left(\frac{\sqrt{v_0} + \sqrt{v_1}}{\sqrt{v_0}} \right), \quad \text{and} \tag{28}$$

$$m_1^* = \frac{(t_{\alpha/2} + t_\kappa)^2 \sigma^2 \left(\frac{1-\rho}{k}\right)}{\delta^2 - (t_{\alpha/2} + t_\kappa)^2 \sigma^2 \left(\frac{2\rho}{k}\right)} \left(\frac{\sqrt{v_0} + \sqrt{v_1}}{\sqrt{v_1}} \right). \tag{29}$$

Finally, we can write down an expression for the minimum total cost, C^*, required in order to achieve a power of κ with a given value of δ, by substituting the relations in equations (28) and (29) into the cost function $C = 2fk + v_0 m_0 k + v_1 m_1 k$:

$$C^* = 2fk + \frac{(t_{\alpha/2} + t_\kappa)^2 \sigma^2 (1 - \rho)(\sqrt{v_0} + \sqrt{v_1})^2}{\delta^2 - (t_{\alpha/2} + t_\kappa)^2 \sigma^2 (\frac{2\rho}{k})} \tag{30}$$

Note that the above closed form solutions were obtained using the assumption that the number of clusters, k, was exogenously given. In practice, it is straightforward to circumvent this assumption by doing a grid search on k, that is, the optimal values of k_0 and k_1 can be

computed for different values of k, and choose the one that minimizes the costs. Hence, the actual important assumption for this special case to be useful is that $k_0 = k_1$.

B.3 Results

Here we report the sample size estimates for the same three case studies as we do in Section 4. The reported sample size estimates are for a double-sided test of means at 5% significance, and for a power of 80%. We set an effect size δ of 0.25, standard deviation σ of 1, and a intracluster correlation of 0.27 for the first example (school grant) but a smaller one, 0.05, for the other two.

B.4 Optimal Sample Size Allocations and Cost Savings

Having discussed the three case studies in detail in the main body of the text, in this section we focus predominantly on the cost savings from our approach (found in Panel B of Tables B1-B3) compared to the non-optimized, balanced allocation approach in Panel A.[26]

Table B1: Heterogenous Fixed Costs per Cluster – School Grant Program

	(1)	(2)	(3)
Fixed Cost Control (f_0)	189	189	189
Fixed Cost Treatment (f_1)	1,000	1,776	3,000
Unit Cost (v)	9.36	9.36	9.36
Target Power	0.80	0.80	0.80
A.) Balanced Allocation			
$k_0 = k_1 = k$	83	79	78
$m_0 = m_1 = m$	13	18	20
Total Cost ($)	118,886	181,886	277,945
B.) Optimal Allocation			
k_0	145	157	198
k_1	59	54	49
m_0	7	8	7
m_1	17	23	30
Total Cost ($)	105,293	148,980	211,154
Savings vs Approach A ($)	13,593	32,906	66,791
Savings vs Approach A (%)	12.9%	22.1%	31.6%

Notes: The optimization algorithm used in both panel A and panel B is based on a grid-search, integer-optimization approach. This yields a directly implementable solution without the need for rounding. The values for number of individuals per cluster (m) and number of clusters (k) are those that achieve 80% power at 5% significance for the cost parameters specified in the top 3 rows. Other assumed parameters: effect size 0.25, standard deviation 1, intra-cluster correlation (ρ) = 0.27.

In Tables B1, B2 and B3, our approach is associated with costs savings of 22%, 21% and 53% respectively for the benchmark cases described in columns (2) of the Tables. These percentage

[26]Note that the optimal sample allocations in column (2) of Tables 1-3 (solution to the primal problem) are not exactly the same as the optimal sample allocations in column (2) of Tables B1 - B3 (dual problem) because we are constraining the optimal solutions to be integers.

Table B2: Heterogenous Variable Costs per Cluster – Unconditional Cash Transfer

	(1)	(2)	(3)
Fixed Cost (f)	250	250	250
Unit Cost Control (v_0)	100	100	100
Unit Cost Treatment (v_1)	500	854	1,200
Target Power	0.80	0.80	0.80
A.) Balanced Allocation			
$k_0 = k_1 = k$	74	94	133
$m_0 = m_1 = m$	4	3	2
Total Cost ($)	214,600	316,028	412,300
B.) Optimal Allocation			
k_0	63	91	91
k_1	71	94	90
m_0	8	6	7
m_1	3	2	2
Total Cost ($)	190,400	261,402	324,950
Savings vs Approach A ($)	24,200	54,626	87,350
Savings vs Approach A (%)	12.7%	20.9%	26.9%

Notes: The optimization algorithm used in both panel A and panel B is based on a grid-search, integer-optimization approach. This yields a directly implementable solution without the need for rounding. The values for number of individuals per cluster (m) and number of clusters (k) are those that achieve 80% power at 5% significance for the cost parameters specified in the top 3 rows. Other assumed parameters: effect size 0.25, standard deviation 1, intra-cluster correlation (ρ) = 0.05.

savings amount to large savings in absolute terms, particularly for graduation-style programs – in Table 3 the absolute saving using our approach exceeds half a million US Dollars for our baseline case. In any case, the concern about possible over rejection under the null because the number of treatment clusters might be too small which we discussed in Section 5 also applies here.

Table B3: Heterogenous Fixed and Variable Costs per Cluster – Graduation Program

	(1)	(2)	(3)	(4)
Fixed Cost Control (f_0)	125	250	500	1,000
Fixed Cost Treatment (f_1)	18,000	18,000	18,000	18,000
Unit Cost Control (v_0)	100	100	100	100
Unit Cost Treatment (v_1)	2,150	2,150	2,150	2,150
Target Power	0.80	0.80	0.80	0.80
A.) Balanced Allocation				
$k_0 = k_1 = k$	32	32	32	32
$m_0 = m_1 = m$	13	13	13	13
Total Cost ($)	1,516,000	1,520,000	1,528,000	1,544,000
B.) Optimal Allocation				
k_0	269	192	114	96
k_1	18	19	19	19
m_0	4	6	9	12
m_1	13	12	13	13
Total Cost ($)	968,325	995,400	1,032,650	1,084,250
Savings vs Approach A ($)	547,675	524,600	495,350	459,750
Savings vs Approach A (%)	56.6%	52.7%	48.0%	42.4%

Notes: The optimization algorithm used in both panel A and panel B is based on a grid-search, integer-optimization approach. This yields a directly implementable solution without the need for rounding. The values for number of individuals per cluster (m) and number of clusters (k) are those that achieve 80% power at 5% significance for the cost parameters specified in the top 3 rows. Other assumed parameters: effect size 0.25, standard deviation 1, intra-cluster correlation (ρ) = 0.05.

C Different Formulations of the Variance of the Treatment Effect

This section presents the equivalence between the variance of the treatment effect in equation (4) above, and equation $A1$ in the Appendix of Shen and Kelcey (2020), which we specify with different sample sizes across treatment conditions at all levels. Table C1 below provides a correspondence between how we define the key parameters, and how Shen and Kelcey (2020) do so.

Following Shen and Kelcey (2020), the variance of the treatment effect is:

$$\sigma_\delta^2 = \frac{\rho(1-R_2^2) + \frac{(1-\rho)(1-R_1^2)}{[\frac{nn^T}{(1-p)n+pn^T}]}}{p(1-p)} . \frac{(1-p)(c_1 n + c_2) + p(c_1^T n^T + c_2^T)}{m} , \tag{31}$$

where the budget function is $m = (1-p)J(c_1 n + c_2) + pJ(c_1^T n^T + c_2^T)$. Solving for J in the budget function, we can write

$$J = \frac{m}{(1-p)(c_1 n + c_2) + p(c_1^T n^T + c_2^T)} . \tag{32}$$

Substituting equation (32) into equation (31), we can rewrite the variance of the treatment effect as:

$$\sigma_\delta^2 = \frac{\rho(1-R_2^2) + \frac{(1-\rho)(1-R_1^2)}{[\frac{nn^T}{(1-p)n+pn^T}]}}{p(1-p)J} . \tag{33}$$

Under the assumption that $R_1^2 = 0$ and $R_2^2 = 0$ the variance of the treatment effect is given by:

$$\sigma_\delta^2 = \frac{\rho + \frac{(1-\rho)}{[\frac{nn^T}{(1-p)n+pn^T}]}}{p(1-p)J} ,$$

which, after some algebra, can be rewritten as:

$$\sigma_\delta^2 = \frac{\rho nn^T + (1-\rho)[(1-p)n + pn^T]}{p(1-p)nn^T J} . \tag{34}$$

The first step is to show that equation (34) is equivalent to equation (4), using the equivalence between parameters in Shen and Kelcey (2020) and our work, which we summarize in Table C1. Subsituting parameters accordingly, we can rewrite equation (34) as a function of our parameters:

$$\sigma_\delta^2 = \frac{\rho m_0 m_1 + (1-\rho)[(1-p)m_0) + pm_1]}{p(1-p)m_0 m_1 J} . \tag{35}$$

From Table C1, we have that $p = k_1/J$ and $(1-p) = k_0/J$. If we substitute these expressions into equation (35) we have:

$$\sigma^2_\delta = \frac{\rho m_0 m_1 + (1-\rho)[\frac{k_0}{J} m_0 + \frac{k_1}{J} m_1]}{\frac{k_1}{J}\frac{k_0}{J} m_0 m_1 J},$$

which after some algebra we can rewritten as:

$$\sigma^2_\delta = \frac{\rho m_0 m_1 J + (1-\rho)[k_0 m_0 + k_1 m_1]}{m_0 k_0 m_1 k_1}. \tag{36}$$

To completely express the variance of the treatment effect in terms of the parameters we use in this paper, we need to substitute J. From the equivalences presented in Table C1 we have $J = K = k_0 + k_1$. Substituting this last expression into equation (36), we get to the expression:

$$\sigma^2_\delta = \frac{\rho m_0 m_1 (k_0 + k_1) + (1-\rho)[k_0 m_0 + k_1 m_1]}{m_0 k_0 m_1 k_1}.$$

Rearranging some terms we have:

$$\sigma^2_\delta = \frac{m_0 k_0 (\rho m_1 + 1 - \rho) + k_1 m_1 (\rho m_0 + 1 - \rho)}{m_0 k_0 m_1 k_1}.$$

Finally, a bit more of algebra leads us to determine that the variance of the treatment effect in Shen and Kelcey (2020), expressed in terms of the paramaters we use, is:

$$\sigma^2_\delta = \frac{1 + (m_0 - 1)\rho}{m_0 k_0} + \frac{1 + (m_1 - 1)\rho}{m_1 k_1}. \tag{37}$$

Recall, equation (4) is:

$$\sigma^2_\delta = \sigma^2 \left[\frac{1 + (m_0 - 1)\rho}{m_0 k_0} + \frac{1 + (m_1 - 1)\rho}{m_1 k_1} \right]. \tag{38}$$

Under the assumption that $\sigma^2 = 1$, equations (37) and (38) are equivalent, which completes this section, as we have shown that the variance of the treatment effect in Shen and Kelcey (2020) coincides with that in this work.

Table C1: Parameter Equivalence Between our formulation and Shen and Kelcey (2020)

	(1)	(2)
Parameter	**Shen and Kelcey (2020)**	**This paper**
Sample units per control cluster	n	m_0
Sample units per treatment cluster	n^T	m_1
Total number of clusters	J	K
Number of control clusters	$(1-p)J$	k_0
Number of treatment clusters	pJ	k_1
Fixed costs per control cluster	C_2	f_0
Fixed costs per treatment cluster	C_2^T	f_1
Variable costs per control cluster	C_1	v_0
Variable costs per treatment cluster	C_1^T	v_1
Total cost	m	C

Notes: p is the proportion of clusters to be assigned to the treatment condition in Shen and Kelcey (2020).