DOI: 10.1111/1475-5890.70005

Fiscal Studies
The Journal of Applied Public Economics
IFS Institute for Fiscal Studies

SYMPOSIUM PAPER

Going beyond simple sample size calculations: a practitioner's guide

Brendon McConnell[1] | **Marcos Vera-Hernández**[2]

[1]City St George's, University of London

[2]University College London; Institute for Fiscal Studies; CEPR

Correspondence
Marcos Vera-Hernández, Department of Economics, University College London, Gower Street, London WC1E 6BT, UK.
Email: m.vera@ucl.ac.uk

Submitted: July 2025

Funding information
Economic and Social Research Council

Abstract

Basic methods to compute required sample sizes are well understood and supported by widely available software. However, researchers often oversimplify their sample size calculations, overlooking relevant features of their experimental design. This paper compiles and systematises existing methods for sample size calculations for continuous and binary outcomes, both with and without covariates, and for both clustered and non-clustered randomised controlled trials. We present formulae accommodating panel data structures and uneven designs, and provide guidance on optimally allocating sample size between the number of clusters and the number of units per cluster. In addition, we discuss how to adjust calculations for multiple hypothesis testing and how to estimate power in more complex designs using simulation methods.

KEYWORDS
power analysis, sample size calculations, randomised control trials, cluster randomised control trials, covariates, multiple outcomes, simulation

JEL CLASSIFICATION
C8, C9

1 | INTRODUCTION

One of the big challenges in economics has been to estimate causal relationships between economic variables and policy instruments. Randomised controlled trials (RCTs) 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, Glennerster and Kremer, 2007).[1] Simpler RCTs are usually set up

[1] See Blundell and Costa Dias (2009) and Imbens and Wooldridge (2009) for reviews on non-experimental methods.

with the objective of estimating the impact of a certain policy or intervention, while more complex RCTs can be implemented to test the competing hypotheses that explain a phenomenon (also known as field experiments; see Duflo (2006) and Levitt and List (2009)).

When setting up an RCT, one of the first important tasks is to calculate the sample size that will be used for the experiment. This is to ensure that the planned sample is large enough to detect expected differences in outcomes between the treatment and control groups. A sample size that is too small leads to an underpowered study, which will have a high probability of overlooking an effect that is real. The implications of small sample sizes go beyond that: low power also means that statistically significant effects are likely to be false positives.[2] Studies with samples larger than required also have their drawbacks: they will expose a larger pool of individuals to an untested treatment, be more logistically complex and be more expensive than necessary.

Basic methods to compute the required sample size are well understood and supported by widely available software. However, the sophistication of the sample size formulae commonly used has not kept pace with the complexity of the experimental designs most often used in practice. RCTs are usually analysed using data collected before the intervention started (baseline data) but this is often ignored by the sample size formulae commonly used by researchers, as is the inclusion of covariates in the analysis. Another departure from the basic design is that interventions are commonly assessed not just on a single outcome variable but on more than one, creating problems of multiple hypotheses testing that should be taken into account when computing the required sample size. Depending on the context and specific assumptions, taking into consideration some of these departures from the basic design will lead to smaller or larger sample sizes.

The objective of this paper is to provide researchers with a practitioner's guide – supported by accompanying software – that enables them to incorporate into their sample size calculations features commonly present in RCTs but often overlooked in practice. Although most of the content is not novel, most of it is dispersedly published in quite diverse notation, making it difficult for the applied researcher to find the right formulae just at the busy time when they are writing the research proposal that will fund the RCT. We also note that understanding the sample size implications of different design features can be very useful when designing the RCT (what waves of data to collect, what information to collect, etc.)

This article will include sample size calculations for both continuous and binary outcomes, starting with the simplest case of individual-level trials, and then cluster randomised trials. We will also cover how to take into account pre-intervention data, as well as covariates. Throughout the paper, we favour simplicity in exposition and attempt to keep the language accessible to the applied researcher who does not have previous exposure to sample size calculations.

The article has three extensions. The first extension discusses how to choose optimally the number of clusters versus the number of units within clusters. The second extension explains how to compute the power using simulation methods, which is useful when there are no existing formulae for the RCT that is being planned. The third extension shows how to adapt the sample size computations when several outcomes are used.

An inherent difficulty in using the sample size formulae that we provide in the paper is that assumptions are needed on some key parameters of the data-generating process, which are not required by the basic formulae. Our view is that the widespread trend towards making data publicly available, including the data used in academic publications, will definitively help researchers to find realistic values for the parameters of interest. Moreover, social science journals might follow the trend set by medical journals on making it compulsory for authors to report certain key estimates that are commonly used in sample size calculations (Schulz, Altman and Moher, 2010).

The paper is organised as follows. Section 2 presents the example of the intervention that will be used throughout the paper, Section 3 provides an overview of basic concepts involved in power

[2] An intuitive explanation using a numeric example can be found in a briefing in *The Economist* (18 October 2013), 'Trouble at the lab', https://www.economist.com/briefing/2013/10/18/trouble-at-the-lab.

calculations, Section 4 considers power calculations for continuous outcomes, Section 5 focuses on discrete outcomes, Section 6 discusses the three extensions and Section 7 concludes. In the online appendices, we provide examples of Stata code to estimate key parameters needed to perform sample size calculations, and code to compute power through simulation. Spreadsheets and Stata do-files to implement the methods discussed in this article can be obtained from https://ifs.org.uk/publications/sample-size-calculators-going-beyond-simple-sample-size-calculations-practitioners.

2 | OVERVIEW OF AN EXAMPLE INTERVENTION

In this section, we will set up an example that we will use for the rest of the paper. Let us assume that we would like to evaluate APRENDE, a fictional job-training programme that will be implemented by the government of EvaluaLand. Such government will run a randomised controlled trial (RCT) to evaluate APRENDE. Our task is to compute the required sample size for such evaluation. The main outcomes of interest are individual earnings, and the proportion of individuals who work at least 16 hours a week.

As will be clear later on, to be able to compute the sample size requirements, we will need some basic parameters, such as average earnings, the standard deviation of earnings, and the proportion of individuals who work at least 16 hours a week. We are at the planning stage, so we have not collected the data yet, and hence we do not know the value of these parameters for our target population. We may use previous studies that report these parameters in our context or in a similar context.

In this case, we have benefited from the availability of a recent labour market survey – the EvaluaLand National Survey of Earnings – which contains the key variables required for the sample size calculations for APRENDE.[3] Specifically, the dataset reports individual earnings, town of residence and a covariate that may be used in the analysis. Importantly, the survey is representative of the target population of APRENDE.

The evaluation of APRENDE may be implemented using either individual-level or cluster-level randomisation. Under individual-level randomisation, a small number of pilot towns would first be selected. Within each town, a list of eligible individuals interested in participating in APRENDE would be compiled, and a lottery would be conducted to determine which individuals are selected to participate in the programme during the pilot phase and which are randomised out. Alternatively, a cluster RCT design could be employed, whereby towns participating in the evaluation are randomly assigned to either the treatment or control group. Eligible individuals residing in treatment towns would then be invited to apply for and participate in APRENDE. In this case, the town constitutes the cluster, as it is the unit of randomisation, even though the data for the evaluation would be collected at a more granular level (i.e. the individual). Common examples of clusters in other contexts include schools, job centres and primary care clinics.

One of the main parameters needed to compute the sample size requirements is the effect size, which is the smallest effect of the policy that we want to have enough power to detect. When considering the effect size for an individual-based RCT, we must take into account that it refers to the comparison in the outcome levels of individuals initially allocated to treatment versus control. Note that this difference will be diluted by any non-compliance (i.e. individuals initially allocated to treatment that eventually decide not to participate), and hence we must adjust the effect size accordingly. For instance, if we think that APRENDE will increase participants' average earnings by 14,000 but 30 per cent of individuals initially allocated to participate in APRENDE decide not to take it up, we must plan for a diluted effect size of 9,800 ($= 14{,}000 \times 0.7$) as this will take into account the non-compliance rate. McKenzie (2025), published in this issue, discusses practical strategies to reduce non-compliance.

In a cluster RCT, the relevant comparison is the difference in outcome levels between the eligible individuals living in treatment towns (irrespective of whether they participated or not) and the eligible

[3] This dataset is included in the Supporting Information to enable readers to implement the code provided in Appendix B.

individuals living in control ones (also irrespective of whether they participated or not). Because not all eligible individuals living in treatment towns will end up participating, the coverage rate of the policy must be taken into account when considering the effect size. Assuming that APRENDE increases participants' average earnings by 14,000, we should plan for an effect size of 8,400 (= 14,000 × 0.6) if the coverage rate is expected to be 60 per cent (it is expected that 40 per cent of the eligible population living in the treatment towns will not participate in APRENDE, either because of capacity constraints or because they are not interested).[4] Of course, the effect size would have to be even smaller if we think that individuals in control towns can travel to treatment towns and participate in APRENDE (contamination). For instance, if 10 per cent of individuals living in control towns could do that, then the planned effect size would have to be 7,000 (= 14,000 × (0.6–0.1)).

3 | BASIC CONCEPTS

One of the most important questions when computing the required sample size for the evaluation of APRENDE is 'What is the smallest effect of the programme on earnings that we want the study to be able to detect?'. The answer to this question defines the effect size – often referred to in the literature as the minimum detectable effect (MDE) – and is denoted by δ.

For those unfamiliar with sample size calculations, this may be a slightly strange concept, as in order to calculate the sample size for a trial, we need to input the impact we expect the trial to have. It is common to refer to existing literature in order to get a sense of this effect size. Of course, the results from previous literature must be contextualised to the study that is being planned. For instance, the researcher might think that APRENDE should be less effective than existing studies, maybe because it targets all ages rather than the youth. Differences in expected non-compliance and contamination between APRENDE and other existing studies will also modify the effect size that we will plan for. Nothing precludes the researcher from conducting sample size calculations with several different values of the effect size to gauge the sensitivity of the results.

Assessing whether an intervention has a genuine effect on the outcome variable is challenging because, in practice, we seldom observe outcomes for the entire population of individuals or clusters assigned to treatment and control. Instead, researchers typically rely on data from a random sample of each group. Even if the intervention has no effect at the population level, the sample average of the outcome in the treatment group will usually differ from that in the control group. This is due to sampling variability – the natural variation in estimates that arises because each sample captures only a subset of the population, and the specific individuals or clusters included in the subset will influence the sample mean. The core inferential task is to determine whether the observed difference in sample means is sufficiently large to suggest a true difference in population means – attributable to the intervention – or whether it is small enough to plausibly reflect random variation from the sampling process alone. This is where hypothesis testing becomes essential. The null hypothesis (H_0) typically states that the population mean of the outcome is equal across treatment and control groups – implying that the intervention was on average ineffective. The alternative hypothesis states that the effect of the intervention is δ (the difference in the population mean of the outcome variable between treatment and control, which we call the effect size).

When conducting the hypothesis test, two possible errors are likely to happen. On the basis of the sample at hand, and the test carried out, the researcher could reject a true null hypothesis, i.e. conclude that the intervention was effective when it was not. This type of 'false positive' error is usually called a Type I error (see Figure 1). The other possible error is to conclude that the intervention had no effect

[4] Conceptually, coverage and compliance are distinct. The coverage rate refers to the proportion of individuals who choose to enrol in the treatment when it is offered to them, whereas the compliance rate refers to the proportion of those enrolled who go on to fully participate in the treatment – that is, who adhere to or complete the intended intervention.

	H_0 is true	H_1 is true
Fail to reject null hypothesis	Correct	Type II error
Reject null hypothesis	Type I error	Correct

FIGURE 1 Type I and Type II errors in hypothesis testing

when one exists (fail to reject the null hypothesis if it is false). This type of 'false negative' error is called a Type II error.

The researcher will never be able to know whether a Type I or Type II error is being committed, because the truth is never fully revealed. But the researcher can design the study so as to control the probability of committing each type of error. Significance, usually denoted by α, is the probability of committing a Type I error (Prob[reject $H_0|H_0$ true] = α). Commonly, α is set to equal 0.05.[5] This means that when the null is true, we will only reject it in 5 per cent of cases. The probability of a Type II error, denoted by β, is the chance of concluding that the intervention has no effect, when one exists (Prob[fail to reject $H_0|H_1$ true] = β). Common values of β are between 0.1 and 0.2.

Power is defined as $1 - \beta$, i.e. Prob[reject $H_0|H_1$ true]. In our context, power refers to the probability of detecting an effect of a given size of APRENDE on earnings, conditional on APRENDE having such an effect. Put more bluntly, power is the probability that a study has of uncovering a true, non-zero, effect. The researcher would like power to be as high as possible; otherwise the study has a high chance of overlooking an effect that is real. Usually, power of 0.8 or 0.9 is considered high enough (consistent with values of β between 0.1 and 0.2).

In addition to specifying the effect size and the desired levels of significance and power, several other key parameters are required to compute the sample size. For binary outcome variables, it is necessary to provide an estimate of the proportion of individuals in the control group who exhibit the outcome of interest (e.g. who are employed, enrolled in school or vaccinated). For continuous outcomes, one must specify the variance of the outcome, denoted σ^2.[6] These values can typically be obtained from existing household surveys (e.g. the EvaluaLand National Survey of Earnings), from previous studies or from a pilot study if one has been conducted.

There is an additional input required when calculating power for cluster RCTs. This is the intra-cluster correlation (ICC), which is a measure of how correlated the outcomes are within clusters. This parameter, denoted here as ρ and defined below, can be estimated from a pilot survey or based on measures found in the existing literature. This parameter plays an important role in sample size calculations for cluster randomised trials, and can lead to one requiring much larger sample sizes than in the individual-level randomisation case.[7] The reason for this is that the larger is the correlation of outcomes amongst individuals within clusters, the less informative an extra individual sampled within the cluster is. Adding an extra cluster of k individuals will result in greater power than including k more individuals across existing clusters.

[5] Later in the paper, we will discuss testing for multiple outcomes, which will affect the value chosen for α.

[6] In the binary case, there is no need to specify the variance separately, as it is fully determined by the mean: the variance of a binary variable equals $p(1-p)$, where p is the mean of the variable.

[7] Where covariates are included, it is the conditional ICC that will be used in the calculations below. This may be harder to obtain from previous studies.

4 | CONTINUOUS OUTCOMES

Here we derive the sample size calculation for the simple case of an RCT in which the treatment, T, is randomised at the individual level and the outcome variable, Y, is continuous. This simple case allows us to focus on the main steps that are necessary to derive the sample size formulae, and it is useful to give a sense of how the other formulae used in this paper are derived.[8] Usually, we test whether T had an effect on Y by testing whether the population means of Y are different in the treatment and control groups. More formally, if we denote the population means in the treatment and control groups by μ_1 and μ_0 respectively, the null hypothesis is H_0: $\mu_1 - \mu_0 = 0$; and the alternative hypothesis is that the difference in the population means equals the MDE, $H_1: \mu_1 - \mu_0 = \delta$.

Assume that we have a sample of n_0 individuals in the control group and a sample of n_1 individuals in the treatment group. We denote by $T_i = 0$ that individual i is part of the control group and by $T_i = 1$ that individual i is part of the treatment group. To test H_0 against H_1, we would estimate the following ordinary least squares (OLS) regression:[9]

$$Y_i = \gamma_0 + \gamma_1 T_i + \epsilon_i,$$

where Y_i is the value of the outcome variable (say earnings in the case of APRENDE) and ϵ_i is an error term with zero mean and variance σ^2, which for the time being we assume is known. The z-statistic associated with γ_1 is given by the OLS estimate of γ_1 divided by its standard error:

$$Z = \frac{\overline{Y_1} - \overline{Y_0}}{\sigma\sqrt{(1/n_0) + (1/n_1)}},$$

where $\overline{Y_1}$ and $\overline{Y_0}$ are the sample averages of Y_i for individuals in the treatment and control group respectively, n_1 and n_0 are the sample size in the treatment and control group respectively, and σ is the standard deviation of Y_i. If the null hypothesis is true, then $\mu_1 = \mu_0$, and Z follows a normal distribution with mean 0 and variance 1. Hence, the null hypothesis will be rejected at a significance level of α if $Z \geq z_{\alpha/2}$ or $Z \leq -z_{\alpha/2}$, where the cumulative distribution function of the standard normal distribution evaluated at $z_{\alpha/2}$ is $1 - \alpha/2$.

As mentioned above, power (denoted by $1 - \beta$) is the probability of rejecting the null hypothesis when the alternative is correct, i.e.

$$1 - \beta = \text{Prob}(Z \leq -z_{\alpha/2} \cup Z \geq z_{\alpha/2} | H_1) = \text{Prob}(Z \leq -z_{\alpha/2} | H_1) + \text{Prob}(Z \geq z_{\alpha/2} | H_1).$$

Because the alternative hypothesis is correct, $\mu_1 - \mu_0$ is no longer zero but δ. Hence, the mean of Z is no longer zero but $\delta / \left(\sigma\sqrt{(1/n_0) + (1/n_1)}\right)$. In this case, $\text{Prob}(Z \leq -z_{\alpha/2} | H_1)$ is approximately zero, and hence we have that[10]

$$1 - \beta = \text{Prob}(Z \geq z_{\alpha/2} | H_1) = 1 - \text{Prob}(Z < z_{\alpha/2} | H_1).$$

[8] The material in this section is standard for statistical textbooks. In this section, we follow Liu (2013) closely.

[9] We use a regression framework to keep the parallelism with forthcoming sections, but a t-test for two independent samples is equivalent.

[10] See, for instance, Liu (2013). Note, however, that Liu (2013) defines $z_{\alpha/2}$ such that the cumulative distribution function of the standard normal distribution evaluated at $z_{\alpha/2}$ is $\alpha/2$ instead of $1 - \alpha/2$.

By subtracting the mean of Z under the alternative hypothesis from both sides of the inequality, we obtain

$$\beta = \text{Prob}\left(Z - \frac{\delta}{\sigma\sqrt{(1/n_0) + (1/n_1)}} < z_{\alpha/2} - \frac{\delta}{\sigma\sqrt{(1/n_0) + (1/n_1)}} \right).$$

Because the left-hand side of the inequality now follows a normal distribution with zero mean and unit variance, it is the case that

$$z_{1-\beta} = z_{\alpha/2} - \frac{\delta}{\sigma\sqrt{(1/n_0) + (1/n_1)}},$$

which implies that[11]

$$z_{\beta} + z_{\alpha/2} = \frac{\delta}{\sigma\sqrt{(1/n_0) + (1/n_1)}}.$$

In the case in which σ is unknown and is estimated using the standard deviation in the sample, a t distribution with $v = n_0 + n_1 - 2$ degrees of freedom must be used instead of the normal distribution. In this case, we have that

$$t_{\beta} + t_{\alpha/2} = \frac{\delta}{\sigma\sqrt{(1/n_0) + (1/n_1)}}.$$

Solving for δ, we obtain the expression for the MDE that can be detected with $1 - \beta$ power at significance level α:

$$\delta = (t_{\beta} + t_{\alpha/2})\sigma\sqrt{\frac{1}{n_0} + \frac{1}{n_1}}. \tag{1}$$

Assuming equal sample sizes in the treatment and control groups, $n_0 = n_1 = n$, the formula simplifies, and the required sample size per arm becomes

$$n = 2(t_{\beta} + t_{\alpha/2})^2\left(\frac{\sigma}{\delta}\right)^2. \tag{2}$$

As is clear from expressions (1) and (2), the required sample size depends solely on the standardised effect size, defined as the ratio of the effect size to the standard deviation of the outcome, δ/σ. As a result, it is not necessary to specify the outcome's mean in the absence of the intervention, nor its variance in absolute terms. This has led to the widespread use of standardised effect sizes in power calculations, as it allows sample size requirements to be expressed without reference to the original scale of the outcome variable.[12]

Finally for this section, we outline the case where variances are unequal, following List, Sadoff and Wagner (2011). This case is not very common in practice, as it is difficult a priori to consider how the

[11] Note that $z_{\beta} = -z_{1-\beta}$.

[12] Cohen (1988) popularised benchmark values for standardised effect sizes, suggesting 0.2 for a small effect, 0.5 for a medium effect and 0.8 for a large effect. Standardised effect sizes can be readily used in standard software by assuming that the variance of the outcome variable is equal to 1.

treatment will affect not just the mean of the outcomes, but the variance too.[13] Under equal variances, the expression for the MDE, equivalent to (1), becomes

$$\delta = (t_\beta + t_{\alpha/2})\sqrt{\frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1}},$$

which leads to the following expressions for the optimal sample (see Appendix A for the full derivation):

$$\begin{aligned} N^* &= (t_\beta + t_{\alpha/2})^2 \frac{1}{\delta^2}\left(\frac{\sigma_0^2}{\pi_0^*} + \frac{\sigma_1^2}{\pi_1^*}\right), \\ N^* &= n_0^* + n_1^*, \\ n_0^* &= \pi_0^* N^*, \\ n_1^* &= \pi_1^* N^*, \end{aligned} \tag{3}$$

where $\pi_0^* = \sigma_0/(\sigma_0 + \sigma_1)$ and $\pi_1^* = \sigma_1/(\sigma_0 + \sigma_1)$, n_0^* (n_1^*) refer to the optimal sample in the control (treatment) group, and N^* refers to the total optimal sample size.[14] These expressions imply that a larger share of the sample should be allocated to the group with the higher variance in the outcome variable, reflecting its greater contribution to the overall sampling variability.

4.1 | Cluster randomisation

In many cases, the outcome variable is measured at the unit level (individual, household, firm, etc.) but the randomisation takes place at the cluster level (school, village, firm, etc.). This may be driven by concerns over spillovers within a cluster, whereby unit-level randomisation would lead to control members' outcomes being contaminated by those of treated individuals. In this case, the sample size formula must be adjusted to reflect that observations from units of the same cluster are not independent, as they may share some unobserved characteristics.

The estimating equation will take the form

$$Y_{ij} = \gamma_0 + \gamma_1 T_j + c_j + u_{ij}, \tag{4}$$

where i denotes units and j denotes clusters. T_j is the treatment indicator. c_j and u_{ij} are error terms at the cluster and unit level respectively. The variances of c_j and u_{ij} are given by $\text{var}(c_j) = \sigma_c^2$ and $\text{var}(u_{ij}) = \sigma_u^2$, and $\sigma_c^2 + \sigma_u^2 = \sigma^2$.

To carry out the sample size calculation in the presence of clustering, we require an additional input – the intra-cluster correlation or ICC, denoted here as ρ:

$$\rho = \frac{\sigma_c^2}{\sigma_c^2 + \sigma_u^2}.$$

[13] One example is the provision of weather-linked insurance to farmers, where we expect the variance of consumption to be lower for the treated individuals. Another example is a migration facilitation programme, where the treatment group may include a higher proportion of migrants, leading to greater heterogeneity in outcomes; see McKenzie (2025) for details.

[14] The notion of optimality employed here is based on minimising the MDE subject to a fixed total sample size. In Section 6.1, we consider an alternative optimality criterion: maximising power subject to a budget constraint.

The ICC thus gives a measure of the proportion of the total variance accounted for by the between-clusters variance component. The intuition behind the ICC is that the larger the fraction of the total variance accounted for by the between-clusters variance component (σ_c^2), the more similar are outcomes within the cluster and the less information is gained from adding an extra individual within the cluster. Proceeding as in the simple case above, and assuming that both the number of clusters and the number of units per cluster are equal across treatment and control groups, the expression for the MDE is[15]

$$\delta^2 = (t_{\alpha/2} + t_\beta)^2 2\left(\frac{m\sigma_c^2 + \sigma_u^2}{mk}\right), \tag{5}$$

where there are k clusters per arm and m units per cluster.[16] Using the definition of the ICC, and rearranging, we arrive at the formula for the total sample per arm:[17]

$$n^* = m^* k^* = (t_{\alpha/2} + t_\beta)^2 2\frac{\sigma^2}{\delta^2}\big(1 + (m-1)\rho\big). \tag{6}$$

Comparing equations (2) and (6), the key difference is the term $(1 + (m-1)\rho)$, which is commonly referred to in the literature as either the design effect or the variance inflation factor (VIF). This term is a consequence of the clustered treatment allocation and leads to larger required sample sizes. There is another difference between equations (2) and (6): in the cluster randomised case, the degrees of freedom for the t-statistic are $2(k-1)$, while in the individually randomised case they are $2(n-1)$. This difference is not taken into account when the sample size for a cluster RCT is computed by first calculating the sample size for an individually randomised trial and then multiplying it by the design effect. In practice, this will usually make little difference unless the number of clusters is small. The methods used in this paper account for the correct degrees of freedom, and hence our results may differ slightly from those produced by software that does not incorporate this adjustment.

In order to get a sense of the interplay between the ICC and the number of units per cluster, Table 1 presents required sample sizes for two different values of the MDE, δ, and six different values of the ICC, ρ. For reference, the standard deviation of earnings and the ICC are 126,383.5 and 0.042 respectively in the EvaluaLand National Survey of Earnings data.[18]

Consider first the upper left quadrant of the table. The case where ICC = 0 represents unit-level randomisation. As the ICC increases, so too does the sample size. The extent of the increase depends also on m, the other key term in the VIF. For instance, for a $\rho = 0.03$ and $m = 60$, a cluster RCT requires almost triple the sample size per arm of a unit-level randomisation equivalent (7,004 compared with 2,508).

Another way to see this is to consider the upper right quadrant. At low levels of the intra-cluster correlation, there is a marked decline in the number of clusters per arm as we increase m (the number of units per cluster). For $\rho = 0.01$, k drops from 274 to 51, 19 per cent of the initial value, as we move from left to right. As the ICC increases, this decline is much shallower. For $\rho = 0.2$, the right-hand

[15] With clustering, and assuming equal variances for the two groups, the standard error of $\hat{\beta}$ takes the form

$$\sqrt{\left(\frac{\sigma_c^2}{k} + \frac{\sigma_u^2}{mk}\right) + \left(\frac{\sigma_c^2}{k} + \frac{\sigma_u^2}{mk}\right)} = \sqrt{2\frac{m\sigma_c^2 + \sigma_u^2}{mk}}.$$

[16] In the clustered case, the degrees of freedom of the t distribution are $2(k-1)$.

[17] To operationalise this formula, one can either solve for m as a function of k or solve for k as a function of m. In the latter case (due to the fact that the degrees of freedom of the t distribution are a function of the number of clusters ($2(k-1)$ in the absence of covariates)), it is necessary to use an iterative process to ensure that the correct degrees of freedom ($2(k^*-1)$) are used to calculate the number of clusters. This issue will be more pronounced when the number of clusters is small.

[18] In Appendix B, we show how to compute the ICC using Stata.

TABLE 1 Sample size requirements per arm for continuous outcomes under cluster-level randomisation

ICC (ρ)	Total sample size per arm (n^*)				Number of clusters per arm (k^*)			
	Number of individuals per cluster (m^*)				Number of individuals per cluster (m^*)			
	10	30	60	100	10	30	60	100
				Effect size = 10,000				
0	2,508	2,508	2,508	2,508	251	84	42	25
0.01	2,743	3,264	4,046	5,089	274	109	67	51
0.03	3,194	4,718	7,004	10,053	319	157	117	101
0.05	3,646	6,173	9,963	15,017	365	206	166	150
0.1	4,774	9,808	17,360	27,428	477	327	289	274
0.2	7,030	17,079	32,153	52,251	703	569	536	523
				Effect size = 20,000				
0	628	628	628	628	63	21	10	6
0.01	693	839	1,058	1,351	69	28	18	14
0.03	806	1,202	1,796	2,589	81	40	30	26
0.05	919	1,565	2,536	3,829	92	52	42	38
0.1	1,201	2,474	4,384	6,931	120	82	73	69
0.2	1,765	4,292	8,083	13,136	177	143	135	131

Note: The cells in the left panels report the sample size per arm (n^*) and those in the right panels report the number of clusters per arm (k^*) required to achieve 80 per cent power at 5 per cent significance if the effect size is either 10,000 (top panel) or 20,000 (bottom panel) and the standard deviation is 126,383.5. The intra-cluster correlation (ρ) is given in the first column of the table.

value for k is 74 per cent of the initial value. It should be clear from this table that it is very important to get accurate measures of the ICC. Small differences in the values of the ICC, such as moving from $\rho = 0.01$ to 0.03, can have significant impacts on the required sample size, particularly when m is large.

Finally, comparing the upper and lower panels of Table 1 illustrates the effect of the MDE: the larger the value of δ, the smaller the sample size required to detect a statistically significant effect.

4.1.1 | Unequal numbers of clusters and units per cluster

Keeping the same number of clusters and units per cluster in the treatment and control arms is common practice, as it minimises the total sample size required to achieve a given level of power. However, there are situations in which departing from this balanced allocation – by allowing for a different number of clusters and/or a different number of units per cluster in treatment and control arms – may be advantageous. One example is when the implementer's capacity constraints limit the number of clusters that can be assigned to treatment. Another is when costs are higher in treatment than in control clusters, and the goal is either to minimise total cost subject to achieving a target power level or to maximise power subject to a fixed budget; see McConnell and Vera-Hernández (2022) for precise methods. In the case of unequal allocation, the required number of treatment clusters (k_1) can be computed as a function of the MDE, δ, the number of control clusters, k_0, and the number of units per cluster, m, using the following formula:

$$k_1 = \frac{(t_{\alpha/2} + t_\beta)^2\left((m\sigma_c^2 + \sigma_u^2)/m\right)}{\delta^2 - (t_{\alpha/2} + t_\beta)^2\left((m\sigma_c^2 + \sigma_u^2)/mk_0\right)}, \tag{7}$$

which assumes that the number of units per cluster is the same in the treatment and control arms. Expression (7) can also be written in terms of the design effect as

$$k_1 = \frac{(t_{\alpha/2} + t_\beta)^2\sigma^2\big((1 + (m-1)\rho)/m\big)}{\delta^2 - (t_{\alpha/2} + t_\beta)^2\sigma^2\big((1 + (m-1)\rho)/mk_0\big)}. \tag{8}$$

The formula for the number of units per treatment cluster (m_1) as a function of the MDE, δ, the number of units per control cluster, m_0, and the number of clusters per arm, k, is given by

$$m_1 = \frac{(t_{\alpha/2} + t_\beta)^2(\sigma_u^2/k)}{\delta^2 - (t_{\alpha/2} + t_\beta)^2(2\sigma_c^2/k + \sigma_u^2/m_0k)}, \tag{9}$$

which assumes that the number of clusters in the treatment arm is the same as in the control arm. Rewriting expression (9) in terms of ρ and σ yields

$$m_1 = \frac{(t_{\alpha/2} + t_\beta)^2\sigma^2\big((1-\rho)/k\big)}{\delta^2 - (t_{\alpha/2} + t_\beta)^2\sigma^2\big((1 + (2m_0 - 1)\rho)/m_0k\big)}. \tag{10}$$

4.2 | The role of covariates

Although, due to randomisation, covariates are not used to partial out differences between treatment and control, they can be very useful in reducing the residual variance of the outcome variable, and subsequently lead to lower required sample sizes.

There are several equivalent ways of expressing the power calculation formula with covariates. Below, we present multiple formulations, as the choice of which to use in practice often depends on the specific inputs available to the researcher.

The simplest or most intuitive version is as follows:

$$n^* = m^*k^* = (t_{\alpha/2} + t_\beta)^2 2\frac{\sigma_x^2}{\delta^2}\big(1 + (m-1)\rho_x\big), \tag{11}$$

where σ_x^2 is the conditional variance (i.e. the residual variance once the covariates have been controlled for) and $\rho_x = \sigma_{x,c}^2/(\sigma_{x,c}^2 + \sigma_{x,u}^2)$ is the conditional ICC.[19] The form of equation (11) mirrors that of the unconditional representation in equation (6). If there are data from a similar context and target population with the relevant variables, as in the case of APRENDE, it is straightforward to get estimates of these conditional parameters.[20] However, if such data are not available, the formulation by Bloom, Richburg-Hayes and Black (2007) might be easier to apply:

$$n^* = m^*k^* = (t_{\alpha/2} + t_\beta)^2 2\frac{\sigma^2}{\delta^2}\big(m\rho(1 - R_c^2) + (1-\rho)(1 - R_u^2)\big), \tag{12}$$

where R_c^2 is the proportion of the cluster-level variance component explained by the covariates and R_u^2 is the unit-level equivalent. This formulation is useful to see the differing impact of covariates at different levels of aggregation, i.e. if the covariates are at the unit or cluster level. For instance, a

[19] In the case with covariates, the number of degrees of freedom of the t distribution is $2(k-1) - J$, where J is the number of covariates.

[20] Refer to Appendix B to see how to estimate these parameters.

TABLE 2 Sample size requirements per arm for continuous outcomes under cluster-level randomisation with a covariate

	Number of individuals per cluster (m) = 100					Number of individuals per cluster (m) = 20					Number of individuals per cluster (m) = 8				
R_c^2	R_u^2					R_u^2					R_u^2				
	0	0.1	0.2	0.4	0.5	0	0.1	0.2	0.4	0.5	0	0.1	0.2	0.4	0.5
								ICC (ρ) = 0.01							
0	1,351	1,289	1,228	1,104	1,043	766	704	642	518	456	679	617	554	430	368
0.1	1,289	1,227	1,165	1,042	981	753	691	629	505	443	674	612	549	425	363
0.2	1,226	1,165	1,103	980	919	741	679	617	493	431	669	607	544	420	358
0.4	1,102	1,040	979	856	795	716	654	592	468	406	659	596	534	410	348
0.5	1,040	978	917	794	733	703	641	579	455	393	654	591	529	405	343
								ICC (ρ) = 0.3							
0	19,342	19,298	19,254	19,166	19,123	4,219	4,176	4,132	4,044	4,000	1,951	1,907	1,863	1,776	1,732
0.1	17,462	17,418	17,374	17,286	17,242	3,843	3,799	3,756	3,668	3,624	1,801	1,757	1,713	1,625	1,581
0.2	15,581	15,537	15,493	15,406	15,362	3,467	3,423	3,379	3,292	3,248	1,650	1,606	1,562	1,475	1,431
0.4	11,820	11,776	11,732	11,645	11,601	2,715	2,671	2,627	2,540	2,496	1,349	1,305	1,262	1,174	1,130
0.5	9,940	9,896	9,852	9,764	9,720	2,339	2,295	2,251	2,163	2,120	1,199	1,155	1,111	1,023	979

Note: Each cell reports the sample size per arm required to achieve 80 per cent power at 5 per cent significance if the effect size is 20,000 and the standard deviation is 126,383.5. The number of units per cluster (m) is 100 in the left panel, 20 in the middle panel and 8 in the right panel. The intra-cluster correlation (ρ) is 0.01 in the top panel and 0.3 in the bottom panel. R_c^2 is the proportion of the cluster-level variance component explained by the covariate, and R_u^2 is the unit-level equivalent. The covariate is a single unit-level variable.

unit-level covariate can affect both R_u^2 *and* R_c^2, whilst a cluster-level covariate can only increase R_c^2. Equation (12) may be useful if R_u^2 and R_c^2 are reported in previous research and the parameters in equation (11) are not. To reiterate, with a series of calculations, it is straightforward to move from equation (12) to equation (11), using R_c^2, R_u^2, σ^2 and ρ to obtain values for σ_x^2 and ρ_x.[21]

Finally, Hedges and Rhoads (2010) present the formula for the inclusion of covariates as

$$n^* = m^*k^* = (t_{\alpha/2} + t_\beta)^2 2\frac{\sigma^2}{\delta^2}\left[\left(1 + (m-1)\rho\right) - \left(R_u^2 + (mR_c^2 - R_u^2)\rho\right)\right].$$

This equation is useful for building intuition into the role of covariates, as the first term in the square brackets is the regular design effect, whilst the second shows how covariates impact the overall variance inflation factor.

Table 2 presents how the inclusion of a covariate impacts the required sample sizes for six different scenarios ($m = 8$, 20 and 100, $\rho = 0.01$ and 0.3). Values for the standard deviation come from the 2024 earnings variable of the EvaluaLand National Survey of Earnings. As is clear from equation (12), the larger either R_u^2 or R_c^2 is, the smaller the sample size per arm is. Note from equation (12) that the influence of R_u^2 is larger when ρ is smaller. For example, in Table 2, when $\rho = 0.01$, $m = 100$ and $R_c^2 = 0$, the sample size per arm decreases from 1,351 to 1,043 (a 23 per cent reduction) when R_u^2 increases from 0 to 0.5. However, the same increase in R_u^2 only translates into a decrease from 19,342 to 19,123 (a 1 per cent reduction) when $\rho = 0.3$. In this sense, increasing R_u^2 is similar to increasing the number of units per cluster, which has little effect on power when ρ is high.

As is also clear from equation (12), the effect of R_c^2 is mediated by $m\rho$, so the reduction in sample size achieved by increasing R_c^2 will be higher when both m and ρ are large. Again, increasing R_c^2 is analogous to increasing the number of clusters. This will have a larger effect when ρ is large and when m is large (because a large m indirectly implies that the number of clusters is small, so we obtain a larger effect when we increase the number of clusters). This is also clear in Table 2: when $\rho = 0.3$, $m = 100$ and $R_u^2 = 0$, the sample size per arm decreases from 19,342 to 9,940 (a 49 per cent reduction) when R_c^2 increases from 0 to 0.5. However, the same increase in R_c^2 only translates into a decrease from 679 to 654 (a 3.7 per cent reduction) when $\rho = 0.01$ and $m = 8$.

A final point to note concerns an issue raised by Bloom, Richburg-Hayes and Black (2007) regarding unconditional versus conditional ICCs. As they emphasise, researchers should not be concerned with the possibility that a unit-level covariate, by reducing the unit-level variance component by a larger extent than the cluster-level component, may lead to a higher conditional ICC. What matters is that by reducing both components, unit-level covariates increase precision and thus lower required sample sizes.

4.2.1 | Unequal numbers of clusters

As shown in Section 4.1.1, the sample size equations can be expressed allowing either the number of clusters or the number of units per cluster to differ between the treatment and control arms. First, consider the expression for k_1 as a function of k_0 and m, written in the form presented by Bloom, Richburg-Hayes and Black (2007):[22]

$$k_1 = \frac{(t_{\alpha/2} + t_\beta)^2\sigma^2\left((m\rho(1-R_c^2) + (1-\rho)(1-R_u^2))/m\right)}{\delta^2 - (t_{\alpha/2} + t_\beta)^2\sigma^2\left((m\rho(1-R_c^2) + (1-\rho)(1-R_u^2))/mk_0\right)}.$$

[21] Using the definition of ρ_x, we note that $\sigma^2(1-\rho)(1-R_u^2) = \sigma_{x,u}^2 = (1-\rho_x)\sigma_x^2$ and $\sigma^2\rho(1-R_c^2) = \sigma_{x,c}^2 = \rho_x\sigma_x^2$. This allows us to write the R^2 terms as functions of ρ, σ^2, ρ_x and σ_x^2: $1 - R_c^2 = \rho_x\sigma_x^2/\rho\sigma^2$ and $1 - R_u^2 = (1-\rho_x)\sigma_x^2/((1-\rho)\sigma^2)$. These expressions are used in intermediate steps to move from equation (12) to equation (11).

[22] We can also write an expression for k_1 in the form of either equation (7), where we replace σ_c and σ_u with $\sigma_{x,c}$ and $\sigma_{x,u}$, or equation (8), where we replace σ and ρ with σ_x and ρ_x.

As before, we can also write an expression for m_1 as a function of k and m_0:[23]

$$m_1 = \frac{(t_{\alpha/2} + t_\beta)^2\sigma^2\left((1-\rho)(1-R_u^2)/k\right)}{\delta^2 - (t_{\alpha/2} + t_\beta)^2\sigma^2\left((2m_0\rho(1-R_c^2) + (1-\rho)(1-R_u^2))/m_0k\right)}.$$

4.3 | Difference-in-differences and lagged outcome as a covariate

Where the researcher has data on the outcome variable not only subsequent to treatment, but also prior to treatment (baseline), it is possible to employ a difference-in-differences approach, as well as to include the baseline realisation of the outcome variable as a covariate, a special case of the approach discussed in Section 4.2. Following Teerenstra et al. (2012), the data-generating process (which includes the panel component) follows

$$Y_{ijt} = \gamma_0 + \gamma_1 T_j + \gamma_2 POST_t + \gamma_3(POST_t \times T_j) + c_j + c_{jt} + u_{ij} + u_{ijt},$$

where i indexes units, j clusters and t time periods ($t = 0$, the pre-intervention period, or $t = 1$, the post-intervention period), $POST_t$ takes value 0 if $t = 0$ and 1 if $t = 1$, and T_j is the treatment indicator. The terms c_j, c_{jt}, u_{ij} and u_{ijt} are assumed to be normally distributed with mean zero and variances σ_c^2, σ_{ct}^2, σ_u^2 and σ_{ut}^2 respectively.[24]

The error terms are structured as two cluster-level components (c_j and c_{jt}) and two unit-level components (u_{ij} and u_{ijt}), where c_j and u_{ij} are time-invariant. Two autocorrelation terms are required in this case, namely the unit-level autocorrelation of the outcome over time, ρ_u, and the analogous cluster-level term, ρ_c:

$$\rho_u = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_{ut}^2} \quad \text{and} \quad \rho_c = \frac{\sigma_c^2}{\sigma_c^2 + \sigma_{ct}^2},$$

where $\text{var}(c_j)=\sigma_c^2$, $\text{var}(c_{jt})=\sigma_{ct}^2$, $\text{var}(u_{ij})=\sigma_u^2$ and $\text{var}(u_{ijt})=\sigma_{ut}^2$. The ICC in this situation is expressed as[25]

$$\rho = \frac{\sigma_c^2 + \sigma_{ct}^2}{\sigma_c^2 + \sigma_{ct}^2 + \sigma_u^2 + \sigma_{ut}^2}.$$

Once these parameters are in hand, we can define the key parameter used in sample size calculations, r, which represents the proportion of the total variance attributable to time-invariant components:

$$r = \frac{\sigma_c^2 + \sigma_u^2/m}{\sigma_c^2 + \sigma_{ct}^2 + \sigma_u^2/m + \sigma_{ut}^2/m} = \frac{m\rho}{1 + (m-1)\rho}\rho_c + \frac{1-\rho}{1 + (m-1)\rho}\rho_u.$$

[23] We can also write an expression for m_0 in the form of either equation (9) or equation (10), replacing unconditional parameters with their conditional versions.

[24] We note the abuse of notation in using the subscript t for the variance terms σ_{ct}^2 and σ_{ut}^2, as these terms are constant across the two time periods.

[25] Appendix Section C.5 details how to estimate these key panel data parameters.

TABLE 3 Sample size requirements per arm for continuous outcomes in panel data models

r	Difference	Difference-in-differences	Lagged outcome
0.1	4,909	8,820	4,860
0.25	4,909	7,354	4,603
0.5	4,909	4,909	3,687
0.75	4,909	2,464	2,159
0.9	4,909	998	949

Note: The ICC is 0.05 and the number of individuals per cluster (m) is set to 20. The effect size is equal to 10,000 and the standard deviation is 126,383.5.

The sample size formula for a difference-in-differences estimation can be written as

$$n^* = m^* k^* = 2(1-r)(t_{\alpha/2} + t_\beta)^2 2\frac{\sigma^2}{\delta^2}\big(1 + (m-1)\rho\big) \tag{13}$$

and the sample size formula for an estimation using the baseline outcome variable as a covariate can be written as

$$n^* = m^* k^* = (1-r^2)(t_{\alpha/2} + t_\beta)^2 2\frac{\sigma^2}{\delta^2}\big(1 + (m-1)\rho\big). \tag{14}$$

In order to see the benefit of using the panel element, it is instructive to compare equations (13) and (14) with equation (6). The most important message is that the sample size requirement is minimised by including the baseline level of the outcome variable as a covariate (note that $1 - r^2 < 1$ and that $1 - r^2 < 2(1 - r)$). Alternatively, given a sample, the highest power is achieved by including the baseline value of the outcome variable as covariate. Hence if baseline data on the outcome variable are available, one should always control for them as a covariate rather than doing difference-in-differences or a simple post-treatment comparison (McKenzie, 2012; Teerenstra et al., 2012).

Also, it is useful to see that the largest reduction on sample size requirements when we include the baseline value as covariate takes place when r is close to 1 (hence $1 - r^2$, which multiplies the sample size formula (14), is close to zero). Intuitively, by conditioning on the baseline value of the outcome variable, we are netting out the time-invariant component of the variance (which is large when r is close to 1).

Note also that if r is close to zero, given a sample, there might be little difference in power between including the baseline value of the outcome as a covariate and just post-treatment differences. Hence, from the point of view of power, it might be better to spend the resources devoted to collect the baseline on collecting a larger sample post-treatment or several post-treatment waves (see McKenzie (2012)).[26] Interestingly, in terms of power, including the baseline value of the variable as covariate always dominates over difference-in-differences. Moreover, baseline data are required for both estimators. Hence, there is little reason in terms of power to justify difference-in-differences.

In Table 3, we report the sample size requirements for the three estimation strategies for various values of r, calibrating the calculations to the likely effect size and variance of the earnings for 2024 of the EvaluaLand National Survey of Earnings. The resulting sample sizes quantify the intuition above – the higher the time-invariant component of the variance, r, the greater the benefit of controlling for baseline differences via covariates or difference-in-differences vis-à-vis single post-treatment difference. For low values of r, the difference-in-differences strategy is highly inefficient. The table

[26] There might be other reasons to collect baseline data than gains in power. These include to check whether the sample is balanced in the outcome variables, to collect information that allows stratification of the sample, and to have the basis for heterogeneity analysis (see McKenzie (2012)).

also clearly illustrates the superiority of controlling for the baseline outcome as a covariate, which consistently outperforms the other two strategies across all values of r.

5 | BINARY OUTCOME CASE

5.1 | Unit-level randomisation

We now move on to discussing the case where the outcome variable is binary – for instance, whether an individual is working or not or whether a student obtained a certain grade level or not. There is a large literature that focuses on the binary outcome case, with several different approaches (e.g. Moerbeek and Maas, 2005; Demidenko, 2007). Some articles deal with effect sizes measured in differences in log odds, others with differences in probability of success between treatment and controls. We follow Schochet (2013), who measures the effect size in terms of differences in the probability of success. We believe that this is more intuitive for most economists, and that the required inputs might be more easily accessible from published studies.[27] One difference between the continuous and binary outcome cases is that in the latter, we do not need the variance. Binary outcomes follow a Bernoulli distribution, so knowing p, the probability of success, also yields the variance: $p(1-p)$.

Using a logistic model, we can write the probability of success for individual i as

$$p_i = \text{Prob}(y_i = 1|T_i) = \frac{e^{\gamma_0+\gamma_1 T_i}}{1+e^{\gamma_0+\gamma_1 T_i}},$$

where y_i is binary (takes value 1 in case of success and 0 in case of failure) and, as before, T_i denotes treatment status. The effect size, δ, can thus be written as $p(y_i = 1|T_i = 1) - p(y_i = 1|T_i = 0)$ or $(p_1 - p_0)$, where the subscripts denote treatment and control status respectively.

Following an analogous procedure to that in the continuous case, we arrive at a sample size equation for the binary case (Donner and Klar, 2010):

$$N^* = \left(\frac{p_1(1-p_1)}{\pi} + \frac{p_0(1-p_0)}{1-\pi}\right)\frac{(z_\beta + z_{\alpha/2})^2}{(p_1-p_0)^2}, \tag{15}$$

where π is the proportion of the sample that is treated, $n_1^* = \pi N^*$ and $n_0^* = (1-\pi)N^*$.[28] Note that equation (15) is equivalent to equation (3), where σ_0^2 and σ_1^2 are replaced with their equivalents in the binary case, $p_0(1-p_0)$ and $p_1(1-p_1)$. In general, these variances will be different, so as we saw in equation (3), the optimal treatment–control split will differ from 0.5. The optimal allocation to treatment status, π^*, can be written as

$$\pi^* = \frac{\sqrt{\big(p_1(1-p_1)\big)/\big(p_0(1-p_0)\big)}}{1+\sqrt{\big(p_1(1-p_1)\big)/\big(p_0(1-p_0)\big)}}.$$

[27] An advantage of the approach we follow is that the impact parameter does not depend on whether covariates are included or not. This is not the case when impact is measured in log odds. See Schochet (2013) for a detailed discussion of this.

[28] If the null hypothesis of zero impact is tested using a Pearson's chi-square test, and $n_1^* = n_0^*$, then

$$n^* = \frac{\left(z_{\alpha/2}\sqrt{2\bar{p}(1-\bar{p})} + z_\beta\sqrt{p_1(1-p_1)+p_0(1-p_0)}\right)^2}{(p_0-p_1)^2},$$

where $\bar{p} = (p_1+p_0)/2$; see Fleiss, Levin and Paik (2003), equation (4.14), as well as equation (4.19) for different sample sizes in treatment and control.

Hence, in the binary outcome case, the optimal split would only equal 0.5 in the special case where $p_0 = 1 - p_1$, e.g. $p_0 = 0.4$ and $p_1 = 0.6$.[29] In the case of an even split between treatment and control status ($\pi = 0.5$), we can write n^* as

$$n^* = \big(p_1(1-p_1) + p_0(1-p_0)\big)\frac{(z_\beta + z_{\alpha/2})^2}{\delta^2}.$$

5.2 | Cluster-level randomisation

Having considered the unit-level treatment case, we now move to cluster randomised treatment, still following Schochet (2013), as we do for the rest of Section 5. For the cluster randomised case, a generalised estimating equation (GEE) approach is followed, where the clustering is accounted for in the variance-covariance matrix, using the ICC, ρ. As before, we can write the probability of success for individual i in cluster j as

$$p_{ij} = \text{Prob}(y_{ij} = 1|T_j) = \frac{e^{\gamma_0 + \gamma_1 T_j}}{1 + e^{\gamma_0 + \gamma_1 T_j}}.$$

For cluster j, the $m \times m$ variance-covariance matrix V_j is written as

$$V_j = A_j^{1/2} R(\rho) A_j^{1/2}, \tag{16}$$

where A_j is a diagonal matrix with diagonal elements $p_{ij}(1 - p_{ij})$ and $R(\rho)$ is a correlation matrix with diagonal elements taking the value of 1 and off-diagonals the value of ρ. Hence $\text{cov}(y_{ij}, y_{km}) = \rho$ when $j = m$ and $\text{cov}(y_{ij}, y_{km}) = 0$ when $j \neq m$. Note the lack of a j subscript on $R(\rho)$ – it is taken as common across clusters, as in the generalised least squares (GLS) approach for a continuous outcome. This means that we no longer specify a random effect for each cluster, which allows us to get closed-form solutions for the sample size equations.[30]

The sample size equation for the binary outcome case with cluster randomisation can be written as

$$N^* = \left(\frac{p_1(1-p_1)}{\pi} + \frac{p_0(1-p_0)}{1-\pi}\right)\frac{(z_{\alpha/2} + z_\beta)^2}{\delta^2}\big(1 + (m-1)\rho\big),$$

where π is the fraction of clusters randomised to receive treatment, $n_1^* = mk_1^* = \pi N^*$ and $n_0^* = mk_0^* = (1 - \pi)N^*$. As above, if the treatment is evenly allocated, we can write this as

$$n^* = mk^* = \big(p_1(1-p_1) + p_0(1-p_0)\big)\frac{(z_\beta + z_{\alpha/2})^2}{\delta^2}\big(1 + (m-1)\rho\big).$$

As before, the sample size equation for the binary outcome mirrors that of the continuous outcome, with the design effect being the only difference between the unit and cluster randomised sample size equations.

Table 4 presents sample size requirements for three different levels of success probability for the control groups, p_0: 0.1, 0.3 and 0.5. The first thing to notice is that the closer p_0 is to 0.5, the larger the

[29] While setting $p_0 = p_1$ would result in an equal allocation between treatment arms, it effectively assumes that the intervention has no impact under the alternative hypothesis. As such, it is unsuitable for power calculations intended to detect a treatment effect.

[30] Results from simulations we ran utilising the GEE approach yielded very similar results to those using a linear probability model with random effects. Schochet (2009) finds similar results using GEE and random effects logit models too.

TABLE 4 Sample size requirements for discrete outcomes under cluster-level randomisation

ICC	Total sample size requirements (N^*)				Number of clusters ($2k^*$)			
	Number of individuals per cluster (m)				Number of individuals per cluster (m)			
	10	30	60	100	10	30	60	100
	Control group success rate (p_0) = 0.1							
0	392	392	392	392	39	13	7	4
0.01	428	506	624	781	43	17	10	8
0.03	498	734	1,087	1,558	50	24	18	16
0.05	569	961	1,550	2,335	57	32	26	23
0.1	746	1,531	2,708	4,278	75	51	45	43
0.2	1,099	2,669	5,023	8,163	110	89	84	82
	Control group success rate (p_0) = 0.3							
0	706	706	706	706	71	24	12	7
0.01	770	911	1,123	1,406	77	30	19	14
0.03	897	1,321	1,957	2,804	90	44	33	28
0.05	1,024	1,731	2,790	4,203	102	58	47	42
0.1	1,342	2,755	4,874	7,700	134	92	81	77
0.2	1,978	4,804	9,042	14,693	198	160	151	147
	Control group success rate (p_0) = 0.5							
0	769	769	769	769	77	26	13	8
0.01	838	992	1,223	1,531	84	33	20	15
0.03	977	1,438	2,131	3,054	98	48	36	31
0.05	1,115	1,885	3,038	4,577	112	63	51	46
0.1	1,461	3,000	5,307	8,384	146	100	88	84
0.2	2,154	5,230	9,846	15,999	215	174	164	160

Note: Effect size is set to 0.1 and treatment is evenly allocated ($\pi = 0.5$). The cells display total – rather than per-arm – sample size requirements and number of clusters.

sample size required. This is because for a binary variable, variance is largest at $p = 0.5$. For example, for $m = 30$ and $\rho = 0.03$, the sample size for $p_0 = 0.5$ is double that for $p_0 = 0.1$. As in the continuous case, we see that higher ICCs and larger cluster sizes, m, lead to larger required total samples. This is due to the design effect.

5.2.1 | Unequal numbers of clusters

It might be useful to have a formula for k_1 as a function of m and k_0, that will provide power of $(1 - \beta)$ for the given m and k_0:

$$k_1 = \frac{\big(p_1(1-p_1)/m\big)(z_{\alpha/2} + z_\beta)^2\big(1 + (m-1)\rho\big)}{\delta - \big(p_0(1-p_0)/mk_0\big)(z_{\alpha/2} + z_\beta)^2\big(1 + (m-1)\rho\big)}.$$

5.3 | The role of covariates

In this subsection, we consider the case of unit-level treatment allocation where one has a single covariate, X_i, that is discrete, but not necessarily binary. In the case where the X_i is continuous, one can discretise the variable. Here, we write p_i as

$$p_i = \text{Prob}(y_i = 1|T_i, X_i) = \frac{e^{\gamma_0+\gamma_1 T_i+\gamma_2 X_i}}{1 + e^{\gamma_0+\gamma_1 T_i+\gamma_2 X_i}}.$$

Where covariates are included, we need several extra inputs into the sample size equation, relating to the distribution of the covariates and how success probabilities change according to the covariate values.

First, assume that X_i can take any of the following Q values, $x_1, \ldots, x_Q$. Define $\theta_q = \text{Prob}(X_i = x_q)$ for $q \in 1, \ldots, Q$, with $0 < \theta_q < 1$ and $\sum_q \theta_q = 1$. Next we need to specify how success probabilities change across the values of X_i. Define $p_{0q} = \text{Prob}(Y_i = 1|T_i = 0, X_i = x_q)$ and $p_{1q} = \text{Prob}(Y_i = 1|T_i = 1, X_i = x_q)$. Then we can define an effect size for a specific value of q, $\delta_q = p_{1q} - p_{0q}$, and an overall effect size, $\delta = \sum_q \theta_q \delta_q$. Schochet (2013) notes that covariate inclusion will improve efficiency if at least two of the p_{0q} or p_{1q} probabilities differ across covariate values.

With these inputs at hand, we can now write the sample size equation as

$$N^* = (\mathbf{g}\mathbf{M}^{-1}\mathbf{g}')\frac{(z_\beta + z_{\alpha/2})^2}{\delta^2},$$

where

$$\mathbf{M} = \begin{bmatrix} m_1 & m_2 & m_3 \\ m_2 & m_2 & m_4 \\ m_3 & m_4 & m_5 \end{bmatrix},$$

$$m_1 = \sum_q \pi\theta_q p_{1q}(1 - p_{1q}) + (1 - \pi)\theta_q p_{0q}(1 - p_{0q}),$$

$$m_2 = \sum_q \pi\theta_q p_{1q}(1 - p_{1q}),$$

$$m_3 = \sum_q x_q\pi\theta_q p_{1q}(1 - p_{1q}) + (1 - \pi)\theta_q p_{0q}(1 - p_{0q}),$$

$$m_4 = \sum_q x_q\pi\theta_q p_{1q}(1 - p_{1q}),$$

$$m_5 = \sum_q x_q^2\pi\theta_q p_{1q}(1 - p_{1q}) + (1 - \pi)\theta_q p_{0q}(1 - p_{0q}),$$

and $\mathbf{g}$ is a 1×3 gradient vector with elements

$$\mathbf{g}[1, 1] = \sum_q \theta_q[p_{1q}(1 - p_{1q}) - p_{0q}(1 - p_{0q})],$$

$$\mathbf{g}[1, 2] = \sum_q \theta_q[p_{1q}(1 - p_{1q})],$$

$$\mathbf{g}[1,3] = \sum_q x_q \theta_q [p_{1q}(1-p_{1q}) - p_{0q}(1-p_{0q})].$$

In Appendix D, we provide a purposefully designed Stata program to carry out this computation for five different values of the covariate.[31]

5.3.1 | Cluster-level randomisation

Finally, we consider a cluster randomised treatment in the presence of a single, discrete cluster-level covariate. Candidates for this could be a discrete cluster characteristic or a continuous variable, such as cluster means of the outcome variable at baseline, which are then discretised. We write the probability of success here as

$$p_{ij} = \text{Prob}(y_{ij} = 1 | T_j, X_j) = \frac{e^{\gamma_0 + \gamma_1 T_j + \gamma_2 X_j}}{1 + e^{\gamma_0 + \gamma_1 T_j + \gamma_2 X_j}}.$$

The variance-covariance matrix in this scenario is very similar to that without a covariate (see equation (16)) with the exception of the use of the conditional ICC, ρ_x, not the raw ICC (ρ), in the correlation matrix. The sample size calculation for this subsection can be expressed as

$$N^* = 2m^*k^* = (\mathbf{g}\mathbf{M}^{-1}\mathbf{g}')\frac{(z_\beta + z_{\alpha/2})^2}{\delta^2}\left(1 + (m-1)\rho_x\right),$$

where ρ_x is the conditional ICC, as we saw in the continuous outcome case with cluster randomisation and covariates. Note that the inclusion of a cluster-level covariate can lead to precision gains through decreasing the total residual variance, as well as by decreasing the conditional ICC. Schochet (2013) suggests that the latter will have more impact on lowering the required sample size.

Table 5 presents the number of clusters required in the binary outcome case for two values of ρ, 0.05 and 0.1, and a binary covariate. What we see here is that the greater is the difference between p_{00} and p_{01} (the difference in control group success rates for the two values of the covariate), the greater is the sample size reduction due to the inclusion of the covariate. The number of clusters required (for $m = 60$, $\rho = 0.05$, $p_0 = 0.5$, and a constant effect size of 0.1 across covariate levels) is 51 in the

TABLE 5 Number of clusters required for binary outcomes under cluster-level randomisation with a binary covariate

Control group success rates for $X_j = 0$ / $X_j = 1$	ICC = 0.05			ICC = 0.1		
	Impacts for $X_j = 0$ / $X_j = 1$			Impacts for $X_j = 0$ / $X_j = 1$		
	0.1 / 0.1	0.05 / 0.15	0.03 / 0.17	0.1 / 0.1	0.05 / 0.15	0.03 / 0.17
0.45 / 0.55	50	49	49	88	86	85
0.4 / 0.6	49	47	47	85	83	81
0.3 / 0.7	42	40	39	74	70	68
0.2 / 0.8	32	29	27	56	50	47

Note: Number of units per cluster (m) is set at 60. The overall base rate in this table is set to 0.5, with the overall impact set to 0.1. Treatment is evenly allocated ($\pi = 0.5$), and $\theta = \text{Prob}(X_j = 1) = 0.5$. The cells display total – rather than per-arm – number of clusters.

[31] Our Stata programs that accommodate two, three, four and five possible values of the covariate can be found at https://ifs.org.uk/publications/sample-size-calculators-going-beyond-simple-sample-size-calculations-practitioners. Schochet (2013) provides a set of SAS programs for sample size calculations for binary outcomes.

absence of a covariate (bottom right section of Table 4). In Table 5, this number falls to 49 when $p_{00} = 0.4$ and $p_{01} = 0.6$, and falls markedly to 32 when $p_{00} = 0.2$ and $p_{01} = 0.8$.

6 | EXTENSIONS

In this section, we provide three extensions that we think are particularly useful for researchers. The first extension discusses how to choose optimally the number of clusters versus the number of units within clusters. The second extension explains how to compute power using simulation methods, which are useful when there are no existing formulae for the RCT that is being planned. The third extension shows how to adapt the sample size computations when several outcomes are used.

6.1 | Choosing the number of clusters versus units per cluster

The same MDE can be obtained with different combinations of k, the number of clusters per arm, and m, the number of units per cluster (see equation (5)). The question arises how to choose amongst the different combinations. One common criterion is to choose the combination that maximises power subject to a budget constraint. Consider the case in which the costs of the RCT comprise a fixed cost per cluster, denoted by f, and a unit constant marginal cost, denoted by v. Hence, the total cost function of the cluster RCT takes the form[32]

$$C = 2k(f + vm).$$

Minimising the square of the MDE (as given in equation (5)) subject to this cost constraint yields optimal values for m:

$$m^* = \sqrt{\frac{f}{v}\frac{\sigma_u^2}{\sigma_c^2}}. \tag{17}$$

Using this formula and the cost function, we derive an expression for the optimal number of clusters per arm, k:

$$k^* = \frac{C}{2\left(f + v\sqrt{(f/v)(\sigma_u^2/\sigma_c^2)}\right)}. \tag{18}$$

As Liu (2013) notes, it may be instructive to use the definition of the ICC to rewrite equations (17) and (18) as

$$m^* = \sqrt{\frac{f}{v}\frac{1-\rho}{\rho}} \quad \text{and} \quad k^* = \frac{C}{2\left(f + v\sqrt{(f/v)((1-\rho)/\rho)}\right)}, \tag{19}$$

[32] This formulation assumes that the fixed cost of a treatment cluster is equal to that of a control cluster and that the marginal cost per unit is the same across both arms. These assumptions are reasonable when the research budget does not cover the cost of delivering the intervention. However, when intervention costs are part of the research budget, treatment is typically more expensive than control. In such cases, the optimal sample allocation will generally require differing numbers of clusters and units in the treatment and control arms; see McConnell and Vera-Hernández (2022) for detailed methods.

where it is clear that the larger the ICC is, the smaller the optimal m is. This is due to the fact that when the ICC is high, outcomes within clusters are highly correlated, and increasing the number of units per cluster, m, adds little in precision gains. Resources are better spent by increasing the number of clusters per arm, k.

By substituting the optimal values m^* and k^* in expression (6) and solving for δ, we can compute the MDE that can be achieved given the budget constraint:

$$\delta^* = (t_{\alpha/2} + t_\beta)\sqrt{2\sigma^2\left(\frac{1+(m^*-1)\rho}{m^*k^*}\right)}. \tag{20}$$

By combining (19) and (20), we derive an expression for the minimum total cost, C^*, required in order to achieve a power of $1-\beta$ for a given value of δ:

$$C^* = \frac{4\sigma^2}{\delta^2}(t_{\alpha/2} + t_\beta)^2\left[f + v\sqrt{\frac{f}{v}\frac{1-\rho}{\rho}}\right]\left(\frac{1+\left(\sqrt{(f/v)((1-\rho)/\rho)}-1\right)\rho}{\sqrt{(f/v)\left((1-\rho)/\rho\right)}}\right).$$

There are cases where the optimal allocation cannot be implemented – for example, when the number of available clusters is smaller than the optimal, or when clusters have fewer units than required. In such instances, expression (6) can be used to solve for the required number of units per cluster, m, given a fixed number of clusters, k, or vice versa. A further consideration is that when the number of clusters is small, standard inference procedures based on cluster-robust t-statistics tend to over-reject under the null hypothesis (Cameron, Gelbach and Miller, 2008; MacKinnon, Nielsen and Webb, 2023). In such cases, inference is typically conducted using the wild cluster bootstrap procedure. However, to the best of our knowledge, there are no widely accepted sample size calculation methods when the analysis is going to be conducted using the wild cluster bootstrap. Hence, it seems preferable to avoid a situation with too few clusters.

6.2 | Simulation

A researcher might need to compute the required sample size for an experiment whose features do not conform to the ones indicated in previous sections. The possibilities of variation are endless. They include experiments in which the number of units per cluster varies across clusters, experiments with more than two treatment arms or using data from more than two time periods, to say a few. In situations where some features of the experimental design vary significantly with respect to the canonical cases given above, simulation methods can be very useful to estimate the power of a given design, and correspondingly adjust the sample of the design to achieve the desired level of power.

To understand the logic of the simulation approach, it is useful to remember the definition of power: the probability that the intervention is found to have an effect on outcomes when that effect is true. In a hypothetical scenario in which the researcher happened to have 1,000 samples as the ones of their study, and if they could be certain that 'the effect is true' in all these samples, then they could estimate such probability (power) by simply counting in how many of these samples they 'find' the effect (the null hypothesis of zero effect is rejected) and dividing it by 1,000.

The simulation approach simply operationalises the above by providing the researcher with 1,000 (or more) computer-generated samples, hopefully similar to the one of their study (or, at least, obtained under the assumptions that the researcher is planning to use in the study). Because these are computer-generated samples, the researcher can obtain these samples imposing the constraint that the effect is

true (and, in particular, it will draw the samples assuming that the effect of the intervention is the same as the effect size, δ, for which the researcher wants to estimate the power).

In general, the steps required to estimate the power of a given design through simulation are as follows (see Appendix C for an example):[33]

Step 1. Define the number of simulations that will be used to estimate the power of the design (say S), as well as the significance level for the tests.

Step 2. Define a model that will be used to draw computer-generated samples to be as 'those in the study'. This model will have a non-stochastic part (sample size, number of clusters, distribution of the sample across clusters, number of time periods, ICC, autocorrelation terms, mean and standard deviation of the outcome variable, effect size, etc.) and a stochastic part (error term).[34] An example of such a model could be equation (4) but with specific values for the effect size, standard deviation and ICC (in Appendix Section C.6, these are set as $\delta = 4$, $\sigma = 10$ and $\rho = 0.3$).

Step 3. Using computer routines for pseudo-random numbers, obtain a draw of the error term (or composite of error terms) for each unit in the sample. It is crucial that the error term is drawn taking into account the stochastic structure of our experiment (the correlation of draws amongst different units and time periods through the ICC or similar parameters). To draw samples from the error terms, a distribution will need to be assumed. Although assuming normality is common, the approach allows the assumption of other distributions that might be more appropriate for the specific experiment.

Step 4. Using the model and parameter values indicated in Step 2 and the sample of the error term (or composite of error terms) generated in Step 3, obtain the values of the outcome variable for the sample. Once this is done, the draws of the error term generated in Step 3 can be discarded.

Step 5. Using the data on outcomes generated in Step 4 and the model of Step 2, test the null hypothesis of interest (usually, that the intervention has no effect).[35] Keep a record of whether the null hypothesis has been rejected or not.

Step 6. Repeat Steps 3–5 for S times.

Step 7. The estimated power is the number of times that the null hypothesis was rejected in Step 5 divided by S.

Although using simulation methods to estimate power has a long tradition in statistics, the approach is not so commonly used in practice (Arnold et al., 2011).[36] We suspect that Step 3 is the most challenging for the applied researcher. In Appendix C, we provide several hints, which could be of some help.

6.3 | Adjusting sample size calculations for multiplicity

A common problem with experiments (and more generally in empirical work) is that more than one null hypothesis is usually tested. For instance, it is common to test the effect of the intervention on more than one outcome variable. This creates a problem because the number of rejected null hypotheses (the number of outcome variables for which an effect is found) will increase – independently of whether they are true or not – with the number of null hypotheses (outcome variables) tested if the significance level is kept fixed with the number of hypotheses.

[33] Feiveson (2002) provides insightful examples for Poisson regression, Cox regression and the rank-sum test.

[34] If a pilot dataset is available, an alternative approach is to bootstrap from these data (see Kleinman and Huang (2017)).

[35] We are assuming that the test for the null hypothesis has the correct size. Otherwise, see Lloyd (2005).

[36] See Hooper (2013), Kumagai et al. (2014) and Kontopantelis et al. (2016) for some recent implementations of the simulation approach to estimate power.

For instance, consider that we are testing the effect of an intervention on three different outcome variables, and that we use an α equal to 0.05 for each test. If we assume that the three outcome variables are independent, then the probability that we do not reject any of the three null hypotheses when they are all true is $(1-0.05)^3$. Hence, the probability that we reject at least one of them if all three are true is $1-(1-0.05)^3 = 0.14$. Why is this a problem? Assume that the intervention will be declared successful if it is found that it improves at least one of the outcomes. The numbers above imply that the intervention will be declared successful with a probability of 0.14 (larger than the normal significance level of 0.05) even if it has no real effect on any of the three outcome variables.

The problem of multiplicity of outcome variables is recognised by regulatory agencies that approve medicines (US Food and Drug Administration, 1998; European Agency for the Evaluation of Medicinal Products, 2002) and has become more common also in applied work in economics (Anderson, 2008; Carneiro and Ginja, 2014; Mohanan et al., 2021).[37] The standard solution requires performing each individual hypothesis test under an α smaller than the usual 0.05 (Ludbrook, 1998; Romano and Wolf, 2005) so that the probability that at least one null hypothesis is rejected when all null hypotheses are true ends up being 0.05.[38] Hence, when doing the sample size calculations, the researcher should also use a smaller α than 0.05, which will increase the sample size requirements.

When the outcome variables are independent, the probability that at least one null hypothesis is rejected when all are true, usually called the family-wise error rate (FWER), is $1-(1-\alpha)^h$, where α is the level of significance of the individual tests and h is the number of null hypotheses that are tested (i.e. number of outcome variables). Hence, if our study needs a FWER of 0.05, then the significance level for each individual test is given by $1-(1-0.05)^{(1/h)}$, which would be 0.0170 in our example of $h = 3$.[39]

In most experiments, the outcome variables will not be independent. Taking into account this dependency will yield higher values of α, and consequently smaller sample size requirements. If one was willing to assume the degree of dependency amongst the different outcome variables, then a time-consuming but feasible approach to compute the required power is to use the simulation methods previously described combined with a method for Step 5 (testing the null hypothesis) that takes into account the multiple tests carried out and the dependence in the data, such as Westfall and Young (1993) or Romano and Wolf (2005). If this was not available, a rule of thumb is to use $\alpha = 1-(1-0.05)^{(1/\sqrt{h})}$, a correction which was popularised by John W. Tukey (Braun, 1994). This will result in an α larger than when independence is assumed, and hence smaller sample size requirements.

7 | CONCLUSION

In this paper, we have reviewed methods for conducting sample size calculations that go beyond the standard textbook example. The extensions discussed include how to balance the trade-off between the number of clusters and the number of units per cluster, how to adjust calculations when multiple outcomes are considered, and how to use simulation techniques to estimate statistical power in more complex designs not addressed by analytical formulae.

Researchers must make more assumptions when applying the more complex methods presented here than when using simpler approaches. Nevertheless, we argue that the growing availability of publicly accessible datasets places researchers in a relatively strong position to make credible assumptions about key parameters. Moreover, researchers can contribute to cumulative knowledge

[37] There is less consensus on whether correcting for multiplicity is necessary when testing multiple treatments (see Wason, Stecher and Mander (2014)).

[38] An alternative way to analyse the data is to test jointly (through an F-test) the null hypothesis that the intervention does not have an impact on any of the outcome variables considered.

[39] A common simplification is to use the Bonferroni correction, which would be $0.05/h$.

by routinely reporting fundamental quantities – such as intra-cluster correlations, unit- and cluster-level autocorrelations, and R-squared values – in their studies. Journal editors could further facilitate progress by coordinating reporting standards, as has been done in the medical sciences.

ACKNOWLEDGEMENTS

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 and Monica Costa Dias 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. All errors are our own responsibility.

CONFLICT OF INTEREST STATEMENT

There are no conflicts of interest.

DATA AVAILABILITY STATEMENT

Spreadsheets and Stata do-files to implement the methods discussed in this article can be obtained from https://ifs.org.uk/publications/sample-size-calculators-going-beyond-simple-sample-size-calculations-practitioners.

ORCID

Brendon McConnell https://orcid.org/0000-0001-6029-9479
Marcos Vera-Hernández https://orcid.org/0000-0003-4128-8883

REFERENCES

Anderson, M. L. (2008), Multiple inference and gender differences in the effects of early intervention: a reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects. *Journal of the American Statistical Association*, 103, 1481–95.

Arnold, B. F., Hogan, D. R., Colford, J. M., & Hubbard, A. E. (2011), Simulation methods to estimate design power: an overview for applied research. *BMC Medical Research Methodology*, 11, 94.

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

Blundell, R., & Costa Dias, M. (2009), Alternative approaches to evaluation in empirical microeconomics. *Journal of Human Resources*, 44, 565–640.

Braun, H. I. (ed.) (1994), *The Collected Works of John W Tukey, Vol. VIII. Multiple Comparisons: 1948–1983*. Chapman & Hall.

Burtless, G. (1995), The case for randomized field trials in economic and policy research. *Journal of Economic Perspectives*, 9 (2), 63–84.

Cameron, A. C., Gelbach, J. B., & Miller, D. L. (2008), Bootstrap-based improvements for inference with clustered errors. *Review of Economics and Statistics*, 90, 414–27.

Carneiro, P., & Ginja, R. (2014), Long-term impacts of compensatory preschool on health and behavior: evidence from Head Start. *American Economic Journal: Economic Policy*, 6, 135–73.

Cohen, J. (1988), *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates.

Demidenko, E. (2007), Sample size determination for logistic regression revisited. *Statistics in Medicine*, 26, 3385–97.

Donner, A., & Klar, N. (2010), *Design and Analysis of Cluster Randomization Trials in Health Research*. Wiley.

Duflo, E. (2006), Field experiments in development economics. In R. Blundell, W. K. Newey and T. Persson (eds), *Advances in Economics and Econometrics*. Cambridge University Press, 322–48.

Duflo, E., Glennerster, R., & Kremer, M. (2007), Using randomization in development economics research: a toolkit. In T. P. Schultz and J. A. Strauss (eds), *Handbook of Development Economics*, Vol. 4. Elsevier, 3895–962.

European Agency for the Evaluation of Medicinal Products (2002), Points to consider on multiplicity issues in clinical trials. CPMP/EWP/908/99, https://www.ema.europa.eu/en/documents/scientific-guideline/points-consider-multiplicity-issues-clinical-trials_en.pdf.

Feiveson, A. H. (2002), Power by simulation. *Stata Journal*, 2, 107–24.

Fleiss, J. L., Levin, B., & Paik, M. C. (2003), *Statistical Methods for Rates and Proportions*. Wiley-Interscience.

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

Heckman, J. J., & Smith, J. A. (1995), Assessing the case for social experiments. *Journal of Economic Perspectives*, 9 (2), 85–110.

Hedges, L. V., & Rhoads, C. (2010), *Statistical Power Analysis in Education Research*. Institute of Education Sciences.

Hooper, R. (2013), Versatile sample-size calculation using simulation. *The Stata Journal*, 13 (1), 21–38.
Imbens, G. W., & Wooldridge, J. M. (2009), Recent developments in the econometrics of program evaluation. *Journal of Economic Literature*, 47, 5–86.
Kleinman, K., & Huang, S. S. (2017), Calculating power by bootstrap, with an application to cluster-randomized trials. *eGEMs (Generating Evidence & Methods to Improve Patient Outcomes)*, 4 (1), 1202.
Kontopantelis, E., Springate, D. A., Parisi, R., & Reeves, D. (2016), Simulation-based power calculations for mixed effects modeling: ipdpower in Stata. *Journal of Statistical Software*, 74 (12), 1–25.
Kumagai, N., Akazawa, K., Kataoka, H., Hatakeyama, Y., & Okuhara, Y. (2014), Simulation program to determine sample size and power for a multiple logistic regression model with unspecified covariate distributions. *Health*, 6, 2973–98.
Levitt, S. D., & List, J. A. (2009), Field experiments in economics: the past, the present, and the future. *European Economic Review*, 53, 1–18.
List, J., Sadoff, S., & Wagner, M. (2011), So you want to run an experiment, now what? Some simple rules of thumb for optimal experimental design. *Experimental Economics*, 14, 439–57.
Liu, X. S. (2013), *Statistical Power Analysis for the Social and Behavioral Sciences: Basic and Advanced Techniques*. Routledge.
Lloyd, C. J. (2005), Estimating test power adjusted for size. *Journal of Statistical Computation and Simulation*, 75, 921–33.
Ludbrook, J. (1998), Multiple comparison procedures updated. *Clinical and Experimental Pharmacology and Physiology*, 25, 1032–7.
MacKinnon, J. G., Nielsen, M., & Webb, M. D. (2023), Cluster-robust inference: a guide to empirical practice. *Journal of Econometrics*, 232, 272–99.
McConnell, B., & Vera-Hernández, M. (2022), More powerful cluster randomized control trials. Institute for Fiscal Studies, Working Paper 22/22.
McKenzie, D. (2012), Beyond baseline and follow-up: the case for more T in experiments. *Journal of Development Economics*, 99, 210–21.
McKenzie, D. (2025), Designing and analysing powerful experiments: practical tips for applied researchers. *Fiscal Studies*, 46, 305–22.
Moerbeek, M., & Maas, C. J. M. (2005), Optimal experimental designs for multilevel logistic models with two binary predictors. *Communications in Statistics – Theory and Methods*, 34, 1151–67.
Mohanan, M., Donato, K., Miller, G., Truskinovsky, Y., & Vera-Hernández, M. (2021), Different strokes for different folks? Experimental evidence on the effectiveness of input and output incentive contracts for health care providers with varying skills. *American Economic Journal: Applied Economics*, 13 (4), 34–69.
Romano, J. P., & Wolf, M. (2005), Stepwise multiple testing as formalized data snooping. *Econometrica*, 73, 1237–82.
Schochet, P. Z. (2009), Technical methods report: the estimation of average treatment effects for clustered RCTs of education interventions. NCEE 2009-0061 rev, US Department of Education.
Schochet, P. Z. (2013), Statistical power for school-based RCTs with binary outcomes. *Journal of Research on Educational Effectiveness*, 6, 263–94.
Schulz, K. F., Altman, D. G., & Moher, D. (2010), CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. *BMJ*, 340, c332.
Teerenstra, S., Eldridge, S., Graff, M., de Hoop, E., & Borm, G. F. (2012), A simple sample size formula for analysis of covariance in cluster randomized trials. *Statistics in Medicine*, 31, 2169–78.
US Food and Drug Administration (1998), *Statistical Principles for Clinical Trials. E9*.
Wason, J. M., Stecher, L., & Mander, A. (2014), Correcting for multiple-testing in multi-arm trials: is it necessary and is it done? *Trials*, 15, 364.
Westfall, P., & Young, S. (1993), *Resampling-Based Multiple Testing? Examples and Methods for P-Value Adjustment*. Wiley.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: McConnell, B., & Vera-Hernández, M. (2025). Going beyond simple sample size calculations: a practitioner's guide. *Fiscal Studies*, 46, 323–348. https://doi.org/10.1111/1475-5890.70005