How Many Cases per Cluster? Operationalizing the Number of Units per Cluster Relative to Minimum Detectable Effects in Two-Level Cluster Randomized Evaluations with Linear Outcomes…

Article *in* American Journal of Evaluation · January 2023

DOI: 10.1177/10982140221134618

CITATIONS
0

READS
5

1 author:

Eric C. Hedberg
Abt SRBI

66 PUBLICATIONS **1,866** CITATIONS

SEE PROFILE

American Journal of Evaluation
1-16

DOI: 10.1177/10982140221134618
journals.sagepub.com/home/aje
SAGE

How Many Cases per Cluster? Operationalizing the Number of Units per Cluster Relative to Minimum Detectable Effects in Two-Level Cluster Randomized Evaluations with Linear Outcomes

E. C. Hedberg[1]

Abstract

In cluster randomized evaluations, a treatment or intervention is randomly assigned to a set of clusters each with constituent individual units of observations (e.g., student units that attend schools, which are assigned to treatment). One consideration of these designs is how many units are needed per cluster to achieve adequate statistical power. Typically, researchers state that "about 30 units per cluster" is the most that will yield benefit towards statistical precision. To avoid rules of thumb not grounded in statistical theory and practical considerations, and instead provide guidance for this question, the ratio of the minimum detectable effect size (MDES) to the larger MDES with one less unit per cluster is related to the key parameters of the cluster randomized design. Formulas for this subsequent difference effect size ratio (SDESR) at a given number of units are provided, as are formulas for finding the number of units for an assumed SDESR. In general, the point of diminishing returns occurs with smaller numbers of units for larger values of the intraclass correlation.

Keywords

Statistical power, sample size, cluster randomized designs, experiments

In cluster randomized evaluations, an intervention is randomly assigned to a subset of clusters, within which there are individual units of observations. For example, n student units each attend M schools, which are assigned to treatment or control, which together constitutes the total sample of $nM = N$. Within this design, researchers planning cluster randomized evaluations commonly: "how many units per cluster are needed for our evaluation design and assumptions?" This

[1] Abt Associates Inc USA

Corresponding Author:
E. C. Hedberg, Abt Associates, 6130 Executive Blvd Rockville, MD 20852 USA.
Email: hedbergec@gmail.com

article alters this question to ask instead: at what number of units per cluster is adding an additional unit to each cluster no longer practically beneficial? The latter question requires both an understanding of how each additional unit improves sensitivity to detect the effect of treatment and how we can measure a practical point at which the improvement in sensitivity is no longer substantively helpful to the evaluation.

For a hypothetical example of the types of decisions in planning evaluations that this article can help rationalize, suppose an evaluation of a nursing home staff anxiety prevention program is being planned. Treatment is to be assigned by nursing home and the dependent variable, an anxiety scale, will be measured for each individual licensed practical nurse (LPN). The number of nursing homes available to be randomized is fixed. It is expected that about 30 percent[1] of the variation in the outcome occurs between nursing home, and a covariate is available that explains about 25 percent of the individual-level variation and 50 percent of the between cluster variation. Please note these values are entirely hypothetical. A concern is that each of the participating nursing homes have on average ten LPNs, but only about eight LPNs per nursing home are expected to participate in the study. Increasing the effort by evaluation team members and survey incentives to add a one nurse per home would add a considerable amount to the evaluation budget. Is this expenditure worth the resources? The results below allow for a computation that indicates the sensitivity of the study would change by 2 percent if the number of LPNs per nursing home would change from eight to nine. Whether 2 percent is meaningful and worth the additional resources, depends on many contextual factors which must also be considered, of course, many of which are discussed below. This article provides computational tools to better understand these decisions from the point of view of cases per cluster, and in turn offers additional structure to evaluation planning.

I begin with reasons why consideration of units per cluster is important. I follow this with a brief overview of statistical power and minimum detectable effect sizes (MDES) for two-level cluster-randomized evaluations, and then present a formula for subsequent difference effect size ratio (SDESR), which summarizes the practical benefit of adding an additional unit to a value n to the MDES. Using derivations which relate the SDESR to the number of units per cluster (n) found in the Online Appendix, I present a formula to find a value of n at which point adding additional units is no longer practically valuable based on a value of the SDESR. I then explore the implications of the formulas and provide examples using empirically derived design parameters. Variations of these formulas are also presented when explanatory covariates are available to the evaluation team. Intermediate expressions for conceptualizing change to expected effect sizes are also included to provide additional tools.

Why Consider Units per Cluster?

Before considering these formulas, it is important to explain why or when researchers should also consider how many units per cluster are utilized in an evaluation, rather than universally promote the need to add clusters. Readers will correctly note that adding clusters will typically produce more benefit than adding units per cluster, which is true when there is any variation associated with the cluster-level (this is discussed at the end of Online Appendix). However, the discussion of how many units per cluster is still important for several reasons. First, in many situations, the number of clusters available, or recruitable, is fixed or practically constrained. For example, the number of available units per cluster may have an impact on which clusters could be used in an evaluation such as in a U.S. state with many rural schools. Assuming only large schools could be included in the evaluation would unnecessarily limit generalizability (because there would be no coverage of smaller schools, from which to generalize). Understanding the point at which larger school-sizes are no longer practically meaningful may expand school eligibility (in the minds of researchers) for evaluations. Beyond schools, cluster-randomized studies focusing on health outcomes (such as those

found at the Prevention Services Clearinghouse—preventionservices.acf.hhs.gov, see Wilson, Price, Kerns, Dastrup, & Brown, 2019) or policing outcomes (White, Mora, Orosco, & Hedberg, 2021) may involve much smaller clusters such as shifts, clinics, or even therapy groups. The inclination of many evaluators may be to find large clusters, and so the work here provides a rational mechanism to evaluate the plausible benefit of these designs and reduce concern that small clusters lead to sub-optimal power.

Second, not all studies are able to rely on administrative data for dependent variable measures. Consider writing sample scores as an example. Oftentimes, the cost of measuring this type of outcome is high and using resources to pay for scoring the writing samples for all students in a school or even an entire classroom is prohibitive. However, this work shows that for many small effects the rational number of units or cluster is less than ten per cluster, requiring only a small set of data to be collected. For additional reasons to consider the sample-size within clusters, I refer readers to Raudenbush's seminal paper on optimal design (1997), which provides other examples, culminating in the statement "Choosing the optimal within-cluster sample size is a prelude to deciding on the total number of clusters" (pg. 174). For Raudenbush, the optimal within-cluster sample size, or number of units per cluster, was related to cost functions and minimizing the sampling variance of the impact estimate. A footnote in Raudenbush (1997) foreshadowed the complexities of optimizing effect sizes, which the present article attempts to unravel.

Through this work I hope to add to the design considerations introduced by Raudenbush, further expanding the set of plausible clusters and dependent measures to support allocating resources for a broader range of studies. To clear, if there are few constraints on the number of units per cluster for an evaluation and the availability of large clusters is adequate, there is little in this paper which will offer utility for refining evaluation design decisions. If, on the other hand, constraints exist and cluster sizes are naturally small, then this paper will offer additional procedures which will improve discussions while planning cluster randomized studies.

Statistical Power and the Minimum Detectable Effect Size

Power is the chance of obtaining a statistically significant result from an evaluation based on the size of the expected estimate, population parameters, sample design, and analysis (see, e.g., Cohen, 1992). The concept of statistical power is based on the two types of statistical error. Assuming that an intervention is not efficacious, an evaluation that concludes that there is indeed an impact is making a Type I error (incorrectly stating that the means of treatment groups are different when they are, in fact, the same). This error is often noted as α in hypothesis testing, with a convention typically set at a 5 percent chance of error in tests of variance (say, in ANOVA models) or splitting a 5 percent chance of error evenly in both directions for tests of comparisons (the so-called two-tailed test for the difference between two means). This convention is associated with typical "critical" values of test statistics associated with degrees of freedom found in the tables located in the back of all reasonably useful introduction to statistics texts.

If, in fact, an intervention is efficacious, but the evaluation concludes otherwise, then a Type II error has occurred (incorrectly stating that the means of treatment groups are the same when they are, in fact, different). This error is often noted as the next Greek letter, β.[2] This error occurs in conjunction with assumed levels of Type I error, because it means the resulting evaluation produced a test statistic that does not meet the critical value set by the assumed Type I error. Power analysis focuses on understanding the sampling distribution of a future evaluation of an intervention with an assumed impact and finding the chance of a statistically significant result. As such, statistical power is the complement of Type II error, or in the notation presented here, $1 - \beta$.

Power analyses result from computing the chance that a statistical test will exceed the critical value under several assumptions that culminate in the expected test statistic, which is based on the

ratio of the mean difference and the standard error of that difference. A brief overview is provided in the Online Appendix, and more detailed summaries can be found in several texts (e.g., Hedberg, 2017b; Liu, 2013; Ryan, 2013). For the general reader, the important points are that the standard error of the mean difference typically includes functions associated with the sample size and the (assumed to be) normally distributed residual variance of the continuous dependent variable. Reorganization of these formulas often converts the mean difference into a standardized mean difference effect size, such as Cohen's d (1992) or Hedges g (1981), representing the difference between treatment groups in units of standard deviations and other non-sample-size parameters into "scale-free" parameters such as correlations and portions of variance associated with various factors. These parameters are combined to form an expected test statistic, which is used with non-central statistical distributions to find a probability of Type II error, and its complement, statistical power.

The expected test can be algebraically equated with a quantity, noted as Q below, which combines values of the standard normal or student's t distribution (with a certain degrees of freedom) based on assumed values of the Type I and Type II error. Given Q, algebraic rules can be used to isolate the key parameters of the expected test (effect size, sample size, and other scale-free parameters) to form expressions of the required sample size to achieve a specified level of power for a specified effect size, or the effect size that satisfies a specified level of power and sample size, which is the focus of this article. This effect size was introduced by Bloom (1995) as the minimum detectable effect size (MDES) and used as a method to understand the sensitivity of a given design, given a sample design and level of power. The MDES works much like letters on an eye exam chart: the better visioned (sensitive) eyes can see smaller letters. It is the MDES for cluster-randomized evaluations that is the focus of our analysis.

Properties of the Minimum Detectable Effect Size for Two-Level Cluster Randomized Evaluations

The literature on statistical power for cluster randomized studies has a long history in health (Murray, 1998) and education (Raudenbush, Martinez, & Spybrook, 2007). Bloom and colleagues further detailed the MDES for cluster randomized evaluations (Bloom, Bos, & Lee, 1999) and detailed the importance of uncorrelated covariates to improve the sensitivity of studies (Bloom, Richburg-Hayes, & Black, 2007). The MDES estimate δ_m, for a two-level cluster randomized design without covariates is a function of the total number of clusters, M, the fraction of clusters in the treatment arm, f, the intraclass correlation, ρ (detailed below), and Q, the scalar that summarizes significance and power through the sum of quantiles from the student's t -distribution using degrees of freedom based on the number of clusters and predictors,[3]

$$\text{MDES}(Q, M, f, \rho, n) = \delta_m \approx Q\sqrt{\frac{\rho}{Mf(1-f)} + \frac{1-\rho}{nMf(1-f)}}. \tag{1}$$

The introduction of covariates uncorrelated with the treatment variable to the analysis model reduces the MDES and this can be represented by adding complements of the R^2 statistics to the numerator in both fractions,

$$\text{MDES}(Q, M, f, \rho, n, R^2_{unit}, R^2_{cluster}) = \delta^*_m \approx Q\sqrt{\frac{\rho(1-R^2_2)}{Mf(1-f)} + \frac{(1-\rho)(1-R^2_1)}{nMf(1-f)}}, \tag{2}$$

where R^2_1 is the reduction in unit variance and R^2_2 is the reduction in cluster variance due to covariates. In this expression, the covariates are assumed to be uncorrelated with treatment assignment. I call this the Variance Component form because it shows two components of variation that drive the MDES:

first, the cluster component $\frac{\rho(1-R_2^2)}{Mf(1-f)}$, which includes only the number of clusters in the denominator, and second, the unit component $\frac{(1-\rho)(1-R_1^2)}{nMf(1-f)}$, which includes the total number of units, nM, in the denominator. As the number of units increases, the second component becomes smaller, which in turn lowers the MDES. Bigger samples are more sensitive.

Another parameter in the MDES is the intraclass correlation, ICC or ρ, which is a measure of natural correlation of units within the same cluster (see Hedberg, 2017a, for a brief introduction) and is also the fraction of the total variance that occurs between clusters. This parameter appears in both components. As we see below, and discussed in detail in the Online Appendix, the ICC is the key parameter for our discussion for understanding the diminishing returns. This can be seen by presenting the MDES in the Design Effect form, which is

$$\text{MDES}(Q, M, f, \rho, n) = \delta_m = Q\sqrt{\frac{1}{Mf(1-f)} \times \frac{1+\rho(n-1)}{n}}. \quad (3)$$

The introduction of covariates uncorrelated with the treatment variable to the analysis model reduces the MDES, and this can be represented by adding functions of the R^2 statistics to the numerator of the second fraction,

$$\text{MDES}(Q, M, f, \rho, n, R_1^2, R_2^2) = \delta_m^* = Q\sqrt{\frac{1}{Mf(1-f)} \times \frac{1+(n-1)\rho-(R_1^2+(nR_2^2-R_2^2)\rho)}{n}}. \quad (4)$$

This form is also important since it showcases how the ICC is an important factor in power and MDES analyses, as the factor $1+\rho(n-1)$ is the typical design effect for cluster samples long noted in the survey and experimental literature (e.g., Hedges & Hedberg, 2007; Kish, 1965); covariates reduce the design effect by $R_1^2+(nR_2^2-R_2^2)\rho$. This formula highlights the result that the sampling variance of the clustered sample mean has a different structure than that of the simple random sample mean. This means that the sampling variance of an impact estimate from a clustered sample is $1+\rho(n-1)$ times larger than the sampling variance incorrectly estimated if a simple random sample is assumed. In turn, the MDES without covariates increases by a factor of $\sqrt{1+\rho(n-1)}$ for a clustered sample with the same total number of observations, $nM = N$, as a simple random sample. Because this expression includes the product of n and ρ, the intraclass correlation is a key parameter in finding the best number of units required per cluster.

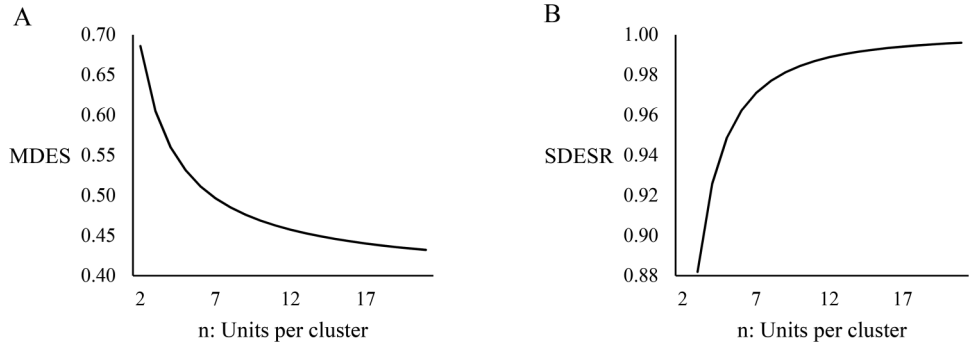

Figure 1. The values of the minimum detectable effect size (MDES, A) and subsequent difference effect size ratios (SDESR, B) by number of units per cluster for 20 clusters in treatment, 20 in control, an ICC of .2, and no covariates, for a two-tailed test ($\alpha = .05$) and statistical power of .8.

Relationship Between the Number of Units Per Cluster (n) and the Minimum Detectable Effect Size for Two-level Cluster Randomized Evaluations

As has been noted, larger values of n lead to smaller values of the MDES. However, this relationship is somewhat complex. Figure 1 presents the MDES (A) and the subsequent difference effect size ratio (SDESR, B) for a cluster randomized design with 20 clusters in treatment and 20 in control with an ICC of .2 for a variety of values for n, the number of units per cluster. If we hold the parameters Q, M, f, ρ, and the R^2 s constant, the MDES decreases as the number of units per cluster increases (see Figure 1 A). However, the change in the MDES is not linear, as each additional unit has a smaller impact on the MDES. We can conceptualize this change with each additional unit as the SDESR, θ,

$$\text{SDESR} = \theta = \frac{\delta_{m|n+1}}{\delta_{m|n}}. \tag{5}$$

For small numbers of units per cluster, the SDESR is noticeably less than 1, and so the benefits to adding units are meaningful. In Figure 1 A, the MDES for this design with two units per cluster is about .685, and the MDES for three units is .605, and the SDESR of .605 to .685 is $\theta = 0.882$ (see Figure 1 B). However, when $n = 10$, the MDES is .469 and when $n = 11$, the MDES is .462, which is a difference in the third decimal and has a SDESR closer to 1 of $\theta = 0.984$. In Figure 1 B you may notice the curve approaching 1 as the number of units increases. Moreover, at the value of $n = 13$, the SDESR for each subsequent value is greater than .99, and beyond what is reported in the figures, the MDES does not result in a value below .4 until $n = 197$, at which point the SDESR is greater than .999.

The point of the preceding paragraph is to illustrate the diminishing returns in sensitivity for adding units. For example, in many cases the practical difference between an MDES of .445 ($n = 15$) and .422 ($n = 30$) may not be worth doubling the sample in cases for which the fixed cost of unit-level measurement is non-trivial. However, offering broad rules of thumb has a mixed history in statistical consultation (Aguinis & Harden, 2010). Instead, I seek to provide some guidance on finding the point of diminishing returns and find which parameters are most influential on that point.

The SDESR metric, however, is still quite abstract and lacks a basic intuition to allow researchers working in the field to find it useful. To that end, suppose that a minimum absolute change (γ), which most researchers using standardized difference effect sizes will find meaningful, is 1 percent of a standard deviation ($\gamma = .01$). Next, suppose a range of effect size benchmarks for which researchers are considering is a tenth of a standard deviation ($\delta_b = .1$) to one and a half standard deviations ($\delta_b = 1.5$). Values of the Benchmarked SDESR ($\hat{\theta}$) corresponding to the minimum absolute change (γ) for an effect size benchmark (δ_b) can be computed with

$$\hat{\theta}(\delta_b, \gamma) = \frac{\delta_b - \gamma}{\delta_b}. \tag{6}$$

Which, for $\delta_b = .1$ and $\gamma = .01$ is $\hat{\theta} = 0.9$, for $\delta_b = .5$ and $\gamma = .01$ is $\hat{\theta} = 0.98$, $\delta_b = 1$ and $\gamma = .01$ is $\hat{\theta} = 0.99$, and $\delta_b = 1.5$ and $\gamma = .01$ is $\hat{\theta} = 0.9933$. Given a range of effect size benchmarks, such as those in Hill, Bloom, Black, and Lipsey (2008) for academic growth, researchers can find a value for the SDESR that represents the minimum practical change to consider. Of course, the value of this change (γ) must also be selected, and a one percent of the standard deviation change is presented as only a practical example.

Relating Subsequent Difference in Effect Size Ratios to the Intraclass Correlation

The goal of this work is to find an answer to this question: at what number of units per cluster is adding additional units no longer practically beneficial, where "no longer practically beneficial" is based on a high SDESR with a value close to 1. To do this, we must formulate a relation between change in the MDES (operationalized as the SDESR) to the units per cluster (n). In the previous section, I described a practical computation of a Benchmarked SDESR that can be based on a reference effect size and a small change. In this section, the focus will be on the relation of this result to the number of units per cluster (n). I do this by first showing the SDESR as a function of n, then showing how n is a function of the SDESR. The next section will bring these two ideas together.

SDESR as a Function of n *and the ICC*

In the previous section, I established the SDESR (θ) as a measure of change to the MDES. In the absence of covariates, the following expression provides a link between the ratio θ, n, and the ICC (ρ)

$$\text{SDESR}(\rho, n) = \theta = \exp\left[\frac{\rho - 1}{2n(1 + (n-1)\rho)}\right]. \tag{7}$$

Details of how this and the other formulas that are introduced here are derived appear in the Online Appendix. Expression (7) allows the researcher to calculate the SDESR as a function of a given number of units per cluster, n, and the intraclass correlation, ρ. Note that this formula does not involve the number of clusters, which allows for the discussion of the benefits of units per cluster to be based on a single a priori expectation of the ICC. This implies that adding or removing clusters will not change the proportional benefit of adding units per cluster for the MDES.

For example, suppose an ICC of $\rho = .2$. The MDES will reduce by roughly 2 percent ($\theta = .98$) when adding a unit to $n = 10$ units resulting in $n = 11$ units

$$\text{SDESR} = \theta = \exp\left[\frac{.2 - 1}{2 \times 10 \times (1 + (10-1) \times .2)}\right] \approx 0.98.$$

However, if the ICC is $\rho = .1$, the impact of adding a unit to $n = 10$ units is larger, the MDES will reduce by roughly 5 percent ($\theta = .95$),

$$\text{SDESR} = \theta = \exp\left[\frac{.1 - 1}{2 \times 10 \times (1 + (10-1) \times .1)}\right] \approx 0.95,$$

which illustrates that adding units is typically more beneficial when ICCs are lower.

In the case of covariates, these expressions have two additional parameters. The first is the proportion reduction in variance at the unit level, noted as R_1^2, and the other parameter is the proportion reduction in variance at the cluster level, noted as R_2^2. The expression for θ in the presence of these factors is

$$SDESR(\rho, n, R_1^2, R_2^2) = \theta^* = \exp\left[\frac{(\rho - 1)(1 - R_1^2)}{2n(1 + (n-1)\rho - (R_1^2 + (nR_2^2 - R_1^2)\rho))}\right] \tag{8}$$

For example, suppose, again, an ICC of $\rho = .2$ and additionally a covariate that explains 25 percent of unit variation ($R_1^2 = 0.25$) and 15 percent of cluster variation ($R_2^2 = 0.15$). The MDES will reduce by roughly 4 percent ($\theta = .96$) when adding a unit to $n = 5$ units resulting in $n = 6$ units

$$SDESR = \theta^* = \exp\left[\frac{(.2 - 1) \times (1 - .25)}{2 \times 5 \times (1 + (5-1) \times .2 - (.25 + (5 \times .15 - .25) \times .2))}\right] \approx 0.96.$$

However, if the unit variation explained by the covariate is $R_1^2 = 0.5$, the impact of adding a unit to $n = 5$ units is smaller, the MDES will reduce by roughly 3 percent ($\theta = .97$),

$$SDESR = \theta^* = \exp\left[\frac{(.2-1)\times(1-.5)}{2\times 5\times(1+(5-1)\times .2-(.5+(5\times .15-.5)\times .2))}\right] \approx 0.97,$$

which illustrates that adding units is typically less beneficial when impacts of unit-level covariates are larger. The benefits of adding units are also smaller (the SDESR is closer to 1) with larger cluster-level covariate impacts for any value of n.

Units per Cluster (n) as a Function of the SDESR and the ICC (or, Finding the Point of Diminishing Returns)

Expressions (7) and (8) can be solved for n to find the function of the ICC and θ (and the R^2 s in the case of covariates) to allow researchers to pre-specify the ratio of change and find the point at which increasing the number of units per cluster is no longer practical. This expression for the point of diminishing returns for units (PDRn) involves the log of the SDESR, $\ln(\theta)$, and the ICC, ρ, without covariates

$$\text{PDRn} = n = \frac{(\rho-1)\ln(\theta) - \sqrt{(\rho-1)\ln(\theta)(2\rho+\ln(\theta)(\rho-1))}}{2\rho\ln(\theta)}, \tag{9}$$

and additional terms for the R^2 s in the case of covariates

$$\text{PDRn}^* = n^* = \frac{(\rho-1)\ln(\theta)(1-R_1^2) - \sqrt{(\rho-1)\ln(\theta)(R_1^2-1)(2\rho(R_2^2-1)+\ln(\theta)(\rho-1)(R_1^2-1))}}{2\rho\ln(\theta)(1-R_2^2)}. \tag{10}$$

For example, suppose researchers decide that an SDESR of $\theta = .9999$ is the maximum change in the MDES of interest. With an ICC of $\rho = .2$, the value of PDRn that will satisfy this criterion is

$$\text{PDRn} = \frac{(.2-1)\times\ln(.9999) - \sqrt{(.2-1)\times\ln(.9999)\times(2\times .2+\ln(.9999)\times(.2-1))}}{2\times .2\times\ln(.9999)} \approx 139\ .$$

However, if the SDESR is $\theta = .99$, the value of PDRn is

$$\text{PDRn} = \frac{(.2-1)\times\ln(.99) - \sqrt{(.2-1)\times\ln(.99)\times(2\times .2+\ln(.99)\times(.2-1))}}{2\times .2\times\ln(.99)} \approx 12.$$

which illustrates that more units are required as $\theta \to 1$, since it required 12 units for $\theta = .99$ ($1-\theta = .01$), but over ten-fold more units, 139, for $\theta = .9999$ ($1-\theta = .0001$, 100 times closer to 1).

In the case of covariates, these expressions also include R_1^2 and R_2^2. If the value of $R_1^2 = .5$ and the value of $R_2^2 = .15$, and an ICC of .2, the value of $PDRn^*$ for an SDESR of .99 is

$$\text{PDRn}^* = \frac{(.2-1)\times\ln(.99)\times(1-.5) - \sqrt{(.2-1)\times\ln(.99)\times(.5-1)\times(2\times .2\times(.15-1)+\ln(.99)\times(.2-1)\times(.5-1))}}{2\times .2\times\ln(.99)\times(1-.15)}$$

$$\approx 10,$$

which illustrates how effective covariates can reduce the PDRn, and in turn, the necessary units for an evaluation.

Table 1. Values of the SDESR (θ) by n and the ICC (ρ) Without Covariates.

	ρ									
n	.01	.02	.03	.04	.05	.1	.15	.2	.25	.3
5	.9092	.9133	.9170	.9206	.9239	.9377	.9483	.9565	.9632	.9687
10	.9556	.9593	.9625	.9653	.9678	.9766	.9821	.9858	.9885	.9906
15	.9715	.9748	.9775	.9797	.9815	.9876	.9909	.9930	.9945	.9955
20	.9794	.9824	.9847	.9865	.9879	.9923	.9945	.9958	.9967	.9974
25	.9842	.9868	.9888	.9903	.9914	.9947	.9963	.9972	.9979	.9983
30	.9873	.9897	.9914	.9926	.9936	.9962	.9974	.9980	.9985	.9988
35	.9895	.9917	.9932	.9942	.9950	.9971	.9980	.9985	.9989	.9991
40	.9911	.9931	.9944	.9953	.9960	.9977	.9985	.9989	.9991	.9993

Table 2. Values of the SDESR (θ) by n and the ICC (ρ) with Covariates Where $R_1^2 = .5$ and $R_2^2 = .25$.

	ρ									
n	.01	.02	.03	.04	.05	.1	.15	.2	.25	.3
5	.9112	.9169	.9220	.9266	.9308	.9469	.9579	.9658	.9718	.9766
10	.9575	.9624	.9664	.9697	.9724	.9814	.9864	.9895	.9917	.9933
15	.9732	.9774	.9805	.9829	.9849	.9905	.9933	.9950	.9961	.9969
20	.9810	.9846	.9871	.9890	.9904	.9942	.9960	.9971	.9977	.9982
25	.9856	.9887	.9908	.9922	.9933	.9961	.9974	.9981	.9985	.9988
30	.9886	.9913	.9931	.9942	.9951	.9972	.9981	.9986	.9990	.9992
35	.9907	.9931	.9946	.9955	.9962	.9979	.9986	.9990	.9992	.9994
40	.9922	.9944	.9956	.9964	.9970	.9984	.9989	.9992	.9994	.9995

Illustrations and Intuitions

Tables 1 through 4 offer further illustrations of these results. These tables were produced in R (R Core Team, 2021) using functions detailed in the Online Appendix. Table 1 presents values for the SDESR, θ, for a variety of values for n and the ICC (ρ). For low values of the ICC such as .01, the benefit of adding to smaller values of n, such as 5, result is nearly ten percent reductions in the MDES, whereas by 40 units, the gains are less than one percent. With larger ICCs such as .25, however, benefits drop below a 1 percent change by 15 units and below a tenth of a percent for 40 units. As seen in Table 2, employing typical covariate values such as $R_1^2 = .5$ and $R_2^2 = .25$, promotes the reduction in benefits at even smaller values of n.

Table 3 presents rounded integer values of PDRn for the same set of ICCs, but for various values of the Benchmarked SDESR ($\hat{\theta}$) that represent change of a percent of a standard deviation ($\gamma = .01$) for benchmark effect sizes ranging from .1 to 1.5. The most striking aspect of this table is the number of small PDRn values, with a .01 standard deviation change associated with a benchmark effect size of .2 or less associated with values of PDRn lower than 10 units for even small ICCs (.01) or higher ICCs such as .25. Table 4 presents PDRn values for even smaller changes to the MDES, half of a standard deviation percent ($\gamma = .005$), which results in values that are typically half to twice as large depending on the value of the ICC (some smaller values are about a third as large). Thus, the smaller the change to the effect size, the larger the PDRn value will be.

Tables 5 and 6, employing typical covariate values such as $R_1^2 = .5$ and $R_2^2 = .25$, has even more small single digit integers, with changes relative to Table 3 that are proportionally larger for larger

Table 3. Values of n by Benchmarked SDESR Values ($\gamma = .01$) and the ICC (ρ) Without Covariates.

				ρ									
$\hat{\theta} = \frac{\delta - .01}{\delta}$				.01	.02	.03	.04	.05	.1	.15	.2	.25	.3
$\delta_b =$	.1	$\hat{\theta} =$	.9000	5	4	4	4	4	3	3	3	3	2
	.2		.9500	9	8	8	7	7	6	5	5	4	4
	.3		.9667	13	12	11	10	10	8	7	6	5	5
	.4		.9750	17	15	14	13	12	10	8	7	6	6
	.5		.9800	21	18	16	15	14	11	9	8	7	7
	.6		.9833	24	21	19	17	16	12	10	9	8	7
	.7		.9857	27	23	21	19	18	14	11	10	9	8
	.8		.9875	30	26	23	21	20	15	12	11	10	9
	.9		.9889	33	28	25	23	21	16	13	12	10	9
	1.0		.9900	36	31	27	25	23	17	14	12	11	10
	1.1		.9909	39	33	29	26	24	18	15	13	11	10
	1.2		.9917	42	35	31	28	26	19	16	14	12	11
	1.3		.9923	45	37	32	29	27	20	17	14	13	11
	1.4		.9929	47	39	34	31	28	21	17	15	13	12
	1.5		.9933	50	41	36	32	29	22	18	15	14	12

Table 4. Values of n by Benchmarked SDESR Values ($\gamma = .005$) and the ICC (ρ) Without Covariates.

				ρ									
$\hat{\theta} = \frac{\delta - .005}{\delta}$				.01	.02	.03	.04	.05	.1	.15	.2	.25	.3
$\delta_b =$	.1	$\hat{\theta} =$	.9500	9	8	8	7	7	6	5	5	4	4
	.2		.9750	17	15	14	13	12	10	8	7	6	6
	.3		.9833	24	21	19	17	16	12	10	9	8	7
	.4		.9875	30	26	23	21	20	15	12	11	10	9
	.5		.9900	36	31	27	25	23	17	14	12	11	10
	.6		.9917	42	35	31	28	26	19	16	14	12	11
	.7		.9929	47	39	34	31	28	21	17	15	13	12
	.8		.9938	52	43	37	33	31	23	19	16	14	13
	.9		.9944	57	46	40	36	33	24	20	17	15	13
	1.0		.9950	62	50	43	38	35	26	21	18	16	14
	1.1		.9955	66	53	46	41	37	27	22	19	17	15
	1.2		.9958	70	56	48	43	39	29	23	20	18	16
	1.3		.9962	74	59	51	45	41	30	24	21	18	16
	1.4		.9964	78	62	53	47	43	31	25	22	19	17
	1.5		.9967	82	65	55	49	45	32	26	23	20	18

values of the ICC. Comparing Table 5 to Table 3, the covariate impacts reduce the PDRn values by at most a third, and for small effect sizes, generally not at all. Comparing Table 6 to Table 5, the increase in the PDRn values is similar to the tables without covariates, with increases ranging from a quarter to 2.25 fold.

Across Tables 3–6 is a pattern where larger benchmark effect sizes have higher PDRn values for the same absolute change (a percent of a standard deviation), as these ratios represent smaller and smaller differences from the benchmark. This is congruent with patterns found in the power analysis of other ratios—such as odds ratios in logistic regression—where additional data is required for

Table 5. Values of n^* by Benchmarked SDESR Values ($\gamma = .01$) and the ICC (ρ) with Covariates Where $R_1^2 = .5$ and $R_2^2 = .25$.

				ρ									
$\hat{\theta} = \frac{\delta - .01}{\delta}$				.01	.02	.03	.04	.05	.1	.15	.2	.25	.3
$\delta_b =$	.1	$\hat{\theta} =$	.9000	4	4	4	4	4	3	3	2	2	2
	.2		.9500	9	8	7	7	6	5	4	4	4	3
	.3		.9667	12	11	10	9	9	7	6	5	5	4
	.4		.9750	16	14	13	11	11	8	7	6	5	5
	.5		.9800	19	16	15	13	12	10	8	7	6	5
	.6		.9833	22	19	17	15	14	11	9	8	7	6
	.7		.9857	25	21	19	17	16	12	10	8	7	7
	.8		.9875	28	23	20	18	17	13	11	9	8	7
	.9		.9889	31	25	22	20	18	14	11	10	9	8
	1.0		.9900	33	27	24	21	20	15	12	10	9	8
	1.1		.9909	36	29	25	23	21	15	13	11	10	8
	1.2		.9917	38	31	27	24	22	16	13	11	10	9
	1.3		.9923	40	32	28	25	23	17	14	12	10	9
	1.4		.9929	42	34	29	26	24	18	14	12	11	10
	1.5		.9933	45	36	31	27	25	18	15	13	11	10

Table 6. Values of n^* by Benchmarked SDESR values ($\gamma = .005$) and the ICC (ρ) with covariates where $R_1^2 = .5$ and $R_2^2 = .25$.

				ρ									
$\hat{\theta} = \frac{\delta - .005}{\delta}$				.01	.02	.03	.04	.05	.1	.15	.2	.25	.3
$\delta_b =$	.1	$\hat{\theta} =$	.9500	9	8	7	7	6	5	4	4	4	3
	.2		.9750	16	14	13	11	11	8	7	6	5	5
	.3		.9833	22	19	17	15	14	11	9	8	7	6
	.4		.9875	28	23	20	18	17	13	11	9	8	7
	.5		.9900	33	27	24	21	20	15	12	10	9	8
	.6		.9917	38	31	27	24	22	16	13	11	10	9
	.7		.9929	42	34	29	26	24	18	14	12	11	10
	.8		.9938	47	37	32	29	26	19	16	13	12	10
	.9		.9944	51	40	35	31	28	20	17	14	12	11
	1.0		.9950	55	43	37	33	30	22	18	15	13	12
	1.1		.9955	58	46	39	35	31	23	19	16	14	12
	1.2		.9958	62	48	41	36	33	24	19	17	15	13
	1.3		.9962	65	51	43	38	35	25	20	17	15	13
	1.4		.9964	69	53	45	40	36	26	21	18	16	14
	1.5		.9967	72	55	47	42	38	27	22	19	16	15

similar changes to extreme base rates (i.e., base rates near 0 or 1) relative to base rates near .5 (see, e.g., Demidenko, 2007).

Returning to the hypothetical nursing home example that started the article, the 30 percent variation in the outcome between nursing homes represents an ICC of .3, and the covariate effects of 25 percent explained at the LPN level is $R_1^2 = .25$ and 50 percent between nursing homes is $R_2^2 = .5$, which for $n = 8$ LPNs produces an SDESR of about .981, or about a 2 percent change. If there was an expected effect size of, say, .4 standard deviations and evaluators wanted a value of n that was within $\gamma = .01$ standard deviations ($\hat{\theta} = .975$), the formula for the PDRn would yield about 7

LPNs (rounded from 6.75 LPNs). This would indicate the optimal sample for the expected effect size would be optimal with 7, not 8 LPNs. Again, please note these parameters are entirely hypothetical and the results of this analysis would be quite different with different values of the expected effect size, ICC, and R^2 statistics.

Proposed Procedure

When deciding the number of units that indicate a reasonable point of diminishing returns, I offer the following suggestion for a sequence of computations during the planning of an evaluation. First, prior to any power analyses, use experience, literature, and benchmarks to select a reasonable expected effect size, δ_b. This should be done regardless of whether the number of units per cluster is a study design consideration. Next, assume either value of the SDESR, θ, based only on proportional change, such as a 1 percent change for SDESR = .99 or a percent of a percent change for an SDESR = .9999. Alternatively, compute a benchmarked SDESR, $\hat{\theta}$, using a expected effect size and a reasonably small change to the effect size in standard units, γ, using expression (6). The final set of parameters to select include defensible and justifiable values for the ICC and covariate R^2 values (if applicable). Generally, SDESR values closer to 1 should be used if the expected effect size is small.

Next, use expression (10) to compute the number of units which represents the point of diminishing returns for the benchmark effect size. Note that with this procedure, as with any power analysis, be sure to increase this value based on expected attrition because this represents the final sample size, not the initially sampled set. At this point researchers also have all the necessary information to compute the number of clusters required for the benchmark effect size, and the supplemental material includes R code for these functions as they can be tedious. Different values of the benchmarked SDESR, as a function of the amount of change to the MDES, γ, will yield different values for n and thus number of clusters. If the available clusters are fixed, then the discussion is focused only on the number of cases per cluster. If the available clusters are negotiable, then this process in conjunction with other optimal design formulas from Raudenbush (1997) can be helpful in determining the best design among several options. In the next section, the results of this exercise are explored based on empirical parameters.

Examples Based on Empirical Work and Other Assumptions

I offer the assumption in this article that the use of "typical" values in place of informed assumptions in planning studies is a counterproductive practice, whether it be for required units per cluster, effect sizes to expect, or even ICC values. Taking this at face value, I then move in this section to showcase how the formulas presented here can inform the planning of studies under various empirically informed scenarios. Suppose researchers are planning an early childhood education evaluation to evaluate an intervention that seeks to increase math scores for third grade students. Data from Hill and colleagues (2008, see Table 5) indicate that the typical impacts from academic intervention studies that they reviewed was a quarter standard deviation (.25). The range of ICCs across geographics and locales in the United States varies widely. For example, in small districts with 3 to 5 schools serving elementary grades, the school-level ICCs tend to be .05 or less, with ICCs of .1 only appearing in larger school districts with 10 schools serving each grade (see Tables 2 and 3 in Hedberg & Hedges, 2014). Across states, the school-level ICCs (without considering district effects) also vary widely with subject and grade, with third grade ICCs for Mathematics scores as high as .24 in Massachusetts, .23 in Colorado and as low as .05 in West Virginia (see Table 2 in Hedges & Hedberg, 2013). Given this empirical evidence, I offer the results of the following exercise.

Suppose four scenarios for planning research, comprising either expected impacts of .25 or .5 standard deviations in populations with ICCs of either .1 or .2. Next, suppose five power analysis

Table 7. Sample Sizes all Meeting Power of .8 for a two-Tailed Test ($\alpha = .05$) Using Different Assumptions by Benchmark Effect Size (δ_b) and ICC (ρ) with Covariates Where $R_1^2 = .5$ and $R_2^2 = .25$.

	ρ					
	.10			.20		
δ	n	M	N	n	M	N
A) Finding the PDRn with $\hat{\theta} = \frac{\delta_b - .01}{\delta_b}$						
.25, $\hat{\theta} = .96$	7	74	518	5	118	590
.50, $\hat{\theta} = .98$	10	18	180	7	30	210
B) Finding the PDRn with $\hat{\theta} = \frac{\delta_b - .005}{\delta_b}$						
.25, $\hat{\theta} = .98$	10	64	640	7	108	756
.50, $\hat{\theta} = .99$	15	16	240	11	26	286
C) Finding the PDRn with $\theta = .999$ across all effect sizes						
.25	52	46	2,392	36	84	3,024
.50	52	14	728	36	24	864
D) Setting $n = 30$						
.25	30	48	1,440	30	86	2,580
.50	30	16	480	30	24	720
E) Setting $n = 60$						
.25	60	44	2,640	60	82	4,920
.50	60	14	840	60	24	1,440

strategies are employed to first find the optimal value of n and then compute the total number of schools with equal allocation to treatment and control: (A) finding the PDRn with $\hat{\theta} = \frac{\delta_b - .01}{\delta_b}$, (B) finding the PDRn with $\hat{\theta} = \frac{\delta_b - .005}{\delta_b}$, (C) Finding the PDRn with $\theta = .999$ across all effect sizes, (D) simply setting $n = 30$, and (E) simply setting $n = 60$. We can judge each scenario by differences in the total sample and differences in the number of clusters required. The method to find the number of clusters required, given a value for n is detailed in Hedberg (2017b). Note that all values of estimated values of n were rounded up and M were rounded up to the nearest even number.

The results of these exercises appear in Table 7, which presents sample sizes that all meet power of .8 for a two-tailed test with the same covariate effectiveness for the respective effect size and ICC values. The first scenario, PDRn with $\hat{\theta} = \frac{\delta_b - .01}{\delta_b}$, produced the smallest overall sample size but required the largest number of clusters and produced the smallest overall sample requirements. The second scenario required slightly higher observations per cluster, fewer clusters, but larger overall sample sizes. Setting the SDESR to .999 produces higher PDRn values, again requires fewer clusters, but also translates into larger overall samples. These results are most like the typical assumptions of $n = 30$ and $n = 60$, which require similar numbers of clusters. These patterns are similar for both ICC values of .1 and .2.

For example, for an effect size of .25 with an ICC of .1, the PDRn is 7 using the first strategy (A) but requires 74 total schools to achieve power of .8. The second strategy (B) increased the PDRn to 10 and lowered the required clusters to 64. The third strategy, (C), produced a much higher PDRn (52) and reduced the number of clusters even more. However, each successive increase in the SDESR increases n, lowers M, but ultimately produces larger samples. The required numbers of clusters are similar for scenarios (D) and (E), which assumed round values of n.

From this exercise, a major takeaway point is that there is a wide variety of situations and scenarios, even with a small selection of empirical settings. As a consequence, the entire prospect of rules of thumb about sample sizes within clusters is rendered inadequate. Instead, rather than present exact answers, this article provides tools and operationalization of the key considerations that can lead

researchers to answers which apply to their studies. The antidote to rules of thumb are tools, which are presented here.

Conclusion

In teaching power analyses for cluster randomized designs, most instructors (including this author) will often note in passing that many different combinations of n and M will yield the same chance of detecting an effect size. Table 7 provides a clear example of this phenomena through a careful consideration of a researcher-controlled parameter of what it means to have diminishing returns, the SDESR. The SDESR can itself be tuned either with a broad threshold (such as .999) or based on changes to a benchmark effect size.

I provide a method to assess how many units are practically beneficial by providing researchers a metric of "beneficial" and employing this metric in a formula to estimate the number of units per cluster. In general, the point of diminishing returns occurs with smaller numbers of units for larger values of the ICC. This is intuitive, as the very design effect that reduces precision in cluster randomized evaluations includes the multiplication of units per cluster (n) and the ICC, ρ. However, with these formulas, intuition is effectively operationalized.

These results hopefully will help researchers avoid broad rules of thumb about one of the important choices in designing cluster randomized evaluations: the number of units per cluster. In my own experience, including being guilty of advising this, the general advice offered is that after 30 units, it does not make sense to continue to add units. As I stated in the first sections of this article, if clusters available to a given evaluation are large and plentiful and the cost of each unit observation are negligible, there is little here which will greatly impact evaluation designs. However, if clusters are smaller, then ICCs are higher, and this work can shed light on answering questions about "how many units do we really need?"

Finally, these results provide evidence to reject rules of thumb for sample sizes. As shown above, the required sample configuration is entirely dependent on various design parameters and on researcher defined goals. This is at the core of most statistical analysis. The ordinary least squares (OLS) regression equations are best for a given criteria, minimizing the total sum of squared deviations between the observations and the prediction. Should regressions need to meet other criteria, such as predicting the best median or more recent algorithms employed by data science researchers, then OLS regression is no longer "best." In this article, I provide expressions for finding values of units per cluster based on the concept of diminishing returns.

Acknowledgments

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305D200045, Abt Associates. The opinions expressed are those of the author and does not necessarily represent the views of the Institute or the U.S. Department of Education. The author thanks Cris Price and Kristen Neishi for helpful comments on earlier drafts, and the constructive feedback of the anonymous reviewers, which markedly improved the manuscript.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Department of Education, (grant number R305D200045).

ORCID iD

E. C. Hedberg https://orcid.org/0000-0003-0679-0720

Supplemental Material

Supplemental material for this article is available online.

Notes

1. This seems to a very high estimate of between-nursing home variation, but similar values for other outcomes have been documented, see, e.g., Min, Park, & Scott (2016).
2. This notation sometimes sows some confusion with unrelated statistics such as regression slopes or standardized regression coefficients as labeled in some software.
3. See Online Appendix for details on derivation of the MDES.

References

Aguinis, H., & Harden, E. E. (2010). Sample size rules of thumb: Evaluating three common practices. In *Statistical and Methodological Myths and Urban Legends* (pp. 267–360). New York, New York: Routledge.

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

Chapter 4

Bloom, H. S., Bos, J. M., & Lee, S. W. (1999). Using cluster random assignment to measure program impacts: Statistical implications for the evaluation of education programs. *Evaluation Review*, *23*(4), 445–469.

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

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155–159.

Demidenko, E. (2007). Sample size determination for logistic regression revisited. *Statistics in Medicine*, *26*, 3385–3397.

Hedberg, E. C., & Hedges, L. V. (2014). Reference values of within-district intraclass correlations of academic achievement by district characteristics: Results from a meta-analysis of district-specific values. *Evaluation Review*, *38*(10), 546–582.

Hedberg, E. C. (2017b). *Introduction to power analysis: two-group studies* (Vol. 176). New York, New York: Sage Publications.

Hedberg, E. C. (2017a). Intraclass Correlations. In M. Allen (Ed.), *The sage encyclopedia of communication research methods (Vols. 1-4)*. Thousand Oaks, CA: Sage Publications, Inc.

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, *6*(2), 107–128.

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

Hedges, L. V., & Hedberg, E. C. (2013). Intraclass correlations and covariate outcome correlations for planning two-and three-level cluster-randomized experiments in education. *Evaluation Review*, *37*(10), 445–489.

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives*, *2*(3), 172–177.

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

Liu, X. S. (2013). *Statistical Power Analysis for the Social and Behavioral Sciences: Basic and Advanced Techniques*. New York, New York: Routledge.

Min, A., Park, C. G., & Scott, L. D. (2016). Evaluating technical efficiency of nursing care using data envelopment analysis and multilevel modeling. *Western Journal of Nursing Research*, *38*(11), 1489–1508.

Murray, D. M. (1998). *Design and Analysis of Group-Randomized Trials* (Vol. 29). USA: Oxford University Press.

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

Raudenbush, S. W., Martinez, A., & Spybrook, J. (2007). Strategies for improving precision in group-randomized experiments. *Educational Evaluation and Policy Analysis*, *29*(1), 5–29.

R Core Team (2021). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing, URL https://www.R-project.org/.

Ryan, T. P. (2013). *Sample Size Determination and Power*. New York, New York: John Wiley & Sons.

White, M. D., Mora, V. J., Orosco, C., & Hedberg, E. C. (2021). Moving the needle: can training alter officer perceptions and use of de-escalation? *Policing: An International Journal*, *44*(43), 418–436.

Wilson, S. J., Price, C. S., Kerns, S. E. U., Dastrup, S. D., & Brown, S. R. (2019). *Title IV-E prevention services clearinghouse handbook of standards and procedures, version 1.0, OPRE Report # 2019–56*. Washington, DC: Office of Planning, Research, and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

Appendix A: Review of the Minimum Detectable Effect Size and the Logic of the Derivations

The impact of an intervention in a randomized experiment is defined as $\Delta = \mu_T - \mu_C$, or the differences in the means (μ) of the treatment (T) and comparison/control (C) groups. To understand the MDES formula, we first must understand that the total amount of variation in a dependent variable in a clustered sample, σ_T^2, is the sum of two sources of variation. The first source is the variation between cluster means, often noted as "level-2" variation with the symbol σ_2^2. Within each cluster, we assume a constant amount of variation within clusters, often noted as "level-1" variation with the symbol σ_1^2. In clustered samples, Raudenbush (1997, see also Bloom et al. 1999; Bloom, 2005) noted that the standard error of the impact, Δ, of a cluster randomized trials (assuming each cluster was equally sized with n units) is

$$SE(\Delta) = \sqrt{\frac{\sigma_2^2}{Mf(1-f)} + \frac{\sigma_1^2}{nMf(1-f)}},$$

where M is the total number of clusters, f is the fraction of the total clusters assigned to the treatment group, and n is the number of units from each cluster included in the analysis (which we assume is the same for each cluster).

Analysis of variance (ANOVA) lessons in most introductory statistics classes state that the total sum of squares (SST) of a variable can be decomposed by a group factor variable into the between-group sum of squares (SSB) and within-group sum of squares (SSW), to form the identity $SST = SSB + SSW$. Variance in outcomes in multi-level models has a similar identity, where the total variation of outcomes in clustered samples is the sum of two variance components (Searl et al. 2009): the level-2 and level-1 variation, $\sigma_T^2 = \sigma_2^2 + \sigma_1^2$. This identity allows for two conversions. The first conversion is for the impact, Δ, which can be converted in a Cohen's d-like effect size, δ, by dividing the observed differences between the averages of the

treatment groups by the estimate of the total population standard deviation (which is the square root of the sum of leve-2 and level-1 variation, see Hedges 2007)

$$\delta = \frac{\Delta}{\sigma_T} = \frac{\Delta}{\sqrt{\sigma_2^2+\sigma_1^2}}.$$

The second conversion follows from this. By standardizing the impact estimate, we must also standardize the separate variances for level-2 and level-1 in the standard error formula by dividing each variance component by the total variance. The standardized form of the level-2 variance component leads to the intraclass correlation (ICC),

$$\rho = \frac{\sigma_2^2}{\sigma_T^2},$$

and the standardized form of the level-1 variance component is the complement of the ICC,

$$1 - \rho = \frac{\sigma_1^2}{\sigma_T^2},$$

leading the identity that $\rho + (1 - \rho) = 1$.

As a result, the standard error of the effect size replaces σ_2^2 with ρ and σ_1^2 with $1 - \rho$ since $\sigma_T^2 = \sigma_2^2 + \sigma_1^2$,

$$SE(\delta) = \sqrt{\frac{\rho}{Mf(1-f)} + \frac{1-\rho}{nMf(1-f)}}.$$

The expected test statistic in standardized form is then

$$\lambda = \frac{\delta}{\sqrt{\frac{\rho}{Mf(1-f)}+\frac{1-\rho}{nMf(1-f)}}}.$$

The addition of uncorrelated covariates reduces residual variance at level-2 by a factor of $1 - R_2^2$ and at level-1 by a factor of $1 - R_1^2$, where R^2 is the portion of variance explained at each level by the covariates. This renders the test when covariances are used

$$\lambda = \frac{\delta}{\sqrt{\frac{\rho(1-R_2^2)}{Mf(1-f)}+\frac{(1-\rho)(1-R_1^2)}{nMf(1-f)}}}.$$

In power analysis, expected test, λ, can be associated with a function based on quantiles from the statistical test distribution, $\lambda \approx Q$ (Ryan 2013). The value of Q is equal to the difference between two quantiles of the t-distribution based on $M - 2 - p$, degrees of freedom (df), where p is the number of cluster level covariates, $Q = H\left[q_{1-\frac{\alpha}{2}}, M - 2 - p\right] - H[q_{1-\text{Power}}, M - 2 - p]$, where $H[q, df]$ is a function that returns quantiles of the t-distribution for a given quantile q and degrees of freedom. For example, with 28 degrees of freedom, the quantile for $1 - \frac{\alpha}{2} = .025$ (for a two-tailed test with $\alpha = .05$) is $H[q_{1-.025}, 28] = 2.05$, and the quantile for the power of .8 is $H[q_{1-.8}, 28] = -0.85$, and so the value of Q is $2.05 - (-.85) = 2.9$. Since $\lambda \approx Q$, we can rearrange the formula for the expected test by replacing λ with Q and solve for δ,

$$MDES = Q\sqrt{\frac{\rho(1-R_2^2)}{Mf(1-f)} + \frac{(1-\rho)(1-R_1^2)}{nMf(1-f)}}.$$

To find how the MDES changes in response to n, one of the several variables included in the MDES function, requires finding a first-order partial derivative of the MDES function in Design Effect form. This form isolates the design effect, $1 + (n-1)\rho - (R_1^2 + (nR_2^2 - R_1^2)\rho)$, in a single element,

$$\text{MDES} = Q\sqrt{\frac{1}{Mf(1-f)} \times \frac{1+(n-1)\rho-(R_1^2+(nR_2^2-R_2^2)\rho)}{n}},$$

which facilitates finding the partial derivative of this function with respect to n.

Since the key element of the MDES function of interest, n, is within a square root function, finding the partial derivative in raw effect-size terms is difficult. Yes, finding the minimum detectable squared effect size is possible, δ_m^2, but changes to the square of a non-linear function (which is not a straight line) requires knowledge of each effect size at a point. That is, algebraically, $(\delta - \gamma)^2 - \delta^2 = \delta(\delta - 2\gamma)$, and likewise with the square-roots. These issues are

side-stepped, however, when we instead find the partial derivative of the natural log of the MDES ($\ln(\delta_m)$).

Using a ratio of MDES values as a metric for change is straightforward with logged values. By taking the derivative for the natural log of the MDES, the derivative function is then in units of change for the natural log of the MDES, $\ln(\delta_{m|n+1}) - \ln(\delta_{m|n})$, which, by the rules of logarithms, is equivalent to the natural log of the SDESR. The log of the ratio of the MDES with an extra unit per cluster ($\delta_{m|n+1}$) to the larger MDES with the base number of units per cluster ($\delta_{m|n}$) is the same value as the difference in natural logarithms of the same MDES values, or

$$\ln(\delta_{m|n+1}) - \ln(\delta_{m|n}) = \ln\left(\frac{\delta_{m|n+1}}{\delta_{m|n}}\right).$$

As a result of these equalities, the procedure used to find expressions (7) and (8) involves the rate of change in the natural log of the MDES with respect to n through partial differentiation. The result for the MDES without covariates is

$$\frac{\partial \ln(\delta_m)}{\partial n} = \frac{\rho - 1}{2n(1+(n-1)\rho)}, \tag{11}$$

and with covariates it is

$$\frac{\partial \ln(\delta_m)}{\partial n} = \frac{(\rho-1)(1-R_1^2)}{2n\left(1+(n-1)\rho-\left(R_1^2+(nR_2^2-R_1^2)\rho\right)\right)}. \tag{12}$$

Expression (7) is the natural exponentiation of (11) and expression (8) is the natural exponentiation of (12), to convert the log ratio into the ratio of actual MDES units.

In both cases, the rate of change in the natural log of the MDES, relative to n, is only a function of n, the ICC (ρ), and the R^2 values if applicable. This is because the variables in the MDES formula that summarize power, significance, number of clusters, and treatment fraction,

$\frac{Q^2}{Mf(1-f)}$, are constant and do not involve n. Also note that each derivative includes in the denominator the design effect, $1 + (n-1)\rho$ for models without covariates, or $1 + (n-1)\rho - (R_1^2 + (nR_2^2 - R_1^2)\rho)$ for models with covariates.

Again, choice of the SDESR metric is because SDESR $= \theta = \frac{\delta_{m|n+1}}{\delta_{m|n}}$, and so $\frac{\partial \ln(\delta)}{\partial n} = \ln(\theta) = \ln(\text{SDESR})$ when $\partial n = 1$. Expressions (9) and (10) result from replacing $\frac{\partial \ln(\delta_m)}{\partial n}$ in expressions (11) and (12) with $\ln(\theta)$ and then solving for (or rearranging to put on the left side) n. Extensions can easily be created for larger increases in the units by dividing $\ln(\theta)$ by an appropriate factor, such as 5 units, 10, or 20 units.

A note on adding clusters vs. adding units

Similar procedures can be used to find the benefits of adding clusters. However, the variable Q is based the number of clusters for degrees of freedom and relies on several complex functions to produce the quantiles of the t-distribtuion, which makes the partial derivative with respect to M complicated. However, if we assume a large sample of clusters, $Q \approx 2.8$ for a typical level of power and significance ($\alpha = .05$ for a two-tailed test and power of .8). Under these circumstances, $\frac{\partial \ln(\delta_m)}{\partial M} = \frac{-1}{2M}$. To see this, suppose we isolate all elements which do not included M in the MDES function

$$\text{MDES} = \sqrt{\frac{1}{M} \times Q^2 \frac{1+(n-1)\rho-\left(R_1^2+\left(nR_2^2-R_2^2\right)\rho\right)}{nf(1-f)}},$$

which renders the element $Q^2 \frac{1+(n-1)\rho-\left(R_1^2+\left(nR_2^2-R_2^2\right)\rho\right)}{nf(1-f)}$ a constant. Our effective function is then

$$g(M) = \ln\left(\left(\frac{1}{M}\right)^{\frac{1}{2}}\right).$$

Using the chain rule of calculus, the first derivative of $\ln\big(g(x)\big)$ is $\frac{1}{(x)}^{g'(x)}$, and we find

$g'^{(M)} = \frac{1}{2}\left(\frac{1}{M}\right)^{-\frac{1}{2}}$, noting that the square root of a quantity is the quantity to the ½ power and applying the power rule, $\frac{d}{dx}(x^a) = ax^{1-a}$, which yields,

$\frac{\partial \ln(\delta_m)}{\partial M} = \frac{-1}{2M}$.

If the ICC is 0, then

$$(\text{MDES}|\rho = 0) = Q\sqrt{\frac{1}{nMf(1-f)}} = \sqrt{\frac{1}{n} \times Q^2 \frac{1}{Mf(1-f)}},$$

which then similarly renders $\frac{\partial \ln(\delta_m)}{\partial n} = \frac{-1}{2n}$; c.f. expression (11) for a non-zero ICC.

In conclusion, the gains to adding clusters are the same as adding units per cluster only when there is no cluster-level variation, that is $\frac{\partial \ln(\delta_m)}{\partial n} = \frac{\partial \ln(\delta_m)}{\partial M}\ \forall\ n = M$ if and only if the ICC is precisely zero ($\rho = 0$) and only for tests with large numbers of clusters. When the ICC is greater than 0, which can be expected, adding a cluster will be more beneficial than adding units per cluster.

Appendix B: Guide to supplemental R code

The supplemental materials provide a script written in R Code (R Core Team 2021), which creates and tests functions, `crt_sdesr` and `crt_n`, for expressions (8) and (10), respectively. These functions can be typed or copy/pasted into an R session. Tables 1-6 were generated using these functions in conjunction with the R "outer" function. Note that in R, "log" defaults to the natural log written as $\ln(.)$. Table 7 also uses these functions in conjunction with a procedure to find a value of clusters, which is also included in the supplemental material script.

Figure B1 presents a screenshot of an example session. For those unfamiliar with R, the code above does not require additional packages to be installed, only the base R program available at cran.r-project.org for installation to most traditional computers or for immediate use

on several websites (I leave it to the reader to search for them as they often change). Once installed and started, a "console" will appear with an input cursor next to the prompt ">". Typing or pasting the code in the supplemental material will load into memory the functions `crt_sdesr`, which will compute the SDESR based on n, the ICC, and the R-squares, and `crt_n`, which computes n based on the SDESR, the ICC, and the R-square values, allowing them to be used.

FIGURE B1 ABOUT HERE

To use these functions in R, simply type the name of the function, an open parenthesis, and each argument followed by an equal sign and its numeric value. Close the line with a close parenthesis and type the enter key. The resulting number will appear after a "[1]", which is simply an indicator for the row of the output (many R functions produce output of numerous rows).

Figure B1: Example R session using the functions which enact expressions

```
> crt_sdesr <- function(rho, # icc
+                       n, # units per cluster
+                       R2_1 = 0, # unit R-square
+                       R2_2 = 0 # cluster R-square
+ ) {
+   numer <- ((rho-1)*(1-R2_1))
+   denom <- 2*n*(1+(n-1)*rho-(R2_1+(n*R2_2-R2_1)*rho))
+   return(
+     exp(numer/denom)
+   )
+ }
> crt_sdesr(rho = .1, n = 10)
[1] 0.9765941
> crt_sdesr(rho = .1, n = 10, R2_1 = .5, R2_2 = .25)
[1] 0.9814247
> crt_n <- function(rho, # icc
+                   sdesr, # theta (SDESR)
+                   R2_1 = 0, # unit R-square
+                   R2_2 = 0 # cluster R-square
+ ) {
+   numer <- (rho - 1)*log(sdesr)*(1-R2_1) -
+     sqrt((rho-1)*log(sdesr)*(R2_1-1)*(2*rho*(R2_2-1) +
+                                       log(sdesr)*(rho-1)*(R2_1-1)))
+   denom <- 2*rho*log(sdesr)*(1-R2_2)
+   return(
+     numer/denom
+   )
+ }
> crt_n(rho = .1, sdesr = .9765941)
[1] 10.00001
> crt_n(rho = .1, sdesr = .9814247, R2_1 = .5, R2_2 = .25)
[1] 10
```