Assessing the Precision of Multisite Trials for Estimating the Parameters Of Cross-site Distributions of Program Effects

Howard S. Bloom
MDRC

Jessaca Spybrook
Western Michigan University

May 10, 2016

Manuscript Under Review.

This paper was funded by the Spencer Foundation and the William T. Grant Foundation. The authors would like to thank Stephen Raudenbush and Luke Miratrix for helpful suggestions about some of the issues that are discussed. However, any opinions expressed or errors made in are solely those of the authors. Lastly, note that the authors are listed in alphabetical order.

Abstract

Multisite trials, which are beginning to be used frequently in education research, provide an exciting opportunity for learning how the effects of education programs are distributed across sites. In particular, these studies can produce rigorous estimates of a cross-site *mean* program effect size, a cross-site *standard deviation* of program effect sizes, and a *difference* between cross-site mean program effect sizes for two subgroups of sites. However, to capitalize on this opportunity will require adequately powering future trials to estimate these parameters. To help researchers do so, we present a simple approach for computing the minimum detectable values of these parameters. The paper then applies this approach to illustrate for each parameter, the precision tradeoff between increasing the number of study sites and increasing site sample size. Findings are presented for multisite trials that randomize individual sample members and for those which randomize intact groups or clusters of sample members.

Introduction

Randomized trials are being used with increasing frequency to estimate causal effects of social and educational interventions or "treatments". For example, since 2002, the National Center for Education Research of the Institute of Education Sciences (IES) has funded more than 160 evaluations that either randomized individuals to treatment and control conditions or randomized intact groups or clusters (e.g. classrooms, schools or early child education centers) to these conditions. For such studies, multi-site trials are becoming the design of choice (Spybrook & Raudenbush, 2009; Spybrook, Shi, & Kelcey, 2016).

In the present paper, we use the phrase multi-site trial (MST) to denote a study that randomizes individuals to experimental conditions within study sites. For example, the national Head Start Impact Study (Puma et. al. 2010) randomized 4,667 program applicants from 378 local Head Start centers (sites) to either receive or not receive a program offer. We use the phrase multi-site cluster-randomized trial (MSCRT) to denote a study that randomizes clusters to experimental conditions within study sites. For example, the evaluation of Reading Comprehension Programs (James-Burdumy et al, 2009) randomized schools to a treatment or control group within each of 10 school districts (sites).

Multi-site trials have several important advantages. First, multiple sites are often needed to produce a sample that is large enough to provide adequate statistical power. Second, multiple sites can strengthen the generalizability of a study's findings. Third, and most relevant for the present paper, multiple sites with random assignment of subjects within sites can make it possible to rigorously study a *cross-site distribution* of program effects.

Consider the evaluation of charter middle schools conducted by Gleason et al., (2010) based on data for roughly 2,100 applicants to 29 over-subscribed charter middle schools from 15 states. Applicants for each charter school (study site) were randomly assigned via lottery to either receive or not receive an offer of admission. From follow-up data for these applicants, Weiss et al. (under review) estimate that the cross-site mean effect size in standard deviation units (*a key parameter* that can be estimated from a multi-site study) of a charter-school admissions offer was virtually zero for math and reading achievement (-0.04σ and -0.04σ for math test scores and -0.02σ and -0.6σ for reading test scores, during the first and second study follow-up years, respectively.)[1] This implies that average student math and reading achievement for charter middle schools in the study were virtually the same as that for their counterfactual alternatives.[2] A policymaker might thus conclude that charter middle schools are uniformly no more or less effective than their alternatives.

However, Weiss et al (under review) also demonstrate that the effectiveness of these charter middle schools, relative to their counterfactual alternatives, varies substantially. They estimate that the cross-site (i.e. cross-school) standard deviation of effect sizes, in standard deviation units, is 0.15σ and 0.25σ for reading achievement and 0.22σ and 0.35σ for math achievement in the first and second study follow-up years, respectively. These findings suggest that charter middle schools in the study ranged from being substantially more effective than their local alternatives to being substantially less effective. The cross-site standard deviation of effect sizes represents a *second important parameter* that can be estimated in a multi-site trial.

[1] These effect sizes are stated as a proportion of the total standard deviation of individual outcomes for control group members in the study sample.

[2] Even though the study was based on a convenience sample of charter middle schools and it is thus not possible to rigorously define the population of charter middle schools that is represents, this population can exist in principle and thus is a valid target of inference (Raudenbush and Bloom, 2015).

This striking variation in charter school effects invites the question: What factors predict or account for the variation? For example, might this variation reflect differences in charter school location (e.g. urban vs. rural), their organizational structures or their educational approaches? A simple way to address questions like this about the "moderation" of charter school effects is to estimate the difference in mean effects for two subgroups of charter schools that are defined in terms of a specific characteristic. This is *a third key parameter* that can be estimated from a multi-site trial.

There is a growing technical literature on the statistical power of MSTs and MSCRTs. Much of this literature focuses on power to detect cross-site mean impacts (e.g. Raudenbush & Liu, 2000; Hedges & Pigott, 2001; Moerbeek & Teerenstra, 2016; Schochet, 2008; Konstantopoulos, 2008; Hedges & Rhoads, 2010; and Dong and Maynard, 2013). Some of this literature also examines power to detect cross-site program effect variation (e.g. Raudenbush & Liu, 2000; Hedges & Pigott, 2004; Konstantopoulos, 2008; and Spybrook, 2014). An even smaller portion of this literature examines power to detect differences in cross-site mean effects for subgroups of sites (e.g. Raudenbush & Liu, 2000; and Hedges & Pigott, 2004).

In light of this, the goals of the present paper are to: (1) consolidate what has been learned from the extant technical literature; (2) add some new insights about factors that influence statistical power and precision, (3) present this new and existing information in a way that can be readily applied by empirical researchers; (4) use the information to illustrate tradeoffs that exist when designing a study to detect the three key parameters of a cross-site impact distribution, and (5) do so for both MSTs and MSCRTs.

To make our discussion concrete, we frame it in terms of statistical *precision* stated as a minimum detectable effect (MDE) for an outcome (Y) that is reported in its natural units (e.g. earnings in dollars or academic attainment in course credits) or a minimum detectable effect size (MDES) for a standardized outcome (Z) that is reported in standard deviation units. By definition, an MDE (Bloom, 1995) or MDES (Bloom, 2005) is the smallest true mean program effect or effect size that a study design can detect at a specified level of statistical significance (typically 0.05 for a two-tailed test) with a specified level of statistical power (typically 80 percent). We thus consider how alternative sample designs for multi-site trials influence: (1) the minimum detectable value of a cross-site *mean* program effect or effect size, (2) the minimum detectable value of a cross-site *standard deviation* of program effects or effect sizes and (3) the minimum detectable *difference* between the mean program effects or effect sizes for two subgroups of program sites. All findings refer to experimental estimates of effects of random assignment to a program, intervention or treatment (effects of "intent to treat").

The paper is organized as follows. Section 2 examines how alternative sample designs influence the precision of program effect estimates from MSTs. Section 3 extends the discussion to MSCRTs. Section 4 reflects on our findings and their implications for research practice.

Precision for Multisite Trials

Assume for simplicity that we are considering the precision of a multisite trial (MST) with J sites, n sample members from each site and proportion $\bar{T}$ of the sample members from each site randomized to treatment.

Estimands

We begin by examining precision for estimating a cross-site *mean* program effect (β) and a cross-site *standard deviation* of program effects (τ) for a population distribution of site-specific mean program effects (B_j) like that illustrated in Figure 1. By definition:

$$\beta \equiv lim_{J_* \to \infty} \sum_{j=1}^{J_*} \frac{B_j}{J_*} \quad (1)$$

and

$$\tau \equiv lim_{J_* \to \infty} \sum_{j=1}^{J_*} \sqrt{\frac{(B_j - \beta)^2}{J_*}} \quad (2)$$

Insert Figure 1 approximately here

We then examine the statistical precision of estimates of a *difference* in cross-site mean program effects ($\Delta \equiv \beta_{II} - \beta_I$) for two-subpopulations of sites (I and II), as illustrated in Figure 2. By definition:

$$\beta_I \equiv lim_{J_I \to \infty} \sum_{j_I=1}^{J_I} \frac{B_{j_I}}{J_I} \quad (3)$$

and

$$\beta_{II} \equiv lim_{J_{II} \to \infty} \sum_{j_{II}=1}^{J_{II}} \frac{B_{jII}}{J_{II}} \quad (4)$$

where B_{j_I} and $B_{j_{II}}$ are the mean program effects for site j in site subpopulations I and II, respectively.

Insert Figure 2 approximately here

Estimation Models

To estimate β and τ from data for our multisite trial we would use the following two-level model with fixed site-specific intercepts (α_j), random site-specific program effect coefficients (B_j) and fixed coefficients (θ_k) for each individual-level baseline covariate (X_k).[3]

[3] Bloom et al. (revise and resubmit) discuss this model and its properties.

Level One: Individuals

$$Y_{ij} = \alpha_j + B_j T_{ij} + \sum_{k=1}^{K} \theta_k X_{kij} + e_{ij} \quad (5)$$

Level Two: Sites

$$\alpha_j = \alpha_j \quad (6)$$

$$B_j = \beta + b_j \quad (7)$$

where $e_{ij} \sim N\left(0, \sigma^2_{|\boldsymbol{X}\alpha_j}\right)$, $b_j \sim N(0, \tau^2)$, Y_{ij} is the observed value of the outcome for individual *i* from site *j*; T_{ij} is 1 if individual *i* from site *j* was randomized to treatment and 0 otherwise; X_{kij} is the value of baseline covariate k for individual *i* from site *j*; θ_k is the coefficient for covariate k; α_j is the conditional population mean control group outcome for site *j*; B_j is the population mean treatment effect for site *j*; β is the cross-site mean treatment effect for the population of sites; e_{ij} is a random error that varies independently and identically across individuals within sites and experimental conditions, with a mean of zero and a variance of $\sigma^2_{|\boldsymbol{X}\alpha_j}$; and b_j is a random error that varies independently and identically across sites with a mean of 0 and a variance of (τ^2).

To estimate a difference between cross-site mean program effects for two population subgroups of sites (Δ) from data for our multisite trial we would modify the preceding model by adding a binary site subgroup indicator (W_j) to Equation 7, yielding:

$$B_j = \beta_I + \Delta W_j + b'_j \quad (8)$$

where β_I is the cross-site mean program effect for site subgroup I, W_j equals one if site *j* is from site subgroup *II* and zero otherwise, and ($b'_j \sim N(0, \tau^2_{|W})$.[4]

A Minimum Detectable Effect Size

[4] To the extent that W predicts cross-site variation in program effects, $\tau^2_{|W} < \tau^2$.

This section derives an expression for the minimum detectable effect size (MDES) of a multisite trial and uses the expression to explore factors that influence the MDES.

The Expression: We refer to the minimum detectable value of a cross-site mean program effect (β) *in its natural units* as a minimum detectable effect or MDE, where:

$$MDE = M_{J-1} se(\hat{\beta}) \tag{9}$$

$se(\hat{\beta})$ is the standard error of the mean program effect estimator ($\hat{\beta}$) and M_{J-1} is a multiplier that, for a two-tailed hypothesis test at the 0.05 significance level with 80 percent power, rapidly approaches 2.8 as J increases (Bloom, 1995).[5] By extension from Raudenbush and Liu (2000):

$$se(\hat{\beta}) = \sqrt{(\frac{1}{J})(\tau^2 + \frac{\sigma^2_{|X\alpha_j}}{n\bar{T}(1-\bar{T})})} \tag{10}$$

where $J, n, \bar{T}$ and τ are defined as before and $\sigma^2_{|X\alpha_j}$ is the individual-level residual error variance for Equation 5, which is assumed for simplicity to be approximately constant across sites and experimental conditions.[6] Therefore:

$$MDE = M_{J-1}\sqrt{(\frac{1}{J})(\tau^2 + \frac{\sigma^2_{|X\alpha_j}}{n\bar{T}(1-\bar{T})})} \tag{11}$$

The two variance components in the MDE (τ^2 and $\sigma^2_{|X\alpha_j}$) are the two sources of uncertainty about our estimator, $\hat{\beta}$. The first variance component, τ^2, reflects sampling error due to the fact that study sites comprise a sample of sites from a population of sites – either implicitly

[5] Pages 158 and 159 of Bloom (2005) explain why the multiplier for a minimum detectable effect (M) equals $t_{\alpha/2} + t_{1-\beta}$, where $t_{\alpha/2}$ is the critical *t* value for a two-tailed hypothesis test and $t_{1-\beta}$ is the corresponding t value for power equal to β.

[6] Assuming that $\sigma^2_{|X\alpha_j}$ is constant across sites does not produce major estimation or inference problems unless $\sigma^2_{|X\alpha_j}$ and site sample sizes vary substantially and are highly correlated (Bloom, et al., revise and resubmit). Assuming that $\sigma^2_{|X\alpha_j}$ is the same for treatment and control group members does not produce major estimation or inference problems unless the values for $\sigma^2_{|X\alpha_j}$ and the sample sizes differ substantially across the two experimental groups (Bloom, et al, revise and resubmit). These simplifying assumptions are made by most, if not all, prior discussions of the statistical power or precision of multisite trials (e.g. Raudenbush and Liu, 2000).

or explicitly. Other things being equal, the more program effects vary across sites, the more uncertainty $\hat{\beta}$ will reflect and the larger the MDE will be. Therefore when planning a multisite trial, it is important to have a basis for "guesstimating" τ^2.[7]

The second variance component in the MDE ($\sigma^2_{|\boldsymbol{X}\alpha_j}$) reflects sampling error in the estimated program effect for each site ($\hat{B}_j$) and thus in the estimated cross-site mean program effect ($\hat{\beta}$). Other things being equal, the larger $\sigma^2_{|\boldsymbol{X}a_j}$ is, the more uncertainty $\hat{\beta}$ will contain, and the larger the MDE will be. For assessing the likely magnitude of this variance component it is useful to restate it as follows.

$$\sigma^2_{|\boldsymbol{X}\alpha_j} = (1-\rho_C)(1-R^2_{C(within)})\sigma^2_C \qquad (12)$$

where σ^2_C is the *total* variance of control group member outcomes (within and between sites) expressed in their natural units; ρ_C is the proportion of total control group outcome variation that is *between* sites (*aka*, the control group intra-class correlation);[8] and $R^2_{C(within)}$ is the proportion of within-site outcome variation for control group members that is explained by our baseline covariates ($\boldsymbol{X}$). Expressing $\sigma^2_{|\boldsymbol{X}\alpha_j}$ in terms of σ^2_C provides a bridge to our discussion of a minimum detectible *effect size* (MDES). Expressing $\sigma^2_{|\boldsymbol{X}\alpha_j}$as a function of ρ_C and $R^2_{C(within)}$ provides a bridge to the empirical literature on values of these parameters.

Substituting Equation 12 into Equation 11 yields:

$$MDE = M_{J-1}\sqrt{(\frac{1}{J})(\tau^2 + \frac{(1-\rho_C)(1-R^2_{C(within)})\sigma^2_C}{n\bar{T}(1-\bar{T})})} \qquad (13)$$

Now consider the implications of Equation 13 for the precision of an estimator of a cross-site mean program effect on a *standardized* outcome measure or "z-score" that is defined, as

[7] Weiss et al. (under review) provide estimates of τ^2 from data for 15 multisite trials.
[8] The parameter, ρ, also represents the proportion of σ^2_C that is explained by the site intercepts, α_j.

follows, in terms of the mean outcome (μ_C) and standard deviation of outcomes (σ_C) for a control group population.[9]

$$Z_{ij} \equiv \frac{Y_{ij}-\mu_C}{\sigma_C} \tag{14}$$

We denote the cross-site mean program effect on this z-score as β_*, where $\beta_* = \frac{\beta}{\sigma_C}$. This mean effect size, is expressed in units of σ_C. We refer to the minimum detectable value of β_* as a minimum detectable *effect size*, or $MDES$ where:

$$\begin{aligned} MDES &= \frac{MDE}{\sigma_C} \\ &= \left(\frac{M_{J-1}}{\sigma_C}\right)\sqrt{\left(\frac{1}{J}\right)\left(\tau^2 + \frac{(1-\rho_C)(1-R^2_{C(within)})\sigma^2_C}{n\bar{T}(1-\bar{T})}\right)} \\ &= M_{J-1}\sqrt{\left(\frac{1}{J}\right)\left(\frac{\tau^2}{\sigma^2_C} + \frac{(1-\rho_C)(1-R^2_{C(within)})}{n\bar{T}(1-\bar{T})}\right)} \\ &= M_{J-1}\sqrt{\left(\frac{1}{J}\right)\left(\tau^2_* + \frac{(1-\rho_C)(1-R^2_{C(within)})}{n\bar{T}(1-\bar{T})}\right)} \end{aligned} \tag{15}$$

and ($\tau_* = \frac{\tau}{\sigma_C}$) is the cross-site standard deviation of effect sizes.

Note that the number of sites (*J*) reduces the influence of both variance components on the MDES, whereas the number of sample members per site (*n*) only reduces the influence of individual-level sampling error. Thus, other things being equal, increasing the number of sites by a given percentage reduces the MDES by more than does increasing the size of site samples by that percentage. Also note that the reduction in the MDES produced by the explanatory power of individual-level covariates ($R^2_{C(within)}$), the intra-class correlation of individual outcomes nested within sites (ρ_C) and the proportion of sample members randomized to treatment ($\bar{T}$) only

[9] In practice, we use the sample control-group mean ($\bar{Y}_C$) and standard deviation (s_C) of outcomes to define a standardized z-score. Weiss et al. (under review) explores issues that arise when z-scores are defined in terms of the mean and standard deviation for other reference population, such as a state or the nation as a whole.

reduces the MDES by reducing uncertainty due to individual-level sampling error. Consequently, the number of sites has more influence than the size of site samples on the MDES. Lastly, note that the amount of cross-site impact variation that exists (τ_*) can have a major influence on the MDES.

In thinking about how these parameters influence the MDES, it is useful to distinguish between ρ_C, $R^2_{C(within)}$ and τ_*, which are mainly determined by the outcome domain, grade level, and nature of the sites involved and thus are a function of the research question being addressed, versus $J, n,$ and $\bar{T}$, which are features of the experimental design used to address the research question.

Illustrative Findings: Equation 15 illustrates that to assess the MDES for a proposed multisite trial (given *J*, *n* and $\bar{T}$) one must assume values for τ_*, ρ_C and $R^2_{C(within)}$. For studies that randomize students within schools to estimate intervention effects on achievement test scores, assumed values for ρ_C and $R^2_{C(within)}$ can be based on extensive estimates of these parameters reported by Bloom, et al., (2007) from data for five urban school districts, by Hedges and Hedberg (2007) from data for several national longitudinal surveys, and by other related sources, such as Zhu et al, (2012), Jacob et al. (2010) and Bloom et al. (1999).[10] Currently however, little empirical information about τ_* exists, the most extensive of which is that provided by Weiss et al. (under review) based on data for 15 MSTs. Some limited additional information is also provided by Olsen et al. (2015).

Consider a multisite trial for estimating the cross-site mean effect size of an educational intervention on third-grade reading scores with: 30 elementary schools (J); 50 third-graders per

[10] Because these two studies examine different populations (urban districts in one case and the nation in another), their estimates of ρ_C and $R^2_{C(within)}$ are not fully comparable. Nonetheless, together they provide a strong basis for extrapolating findings to other settings.

school (n); 0.60 of the third graders at each school randomized to treatment ($\bar{T}$), and individual second-grade reading scores as a baseline covariate (X). For third-grade reading test scores, Table 8 in Bloom et al. (2007) reports estimated values of ρ_C ranging from 0.15 to 0.22 across the five school districts examined, with a mean of 0.18. The table also reports corresponding estimates of $R^2_{C(within)}$ that range from 0.22 to 0.52, with a mean of 0.38.[11] So let's assume these mean values for ρ_C and $R^2_{C(within)}$.

Now we must assume a value for τ_*. At one extreme, Weiss et al. (under review) report an estimate of $0.25\sigma_C$ for the cross-school standard deviation of small-class effects on reading achievement for elementary-school students in the Tennessee STAR class size experiment (Word et al., 1990)). At the other extreme, the paper reports an estimate of τ_* equal to zero from data for a large-scale experimental evaluation of Teach for America in elementary schools **(**Clark, et al., 2015). Given the preceding assumptions for $J, n, \bar{T}, \rho_C \; and \; R^2_{C(within)}$ plus a value of τ_* equal to 0.25, Equation 15 implies that our MDES would be $0.17\sigma_C$. If instead we assume that τ_*equals zero, Equation 15 implies that our MDES would be $0.11\sigma_C$. This comparison illustrates the powerful influence of τ_* on the MDES,

Now consider how Equation 15 enables us to examine the relative influence of the number of sites (*J*) and the size of site samples (*n*) on the MDES. Assume that $\bar{T} = 0.5$, $\tau_* = 0.15, \rho_C = 0.15\sigma_C, R^2_{C(within)} = 0.4$, our statistical significance threshold is 0.05 for a two-tail test and our desired statistical power level is 80 percent. To compare the influences of *J* and *n* on the MDES, look at the shaded MDES values on the diagonal of Table 1. These values represent a constant total sample of 1,000 persons allocated to widely varying numbers of sites. As can be

[11] Table 3 (p. 69) of Hedges and Hedberg (2007) reports a national estimate of ρ_C for third-grade reading test scores equal to 0.27 and a corresponding estimate of $R^2_{C(within)}$ equal to 0.48.

seen, the MDES is far more sensitive to the number of sites than to the size of site samples. For example, at one extreme, with only five sites and 200 sample members per site, the MDES is 0.30σ, whereas at the other extreme, with 200 sites and only 5 persons per site, the MDES is 0.13σ, where to simplify reporting and make it comparable to common practice, we denote the standard deviation of individual-level outcomes for a study's control group as σ instead of σ_C.

Insert Table 1 approximately here

Three further points are important to note when designing the sample for a multisite trial. First, the precision tradeoff between the number and sample size of sites depends on the amount of cross-site impact variation that exists (τ_*). For example, if τ_* were zero – instead of 0.15σ in Table 1 – increasing the number of sites by a given percentage would have *the same* effect on precision as increasing site sample sizes by that percentage. This illustrates why information about the likely value of τ_* is essential for planning an MST. Second, one must temper statistical precision concerns by the practical tradeoffs that exist between the cost and feasibility of recruiting and operating more sites versus increasing the size of site samples (see Raudenbush and Liu, 2000). Third, to convert an MDES for a standardized outcome measure (Z) that is expressed in units of σ to an MDE for its corresponding outcome measure (Y) expressed in natural units simply multiply the MDES by σ.

A Minimum Detectable Effect Size Standard Deviation

This section considers how sample designs for multisite trials influence our ability to quantify cross-site impact variation.

The Expression: Appendix A develops the following expression for a minimum detectable cross-site standard deviation of program effect sizes.

$$MDESSD = \sqrt{\left(\frac{(1-\rho_C)\left(1-R^2_{C(within)}\right)}{n\bar{T}(1-\bar{T})}\right)\left(\frac{F_{0.05}}{F_{0.80}} - 1\right)} \quad (16)$$

where ρ_C, $R^2_{C(within)}$, n and $\bar{T}$ are defined as before, $F_{0.05}$ is the 0.05 critical value of an F statistic with J-1 numerator degrees of freedom and J(n-2)-K denominator degrees of freedom, and $F_{0.80}$ is the value of an F statistic with J-1 numerator degrees of freedom and J(n-2)-K denominator degrees of freedom that is below 80 percent of the values in the distribution of that statistic.[12]

Equation 16 illustrates that site sample size (n) has a direct and pronounced influence on the MDESSD. For example, quadrupling the value of *n* cuts the MDESSD in half. Likewise, the individual-level explanatory power (ρ_C and $R^2_{C(minimum)}$) of site indicators and individual-level covariates can play an important role in reducing the MDESSD. In contrast, as illustrated below, the number of study sites (J) – which influences the expression $\left(\frac{F_{0.05}}{F_{0.80}}\right)$ through its effect on the number of numerator and denominator degrees of freedom for $F_{0.05}$ and $F_{0.80}$, has much less of an effect on the MDESSD.

Illustrative Findings: Table 2 presents estimates of the MDESSD for the combinations of *J* and *n* that were examined for the MDES in Table 1 and for the values of $\bar{T}$, ρ_C, $R^2_{C(within)}$, statistical significance and statistical power that were assumed for Table 1. Note that the pattern of shaded findings along the diagonal of Table 2 for the MDESSD is the reverse of that for its counterpart in Table 1 for the MDES. In Table 2 we see that increasing site sample size by a given proportion reduces the MDESSD by more than increasing the number of sites by that proportion. Thus one must prioritize estimates of β_* or τ_* when planning a multisite trial.

[12] To convert an MDESSD (in units of σ) to a minimum detectable effect standard deviation (in natural units of the outcome involved) multiple the MDESSD by σ.

Insert Table 2 approximately here

A Minimum Detectable Effect Size Difference

When program effects vary across sites it is important to understand the features of sites that predict or "moderate" this variation. The simplest form of such a predictor is a binary site characteristic. For example, we might want to compare the academic achievement effects of urban versus non-urban charter schools (Angrist et al., 2013) or the socio-emotional effects of Head Start programs that are rated as "high quality" versus other Head Start programs (Peck and Bell, 2014). In each case, we are interested in comparing cross-site impact distributions for two subgroups of sites (see Figure 2).

In particular, we might be interested in estimating the *difference* between the cross-site *mean* program effects or effect sizes for the two subgroups of sites. As noted earlier, we could estimate the difference in mean program *effects* ($\Delta = \beta_{II} - \beta_I$) using the model represented by Equations 5, 6 and 8 for the outcome measure, *Y,* in its natural units. To convert these findings to standardized effect sizes, we can divide them by σ_C (i.e. $\Delta_* = \frac{\Delta}{\sigma_C} = \frac{\beta_{II}}{\sigma_C} - \frac{\beta_I}{\sigma_C}$). Thus when planning our study it is important to try to determine the minimum detectable value of Δ_*. We refer to this parameter as a minimum detectable effect size *difference* or MDESD.

Some Intuition: Unlike an MDES or an MDESSD, an MDESD does not have a closed form expression. Thus to compute it requires an iterative process. Before describing this process it is useful to develop some intuition about it. For this purpose, consider a binary site-level characteristic (W) that is equal to one for sites from subgroup II (e.g. urban charter schools) and zero for sites from subgroup I (e.g. non-urban charter schools). Assume that W predicts a proportion (R_W^2) of the total cross-site variation in program effect sizes (τ_*^2).

Appendix B derives the following relationship between Δ_*, τ_*^2, R_W^2 and the proportion of study sites that are in each subgroup (π for subgroup II and $(1-\pi)$ for subgroup I).

$$\Delta_* = \sqrt{\frac{R_W^2 \tau_*^2}{\pi(1-\pi)}} \tag{17}$$

This relationship demonstrates several important points. *First, the amount of cross-site impact variation that exists (τ_*^2) determines the maximum subgroup impact difference that is possible ($\Delta_{*(max)}$).* For example, if all sites have the same program effect ($\tau_*^2 = 0$) then there is no margin for a site subgroup impact difference ($\Delta_{*(max)} = 0$). However, as cross-site impact variation increases, the margin for a site subgroup impact difference increases.

Second, for a given total amount of cross-site impact variation (τ_*^2), the explanatory power (R_W^2) of our site-level program-effect predictor (W) is closely related to the subgroup mean impact difference (Δ_*). Indeed, they are like two sides of the same coin. For example, if our subgroup indicator predicts no cross-site impact variation ($R_W^2 = 0$) then Δ_* must equal zero. But as R_W^2 increases, Δ_* must increase accordingly (given τ_*^2 and π).

Third, the maximum possible value of Δ_*, given τ_*^2 and π, occurs when our site subgroup indicator predicts all of the existing cross-site impact variation (when $R_W^2 = 1$). Thus:

$$\Delta_{*(max)} = \sqrt{\frac{\tau_*^2}{\pi(1-\pi)}} \tag{18}$$

A Candidate Expression for the MDESD, When It Exists: Appendix B derives the following expression for a minimum detectable effect size difference, when it exists.

$$MDESD = M_{J-2} \sqrt{\left(\frac{1}{\pi(1-\pi)J}\right)\left((1-R_W^2)\tau_*^2 + \frac{(1-\rho_C)\left(1-R_{C(within)}^2\right)}{n\bar{T}(1-\bar{T})}\right)} \tag{19}$$

As can be seen, Equation 19 for an MDESD is similar to Equation 15 for an MDES. This is because an MDESD is defined for a difference between two cross-site mean impacts and an MDES is defined for an overall cross-site mean impact. However there are several important differences between Equations 19 and 15. First, the multiplier, M_{J-2}, in Equation 19 replaces the multiplier, M_{J-1}, in Equation 15 to account for the difference in the number of degrees of freedom involved. Second, the term, $\pi(1-\pi)J$, in Equation 19 replaces the term, J, in Equation 15 to reflect the difference that exists in the precision effect of the number of sites involved. Third, the term, $(1-R_W^2)\tau_*^2$, in Equation 19 replaces the term, τ_*^2, in Equation 15 to reflect the precision effect of the predictive power of *W*.

When Might the MDESD Not Exist? The problem with using Equation 19 to determine an MDESD is – as noted earlier – that the MDES can be smaller than $\Delta_{*(max)}$ and thus might not exist. One way to see this is to consider a program with no cross-site impact variation ($\tau_*^2 = 0$), where:

$$\Delta_{*(max)} = \sqrt{\frac{\tau_*^2}{\pi(1-\pi)}} = \sqrt{\frac{0}{\pi(1-\pi)}} = 0 \tag{20}$$

and

$$\begin{aligned} MDESD &= M_{J-2}\sqrt{\left(\frac{1}{\pi(1-\pi)J}\right)\left((1-R_W^2)0 + \frac{(1-\rho_C)\left(1-R_{C(within)}^2\right)}{n\bar{T}(1-\bar{T})}\right)} \\ &= M_{J-2}\sqrt{\left(\frac{1}{\pi(1-\pi)J}\right)\left(\frac{(1-\rho_C)(1-R_{C(within)}^2)}{n\bar{T}(1-\bar{T})}\right)} \\ &\neq 0 \end{aligned} \tag{21}$$

Thus even though the MDESD is positive due to individual-level sampling error ($\frac{(1-\rho_C)\left(1-R^2_{C(within)}\right)}{n\bar{T}(1-\bar{T})}$), the maximum possible subgroup impact difference is zero. *Consequently, the minimum detectable value of* Δ_* *is larger than its maximum possible value.*

Why Can We Not Use Equation 19 to Compute an MDESD? Equation 19 implies that:

$$se\left(\hat{\Delta}_*\right) = \sqrt{\left(\frac{1}{\pi(1-\pi)J}\right)\left((1-R^2_W)\tau^2_* + \frac{(1-\rho_C)\left(1-R^2_{C(within)}\right)}{n\bar{T}(1-\bar{T})}\right)} \tag{22}$$

In addition, note that rearranging terms in Equation 17 yields:

$$R^2_W = \frac{\Delta^2_*\pi(1-\pi)}{\tau^2_*} \tag{23}$$

Substituting Equation 23 into Equation 22 yields:

$$\begin{aligned} se\left(\hat{\Delta}_*\right) &= \sqrt{\left(\frac{1}{\pi(1-\pi)J}\right)\left((1-\frac{\Delta^2_*\pi(1-\pi)}{\tau^2_*})\tau^2_* + \frac{(1-\rho_C)\left(1-R^2_{C(within)}\right)}{n\bar{T}(1-\bar{T})}\right)} \\ &= \sqrt{\left(\frac{1}{\pi(1-\pi)J}\right)\left(\tau^2_* - \Delta^2_*\pi(1-\pi) + \frac{(1-\rho_C)\left(1-R^2_{C(within)}\right)}{n\bar{T}(1-\bar{T})}\right)} \end{aligned} \tag{24}$$

Equation 24 illustrates the fundamental problem that we face when trying to compute an MDESD from Equation 19: *the fact that the standard error of our estimator,* $se(\hat{\Delta}_*)$*, depends on the value of our estimand (*Δ_**).*

So How Can We Determine an MDESD? In light of the preceding problem, we recommend the following six-step process to determine an MDESD.

- Step 1: Given τ^2_* and π, determine the maximum possible value of Δ_*. This is the value of Δ_* when $R^2_W = 1$. Thus $\Delta_{*(max)} = \sqrt{\frac{\tau^2_*}{\pi(1-\pi)}}$.

- Step 2: Given J, n, π, τ_*^2, ρ_C, $R^2_{C(within)}$ and the fact that $R^2_W = 1$, compute power for the alternative hypothesis that the true value of Δ_* equals $\Delta_{*(max)}$, where.

  $Power = 1 - \Phi\left(t_{critical} - \frac{\Delta_*}{se(\hat{\Delta}_*|R^2_W)}\right)$,

  Φ is the cumulative t distribution with J-2 degrees of freedom and

  $se\left(\hat{\Delta}_{*|R^2_W}\right) = \sqrt{\left(\frac{1}{\pi(1-\pi)J}\right)\left((1-R^2_W)\tau_*^2 + \frac{(1-\rho_C)\left(1-R^2_{C(within)}\right)}{n\bar{T}(1-\bar{T})}\right)}$

- Step 3: If power to detect $\Delta_{*(max)}$ is less than 0.80, there is no MDESD for the assumed sample design and data structure. So stop here! If power to detect $\Delta_{*(max)}$ is greater than 0.80 continue to step 4.
- Step 4: Try a new value of Δ_* that is less than the previous value and compute $R^2_W = \frac{\Delta_*^2\pi(1-\pi)}{\tau_*^2}$.
- Step 5: Calculate power for the new trial value of R^2_W, as in step 2.
- Step 6: If power is greater than 0.80, try a smaller value of Δ_* (and thus R^2_W) and repeat steps 4 and 5. If power is less than 0.80, try a larger value of Δ_* (and thus R^2_W) and repeat steps 4 and 5. Continue until your trial value for Δ_* (and thus R^2_W) has power that is "close enough" to 0.80 to stop.

Illustrative Findings: Table 3 reports MDESD findings for the scenarios that are reflected for the MDES and MDESSD in Tables 1 and 2. In addition, to determine an MDESD it is necessary to specify a value for π, which we set at 0.60.

Two findings are reported in each cell of Table 3. The first is the MDESD. Note that each MDESD in Table 3 is roughly twice the size of its counterpart MDES in Table 1. This represents the well-known fact that precision is much greater for estimates of a full sample mean than for

corresponding estimates of a difference between two subsample means. For example, the MDESD in Table 3 is 0.27σ for a study with 50 sites and 20 sample members per site versus 0.14σ for an MDES in Table 1. This "fact of statistical life" is doubly problematic because: (1) not only does it take a much larger sample to attain a given degree of precision for a subsample difference of means than for a full-sample mean, but often the difference between mean subgroup impacts is smaller (potentially by a lot) than the mean full-sample impact.

Second, many cells in Table 3 have no values because they do not have a feasible MDESD. This occurs because the minimum detectable effect size difference is larger than the maximum possible effect size difference. Thus it is not possible to attain 80 percent power for a feasible value of Δ_*.

Third, the tradeoff between the number and sample size of sites (J and n) is the same in Table 3 as in Table 1 because both tables examine statistical properties of cross-site means. Specifically, increasing the number of sites by a given percentage improves precision by more than increasing site sample size by the same percentage. This can be seen most clearly by comparing findings in the shaded cells along the diagonal of the tables.

The second finding reported for each cell in Table 3 (in parentheses) is the value of R_W^2 that corresponds to the minimum detectable value of Δ_*. Recall that these two parameters are two sides of the same coin – in that the value of one determines the value of the other, given τ_*^2 and π (see Equation 3). Hence the values for R_W^2 in Table 3 represent the minimum predictive power of our site-level impact predictor (W) that is needed to detect a non-zero value of Δ_*.

Insert Table 3 about here

Precision for Multisite Cluster Randomized Trials

This section extends the preceding discussion for multisite trials to that for multisite cluster-randomized trials.

Estimation Models

Assume that we have *J* sites, *m* clusters per site, *n* sample members per cluster and proportion $\bar{T}$ of the clusters from each site randomized to treatment. Once again, our estimands are: the cross-site *mean* program effect size (β_*), the cross-site *standard deviation* of program effect sizes (τ_*) and (3) the *difference* in cross-site mean program effect sizes for two subgroups of sites ($\Delta_* = \beta_{*(II)} - \beta_{*(II)}$). Consequently, we want to assess a minimum detectable effect size (MDES), a minimum detectable effect size standard deviation (MDESSD) and a minimum detectable effect size difference (MDESD) for multisite cluster randomized trials.

The impact estimation model for an MSCRT can be written as a three-level model with fixed site-specific intercepts (α_j), random site-specific impact coefficients (B_j) and fixed coefficients (θ_k) for cluster-level baseline covariates (X_k). Note that in cluster randomized trials, a cluster-level covariate (such as a lagged school-level mean test score) is often used to increase precision because: (1) such covariates explain variation in mean outcomes across clusters, which is usually the primary limitation on precision for cluster randomized designs (Raudenbush, 1997 and Bloom, 2005), (2) they can have substantial explanatory power (Bloom et al., 2007) and (3) data for them can be easy to obtain.

Level One: Individuals

$$Y_{imj} = \alpha_{mj} + e_{imj} \quad (25)$$

Level Two: Clusters

$$\alpha_{mj} = \alpha_j + B_j T_{mj} + \sum_{k=1}^{K} \theta_k X_{kmj} + r_{mj} \quad (26)$$

Level Three:

$$\alpha_j = \alpha_j \tag{27}$$

$$B_j = \beta + b_j \tag{28}$$

where $e_{imj} \sim N\left(0, \sigma^2_{|\alpha_{mj}}\right)$, $r_{mj} \sim N\left(0, \varphi^2_{|X\alpha_j}\right)$ and $b_j \sim N(0, \tau^2)$, Y_{imj} is the observed value of the outcome for individual *i* from cluster *m* in site *j*; T_{mj} equals one if cluster *m* in site *j* was randomized to treatment and zero otherwise; X_{kmj} is the value of baseline covariate *k* for cluster *m* from site *j*; θ_k is a fixed coefficient for cluster-level covariate *k*; α_{mj} is the mean outcome for cluster *m* in site *j;* α_j is the conditional population mean control group outcome for site *j*, which is fixed for each site; B_j is the population treatment effect for site *j*, which varies randomly across sites; β is the cross-site mean treatment effect for the population of sites; e_{imj} is a random error that varies independently and identically across individuals within clusters, with a mean of zero and variance $\sigma^2_{|\alpha_{mj}}$; r_{mj} is a random error that varies independently and identically across clusters within sites and experimental conditions, with a mean of zero and variance $\varphi^2_{|X\alpha_j}$; b_j is a random error that varies independently and identically across sites with a mean of zero and variance τ^2.

As we did for an MST, we can estimate a *difference* in cross-site mean program effects for two subgroups of sites from an MSCRT by adding a binary site subgroup indicator, W_j to the site-level model (Equation 28), which then becomes:

$$B_j = \beta_I + \Delta W_j + b'_j \tag{29}$$

where β_I is the cross-site mean program effect for site subgroup I, W_j equals one if site j is from subgroup *II* and zero otherwise and $b'_j \sim N(0, \tau^2_{|W})$.

A Minimum Detectable Effect Size

The minimum detectable effect (MDE) for our multisite cluster randomized trial (MSCRT) is:

$$MDE = M_{J-1}\sqrt{\left(\frac{1}{J}\right)\left(\tau^2 + \frac{\varphi^2_{|X\alpha_j}}{m\bar{T}(1-\bar{T})} + \frac{\sigma^2_{|\alpha_{mj}}}{mn\bar{T}(1-\bar{T})}\right)} \quad (30)$$

where all terms have already been defined. Note that this MDE contains three variance components. The first variance component, τ^2, reflects random error due to the sampling of study sites from a population of sites. The next two variance components, $\varphi^2_{|X\alpha_j}$, and $\sigma^2_{|\alpha_{mj}}$, reflect random error due to the sampling of clusters within sites and individuals within clusters, respectively.

The two within-site variance components can be re-expressed in terms of the *total* variance of control group outcomes σ^2_C, the proportion of this variance that is between sites (the control group intra-class correlation at the site level), ρ_{CS}, the proportion of the remaining control-group outcome variance that is between clusters within sites (the control-group intra-class correlation at the cluster level), ρ_{CC}, and the proportion of the between-cluster within-site outcome variance for control group members (R^2_{CC}) that is explained by cluster-level covariates (*X*), where:

$$\sigma^2_{|\alpha_{mj}} = (1 - \rho_{CS} - \rho_{CC})\sigma^2_C \quad (31)$$

$$\varphi^2_{|X\alpha_j} = \rho_{CC}(1 - R^2_{CC})\sigma^2_C \quad (32)$$

Similar to the MST, expressing variance components for the MSCRT in terms of the total control group outcome variance allows one to use the existing empirical literature to guesstimate the proportion of the total control group outcome variance that is at different levels (ρ_{CS}, ρ_{CC}) and the proportion of the cluster-level outcome variance that is explained by baseline covariates, R^2_{CC} . The re-formulated MDE is:

$$MDE = M_{J-1}\sqrt{\left(\frac{1}{J}\right)\left(\tau^2 + \frac{\rho_{CC}(1-R^2_{CC})\sigma^2_C}{m\bar{T}(1-\bar{T})} + \frac{(1-\rho_{CS}-\rho_{CC})\sigma^2_C}{mn\bar{T}(1-\bar{T})}\right)} \quad (33)$$

We can also translate the outcome measure (Y_{ijk}) in its original units to a standardized outcome measure, $Z_{ijk} = \frac{Y_{ijk}-\mu_C}{\sigma_C}$ that is defined in terms of control group population parameters, μ_C, and σ_C. We thus define the resulting standardized cross-site mean effect size as $\beta_* = \frac{\beta}{\sigma_C}$ and the resulting cross-site standard deviation of effects sizes as $\tau_* = \frac{\tau}{\sigma_C}$. Dividing the MDE by the control group standard deviation, σ_C, yields an MDES of:

$$MDES = M_{J-1}\sqrt{\left(\frac{1}{J}\right)\left(\tau^2_* + \frac{\rho_{CC}(1-R^2_{CC})}{m\bar{T}(1-\bar{T})} + \frac{(1-\rho_{CS}-\rho_{CC})}{mn\bar{T}(1-\bar{T})}\right)} \quad (34)$$

Much like Equation 15 for an MST, Equation 34 for an MSCRT suggests that the influence of sample size at the various levels is not the same. Specifically, the number of sites has the greatest impact on precision, followed by the number of clusters per site, followed by the number of individuals per cluster. This can be seen by examining the influence of each sample size parameter on the three variance components in the MDES.

Illustrative Findings: Suppose we are designing a study with eighth-grade students in middle schools that are randomized within districts to a science education intervention or a control group, with subsequent science test scores as the outcome of interest. Thus, ρ_{CS} is the proportion of the total variance of future science test scores for control group members that is between districts, ρ_{CC} is the proportion of the variance of future science test scores for control group members within districts that is between schools, and R^2_{CC} is the proportion of the variance of mean future science test scores across schools in districts that is explained by our school-level baseline covariate, which we assume is each school's mean science test score in the previous

year. Lastly, note that τ_* is the *cross-district* standard deviation of program effect sizes on science test scores.

We can obtain approximate values for the first three parameters from recent empirical research (e.g. Hedges & Hedberg, 2013; Spybrook, Westine, & Taylor, 2014; Westine, Spybrook, & Taylor, 2013). For example, Spybrook, Westine, & Taylor (2014) provide relevant empirical findings for eighth-grade science achievement based on administrative data from three states. Their estimates of ρ_{CS} range from 0.04 to 0.12, with a mean of 0.07. Their estimates of ρ_{CC} range from 0.10 to 0.11, with a mean of 0.10. Their estimates of R^2_{CC} for a baseline covariate that is the previous year's school-level mean science test score ranges from 0.64 to 0.82, with a mean of 0.74. To our knowledge, there is no existing empirical information about the value of τ_*, when it is defined as the standard deviation of program effects across school districts. So to guesstimate this parameter, we can start with findings from Weiss et al. (under review) about the standard deviation of education program effects across schools. However, we would anticipate that there would be less variability in program effects across districts than across schools. Hence we assume that $\tau_* = 0.10\sigma$.

Table 4 presents the MDES for scenarios that comprise different numbers of sites and clusters per site. These scenarios assume that $\bar{T} = 0.5, \tau_* = 0.10\sigma_C,\ \rho_{CS} = 0.07, \rho_{CC} = 0.10, R^2_{CC} = 0.74$, *n*=200, alpha = 0.05 for a two-tailed test and power = 0.80. Note that the number of sites (school districts) for the MSCRT in Table 4 is much smaller than the number of sites (schools) for the MST in Table 1. This is because the often high costs of MSCRTs make it necessary to minimize the number of sites involved. For example, the National Center for Educational Evaluation and Regional Assistance recently funded two major MSCRTs of math and reading curricula (Agodini et al., 2010; James-Burdumy et al., 2009). The sample for each

study included about 10 school districts (sites) with approximately 10 schools (clusters) per district, for a total of 100 schools. Both studies were large and costly undertakings.

Insert Table 4 about here

The total sample size is constant for the two light gray cells and the two dark gray cells along the shaded diagonal in Table 4. At one extreme on this diagonal there are 6 sites with 20 clusters per site; at the other extreme, there are 20 sites with 6 clusters per site. As can be seen, the MDES decreases as we move from a small number of sites with a large number of clusters each to a large number of sites with a small number of clusters each. Thus allocating resources to ensure more sites rather than more clusters per site can improve the power of an MSCRT to detect a cross-site mean program effect size.

A Minimum Detectable Effect Size Standard Deviation

The minimum detectable effect size standard deviation (MDESSD) for a multisite cluster-randomized trial (MSCRT) can be expressed as:

$$MDESSD = \sqrt{\left(\frac{\rho_{CC}\left(1-R^2_{CC}\right)}{m\bar{T}(1-\bar{T})} + \frac{(1-\rho_{CS}-\rho_{CC})}{mn\bar{T}(1-\bar{T})}\right)}\left(\frac{F_{0.05}}{F_{0.80}} - 1\right) \quad (35)$$

where *n*, *m*, $\bar{T}$, ρ_{CC}, ρ_{CS}, R^2_{CC} were defined previously, $F_{0.05}$ is the critical value of an F statistic with J-1 and J(m-2)-K degrees of freedom in its numerator and denominator, respectively, and $F_{0.80}$ is the value of a corresponding F statistic that is below 80 percent of its distribution.

Previous findings for the MST suggest that for the MSCRT the number of clusters per site has the most influence on the MDESSD; and as can be seen from Equation 35, this is the case. Specifically, the number of clusters per site, *m*, reduces the uncertainty produced both by the outcome variance across clusters within sites and the outcome variance across individuals within clusters. In contrast, the number of individuals per cluster, *n*, only reduces the uncertainty

produced by the outcome variance across individuals within clusters. The number of sites only influences the MDESSD through the number of degrees of freedom for $F_{0.05}$ and $F_{0.80}$.

Illustrative Findings: Table 5 reports the MDESSD for the parameter assumptions and sample-size combinations represented by Table 4 for the MDES. As expected, the MDESSD is smaller for a design with 6 sites and 20 clusters per site (MDESSD = 0.15σ) than for a design with 20 sites and 6 clusters per site (MDESSD=0.17σ). However, the difference is small. This is partly because we assumed that only a small portion of the within-site control-group outcome variance is across clusters ($\rho_{CC} = 0.10$). In addition, we assumed that much of this small cross-cluster variance is explained by the cluster-level baseline covariate ($R^2_{CC} = 0.74$). Lastly, Table 5 has a limited range of values for *J* and *m* relative to the much larger range of values for *J* and *n* in Table 2 for an MST. This is because it is very uncommon to see an MSCRT with a large number of sites.

Insert Table 5 about here

A Minimum Detectable Effect Size Difference

We turn now to the precision of an MSCRT for estimating a difference in mean program effect sizes for two subgroups of sites, ($\Delta_* = \frac{\Delta}{\sigma_C} = \frac{\beta_{II}}{\sigma_C} - \frac{\beta_I}{\sigma_C}$) where Δ is the coefficient for W_j in Equation 29. In other words, we turn to the minimum detectable effect size difference or MDESD. Once again, note that there is no closed form expression for this parameter because the standard error of the estimator, $se(\hat{\Delta}_*)$, depends on the value of the estimand, Δ_*. Hence the six-step procedure outlined for an MST should also be used for an MSCRT. The only difference is with respect to calculating power in Step 2 because the standard error of our parameter estimator ($\hat{\Delta}_*$) is:

$$se(\hat{\Delta}_*) = \sqrt{\left(\frac{1}{\pi(1-\pi)J}\right)\left((1-R_W^2)\tau_*^2 + \frac{\rho_{CC}(1-R_{CC}^2)}{m\bar{T}(1-\bar{T})} + \frac{(1-\rho_{CS}-\rho_{CC})}{mn\bar{T}(1-\bar{T})}\right)} \quad (36)$$

where π is the proportion of sites in each subgroup, $R_{|W}^2$ is the explanatory power of the site-level impact predictor (W), and the remaining terms were defined previously. In terms of sample size, we see the same pattern as for the MDES. The total number of sites is the most influential sample size, followed by the number of clusters per site, and the number of individuals per cluster.

Illustrative Findings: Table 6 reports MDESDs for the parameter assumptions in Tables 4 and 5 plus the additional assumption that $\pi = 0.60$. Dashes in Table 6 represent sample designs that have no MDESD because their maximum possible effect size difference is less than their minimum detectable effect size difference. Table 6 also reports (in parentheses) the value of R_W^2 for each MDESD. Note that like the multisite randomized trial, the MDESD for a multisite cluster randomized trial is nearly twice the size of its corresponding MDES. Similarly, the number of sites for a multisite cluster-randomized trial is the sample size with the greatest influence on precision for the MDESD.

Insert Table 6 about here

Implications for Research Practice

Multisite randomized studies offer an exciting opportunity to address a series of important questions about the effectiveness of educational and related programs. For example, they make it possible to assess the "bottom line" of a program's effectiveness by identifying its cross-site mean effect. In addition, they make it possible to assess the variability of a program's effectiveness by identifying the cross-site standard deviation of its effects. They also make it possible to explore site-level factors that influence a program's effectiveness, by identifying differences that exist between mean program effects for subgroups of sites with specific

characteristics. However to capitalize on this opportunity requires adequately powering future studies to detect these three features of cross-site impact distributions.

Toward this end, the present paper provides simple expressions and procedures for assessing the minimum detectable effect size (MDES), the minimum detectable effect size standard deviation (MDESSD), and the minimum detectable effect size difference (MDESD) of a multisite randomized trial and a multisite cluster-randomized trial. These diagnostics are a function of sample design parameters, which must be chosen by researchers who are planning a study. In addition, they also depend on features of the data to be used (e.g. the outcome measure of interest), the student population to be studied (e.g. their grade level) and the educational environment in which a study will be conducted (e.g. regular public schools, charter schools and/or private schools), which are determined by the research questions to be addressed.

To provide practical guidance for dealing with these issues, we have tried to: (1) develop intuition about the factors that influence precision, why they influence precision, and how their influence differs for different estimands; (2) point researchers to relevant empirical sources that can help them determine the parameter values that must be assumed in order to assess the precision of alternative research designs; and (3) illustrate the tradeoffs that exist when doing so.

Four main messages emerge from an examination of our findings.

1. *The number of sites* is the sample size with the most influence on a minimum detectable effect size (MDES) or a minimum detectable effect size difference (MDESD).
2. *The number of individuals per site* (for a multisite randomized trial) and the *number of clusters per site* (for a multisite cluster-randomized trial) are the sample sizes with

the most influence on a minimum detectable effect size standard deviation (MDESSD).

3. In general, *large studies* are needed to provide adequate power for rigorously studying cross-site impact variation. This is especially true for studying site-subgroup differences in mean program effects.
4. In general, because of their large sample size requirements, it probably will be difficult to implement *multisite cluster randomized trials that randomize schools within school districts* with adequate power for studying cross-site impact variation.

The first two points highlight the tradeoff that exists between designing a study to estimate mean program effects or site-subgroup mean impact differences and designing a study to estimate cross-site impact variation. In short, designing a study with adequate power for one of these estimands does not ensure adequate power for the others. This fact was demonstrated both by comparing expressions for the MDES, MDESSD and MDESD and by comparing illustrative findings for them.

The third point highlights the need for large studies in order to rigorously examine cross-site impact variation. Consider first the size of a multisite randomized trial that is needed to detect with confidence a cross-site mean program effect of 0.15σ, which by many standards would be considered a successful intervention (e.g. Bloom et al., 2008 and Hill et al., 2008). Findings in Table 1 suggest that with 20 sites, this would require between 50 and 100 sample members per site or between 1,000 and 2,000 total sample members.

Consider next the size of a multisite randomized trial that would be needed to detect with confidence a cross-site standard deviation of program effects equal to 0.15σ, which is substantial. For example, assuming approximately normally, this implies that 95 percent of the

sites for an intervention would have a program effect size that was within $\pm$ 0.30σ of the cross-site mean – quite a wide range of possible values. Findings in Table 2 suggest that 100 sample members at each of 20 sites or 200 sample members at each of 10 sites (2,000 total sample members) would be needed to meet this objective.

Lastly, consider the size of a multisite randomized trial that would be needed to detect with confidence a site subgroup difference in program effect sizes equal to 0.15σ, which is substantial, unless program effects are near zero or negative for one of the two subgroups. Table 3 suggests that 50 sites with between 100 and 200 persons per site (5,000 to 10,000 total sample members) would be needed to meet this objective.

Of course the preceding findings are based on specific assumptions and they will differ as these assumptions differ. Therefore a detailed analysis is required to judge the sample size requirements and thus the feasibility and cost of any given study. Nonetheless, we believe that the present findings might represent a broad range of potential situations. Fortunately, at least with respect to estimating a cross-site mean and standard deviation of program effects for an MST, a number of studies of adequate scale have been conducted to date, as demonstrated by findings from Weiss et al. (under review). For specific studies that are not large enough, one way to provide samples with adequate power for studying cross-site impact variation is to pool data across studies with comparable interventions, student populations and outcome measures. One precedent for this approach is the study of variation in the effects of welfare-to-work programs conducted by Bloom et. al. (2003), who pooled findings from three major evaluations conducted over the course of a decade by a single organization (MDRC).[13]

[13] The three studies pooled were: (1) an evaluation of the Greater Avenues for Improvement Program conducted in 22 local welfare offices (sites) in California (Riccio and Friedlander, 1992), (2) an evaluation of Project Independence conducted in 10 local welfare offices (sites) in Florida (Kemple and Haimson, 1994), and (3) the

The fourth point noted above highlights the challenge that exists with respect to studying cross-site impact variation using trials that randomize schools (clusters) to experimental conditions within districts (sites).[14] Not only does this approach require especially large samples to detect cross-site impact variation of a given magnitude, but it is also likely that the impact variation which exists across districts is smaller than that which exists across organizational units like schools. But even if the cross-site standard deviation of program effects were only 0.10σ, (which implies a cross-site impact variance of 0.01σ assuming approximately normality, this implies that 95 percent of the sites for an intervention would have an effect size that was within $\pm$ 0.20σ of the cross-site mean, which reflect a substantial range. This problem becomes even harder to overcome when designing a multisite cluster randomized trial to detect meaningful effect size *differences* across subgroups of sites.

Consequently, pooling data across studies probably will be critical for any research on cross-site impact variation that uses multi-site cluster randomized trials. And to successfully combine these studies it will be essential to keep this goal clearly in mind when planning them. Only in this way will it be possible to collect common measures and select sufficiently heterogeneous sites across a given group of studies. Perhaps in some cases, this level of coordination might be possible through large scale federal funding programs like the Investing in Innovation Fund (http://www2.ed.gov/programs/innovation/index.html).

National Evaluation of Welfare-to-Work programs conducted in 27 local welfare offices (sites) from six states (Hamilton, 2002).

[14] In the future, it might be possible to study cross-site impact variation using MSCRTs that randomize intact classrooms (clusters) within schools (sites). However to date, very few such studies have been conducted. And even these studies would need to be quite large because of the typically small numbers of classrooms per school – especially for elementary or middle schools.

Nonetheless, we believe that multisite randomized trials have a great potential for enabling researchers, policy makers and practitioners to learn much more "about and from" cross-site variation in the effects of educational and related programs. In this way, we hope that it will soon become possible to better understand what works for whom and under what conditions. But in order for this potential to be realized it will be essential to continue to improve future research designs, measurement tools, data collection strategies and statistical analysis methods. Toward this end, we hope that the present paper will contribute to the first of these objectives.

REFERENCES

Agodini, R., Harris, B., Thomas, M., Murphy, R., Gallagher, L., Pendleton, A. (2010). *Achievement effects of four early elementary school math curricula* (NCEE 2011-4001). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Angrist, J. D., Pathak, P. A., & Walters, C. R. (2013). Explaining charter school effectiveness. *American Economic Journal. Applied Economics, 5*(4), 1-27.

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

Bloom, H. S. (2005). Randomizing groups to evaluate place-based programs. In H.S. Bloom (Ed), *Learning More from Social Experiments: Evolving Analytic Approaches* (115-172). New York: Russell Sage Foundation.

Bloom, H. S., Hill, C. J., & Riccio, J. A. (2003). Linking program implementation and effectiveness: Lessons from a pooled sample of welfare-to- work experiments. *Journal of Policy Analysis and Management, 22*(4), 551-575.

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.

Bloom, H.S., Raudenbush, S.W., Weiss, M. & Porter, K. (revise and resubmit). Using multi-site experiments to study cross-site variation in effects of program assignment.

Bloom, H., Hill, C., Black, A. R., & Lipsey, M. (2008). Performance trajectories and performance gaps as achievement effect- size benchmarks for educational interventions. *Journal of Research on Educational Effectiveness, 1*(4), 289-328.

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

Clark, M. A., Isenberg, E., Liu, A. Y., Makowsky, L., and Zukiewicz, M. (2015). Impacts of the Teach For America Investing in Innovation Scale-Up. Princeton, NJ: Mathematica Policy Research.

Dong, N., & Maynard, R. A. (2013). "PowerUp"!: A tool for calculating minimum detectable effect size differences and minimum required sample sizes for experimental and quasi-experimental design studies. *Journal of Research on Educational Effectiveness, 6(1),* 24-67. doi: 10.1080/19345747.2012.673143.

Gleason, P., Clark, M., Tuttle, C. C., & Dwoyer, E. (2010). *The Evaluation of Charter School Impacts: Final Report* (NCEE 2010-4029). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Hamilton, G. (2002) *Moving People from Welfare to Work: Lessons from the National Evaluation of Welfare-to-Work Strategies*. Washington, DC: U.S. Department of Health and Human Services. Administration for Children and Families, Office of the Assistant Secretary for Planning and Evaluation & U.S. Department of Education, Office of the Under Secretary, Office of Vocational and Adult Education.

Hays, W. L. (1973) *Statistics for the social sciences* (2d ed). New York: Holt, Rinehart and Winston.

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*(6), 445-489.

Hedges, L. V., & Pigott, T. (2001). The power of statistical tests in meta- analysis. *Psychological Methods; Psychological Methods, 6*(3), 203-217.

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., & Spybrook, J. (2013). Intraclass correlations and covariate outcome correlations for planning two- and three- level cluster- randomized experiments in education. *Evaluation Review, 37*(6), 445-489.

Hedges, L. V., & Pigott, T. D. (2004). The power of statistical tests for moderators in meta-analysis. *Psychological Methods, 9*(4), 426.

Hedges, L. V. & Rhoads, C. (2010). *Statistical power analysis in education research* (NCSER 2010-3006). Washington, DC: National Center for Special Education Research, Institute of Education Sciences, U.S. Department of Education.

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.

Jacob, R., Zhu, P., & Bloom, H. (2010). New empirical evidence for the design of group randomized trials in education. *Journal of Research on Educational Effectiveness, 3*(2), 157-198.

James-Burdumy, S., & National Center for Education Evaluation and Regional Assistance. (2009). *Effectiveness of selected supplemental reading comprehension interventions: Impacts on a first cohort of fifth-grade students.* Washington, D.C.: National Center for

Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Kemple, J. J., & Haimson, J. (1994). *Florida's Project Independence: Program Implementation, Participation Patterns, and First-year Impacts*, New York: MDRC.

Konstantopoulos, S. (2008). Computing power of tests of the variance of treatment effects in designs with two levels of nesting. *Multivariate Behavioral Research, 43*(2), 327-352.

Liu, X. S., (2014). *Statistical power analysis for the social and behavioral sciences: Basic and advanced techniques*. New York: Routledge.

Morebeek, M., & Teerenstra, S. (2016). *Power analysis of trials with multilevel data*. Boca Raton: CRC Press.

Olsen, R., Bein, E., & Judkins, D. (2015). *Sample size requirements for education multi-site RCTs that select sites randomly*. Paper presented at 37th Annual Conference of the Association for Public Policy Analysis and Management, Miami, FL.

Peck, L. R., & Bell, S. H. (2014). *The Role of Program Quality in Determining Head Start's Impact on Child Development*, OPRE Report #2014-10. Washington DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

Puma, M., Bell, S., Cook, R., Heid, C., Shapiro, G., Broene, P., et al. (2010). Head start impact study. Final report. Washington, DC: U.S. Department of Health and Human Services, Administration for Children & Families.

Raudenbush, S. W., & Bloom, H. S. (2015). Learning about and from a distribution of program impacts using multisite trials. *American Journal of Evaluation*, *36*(4), 475-499.

Raudenbush, S. W., & Liu, X. (2000). Statistical power and optimal design for multisite randomized trials. *Psychological Methods, 5*(2), 199-213.

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

Riccio, J. A., & Friedlander, D. (1992). *GAIN: Program strategies, participation patterns, and first-year impacts in six counties: Executive summary.* New York: MDRC.

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

Spybrook, J. (2014). Detecting intervention effects across context: An examination of the precision of cluster randomized trials. *The Journal of Experimental Education, 82*(3), 334-357.

Spybrook, J., & Raudenbush, S. W. (2009). An examination of the precision and technical accuracy of the first wave of group- randomized trials funded by the institute of education sciences. *Educational Evaluation and Policy Analysis, 31*(3), 298-318.

Spybrook, J., Shi, R., & Kelcey, B. (2016). Progress in the past decade: An examination of the precision of cluster randomized trials funded by the U.S. institute of education sciences. *International Journal of Research & Method in Education, 39*(3), 1-13.

Spybrook, J., Westine, C. D., & Taylor, J. A. (2016). Design parameters for impact research in science education: A multi-state analysis. *AERA Open, 2*(1), doi: 10.1177/2332858415625975.

Weiss et. al. (under review) How Much Do the Effects of Education and Training Programs Vary Across Sites? Evidence from Existing Multisite Randomized Trials.

Weiss, M. J., Bloom, H. S., & Brock, T. (2014). A conceptual framework for studying the sources of variation in program effects. *Journal of Policy Analysis and Management 33*(3), 778-808.

Westine, C., Spybrook, J., & Taylor, J. (2013). An empirical investigation of design parameters for planning cluster randomized trials of science achievement. *Evaluation Review, 37*(6), 490-519.

Word, E., Johnston, J., Bain, H., Fulton, D.B., Boyd-Zaharias, J., Lintz, M.N., Achilles, C.M., Folger, J., & Breda, C. (1990). *Student/teacher achievement ratio (STAR): Tennessee's K-3 class-size study*. Nashville, TN: Tennessee State Department of Education.

Zhu, P., Jacob, R., Bloom, H., & Xu, Z. (2012). Designing and analyzing studies that randomize schools to estimate intervention effects on student academic outcomes without classroom-level information. *Educational Evaluation and Policy Analysis, 34*(1), 45-68.

Table 1. MDES for a cross-site mean program effect size for a randomized multisite trial

Number of Individuals per Site (*n*)	Number of Sites (*J*)					
	5	10	20	50	100	200
5	1.10	0.65	0.43	0.27	0.19	0.13
10	0.80	0.47	0.31	0.19	0.14	0.10
20	0.59	0.35	0.23	0.14	0.10	0.07
50	0.42	0.25	0.17	0.10	0.07	0.05
100	0.35	0.21	0.14	0.08	0.06	0.04
200	0.30	0.18	0.12	0.07	0.051	0.04
500	0.27	0.16	0.11	0.07	0.05	0.03

NOTES: Values in the table are for two-tail significance = 0.05, power = 80 percent, a single level-one baseline covariate, $R^2_{C(within)} = 0.4$, $\bar{T} = 0.5$, , $\rho_C = 0.15$, constant *n* within sites, and $\tau_* = 0.15$ (or $\tau_*^2 = 0.0225$).

Table 2. MDESSD for a cross-site standard deviation of program effect sizes for a randomized multisite trial

Number of Individuals per Site (n)	Number of Sites (J)					
	5	10	20	50	100	200
5	1.65	1.07	0.78	0.57	0.45	0.37
10	1.05	0.70	0.52	0.38	0.30	0.35
20	0.72	0.48	0.36	0.26	0.21	0.17
50	0.45	0.30	0.22	0.16	0.13	0.11
100	0.31	0.21	0.16	0.11	0.09	0.08
200	0.22	0.15	0.11	0.08	0.07	0.05
500	0.14	0.09	0.07	0.05	0.03	0.03

NOTES: Values in the table are for two-tail significance = 0.05, power = 80 percent, a single level-one baseline covariate, $R^2_{C(within)} = 0.4$, $\bar{T} = 0.5$, , $\rho_C = 0.15$, and constant n within sites.

Table 3. MDESD for the coefficient on a binary site-level moderator for a randomized multisite trial

Number of Individuals per Site (*n*)	Number of Sites (*J*)					
	5	10	20	50	100	200
5	--	--	--	--	--	0.26 (0.72)
10	--	--	--	--	0.26 (0.72)	0.19 (0.39)
20	--	--	--	0.27 (0.78)	0.20 (0.43)	0.14 (0.21)
50	--		0.28 (0.84)	0.19 (0.39)	0.14 (0.21)	0.10 (0.11)
100	--	0.30 (0.96)	0.23 (0.56)	0.16 (0.27)	0.12 (0.15)	0.08 (0.07)
200	--	0.26 (0.72)	0.20 (0.43)	0.14 (0.21)	0.10 (0.11)	0.07 (0.05)
500	0.30 (0.96)	0.24 (0.61)	0.18 (0.35)	0.13 (0.18)	0.09 (0.09)	0.07 (0.05)

NOTES: Values in the table are for two-tail significance = 0.05, power = 80 percent, a single level-one baseline covariate, $R^2_{C(within)} = 0.4$, $\bar{T} = 0.5$, , $\rho_C = 0.15$, constant *n* within sites, and $\tau_* = 0.15$ (or $\tau_*^2 = 0.0225$), and $\pi = 0.60$. Values in parentheses are the R-square that corresponds to each minimum detectable effect size difference.

Table 4. MDES for a cross-site mean program effect size for a multisite cluster randomized trial

Number of Clusters per Site (m)	Number of Sites (J)					
	4	6	8	10	12	20
4	0.43	0.29	0.23	0.20	0.18	0.13
6	0.37	0.25	0.20	0.17	0.15	0.11
8	0.34	0.23	0.18	0.16	0.14	0.10
10	0.32	0.21	0.17	0.15	0.13	0.10
12	0.30	0.20	0.16	0.14	0.13	0.09
20	0.27	0.18	0.15	0.13	0.11	0.08

NOTES: Values in the table are for two-tail significance = 0.05, power = 80 percent, no level-one covariates, a single level-two covariate, $R^2_{CC} = 0.74$, $\bar{T} = 0.5$, , $\rho_{CS} = 0.07$, $\rho_{CC} = 0.10$, constant *n=200* within each cluster, constant *J* within each site, and $\tau_* = 0.10$ (or $\tau^2_* = 0.01$).

Table 5. MDESSD for a cross-site standard deviation of program effect sizes for a multisite cluster randomized trial

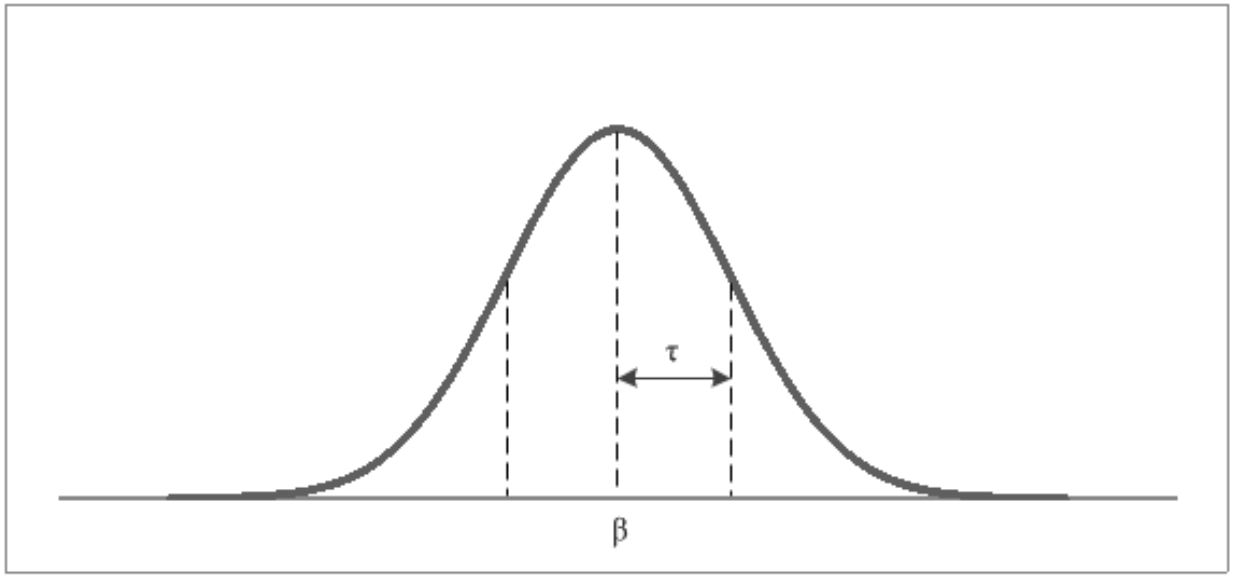
Number of Clusters per Site (m)	Number of Sites (J)					
	4	6	8	10	12	20
4	0.60	0.43	0.35	0.31	0.28	0.23
6	0.41	0.31	0.26	0.23	0.21	0.17
8	0.35	0.26	0.22	0.19	0.18	0.14
10	0.31	0.23	0.19	0.17	0.16	0.13
12	0.27	0.20	0.17	0.15	0.14	0.11
20	0.21	0.15	0.13	0.12	0.11	0.09

NOTES: Values in the table are for two-tail significance = 0.05, power = 80 percent, no level-one covariates, a single level-two covariate, $R^2_{CC} = 0.74$, $\bar{T} = 0.5$, , $\rho_{CS} = 0.07$, $\rho_{CC} = 0.10$, constant *n=200* within each cluster, and constant *J* within each site.

Table 6. MDESD for the coefficient on a binary site-level moderator for a multisite cluster randomized trial

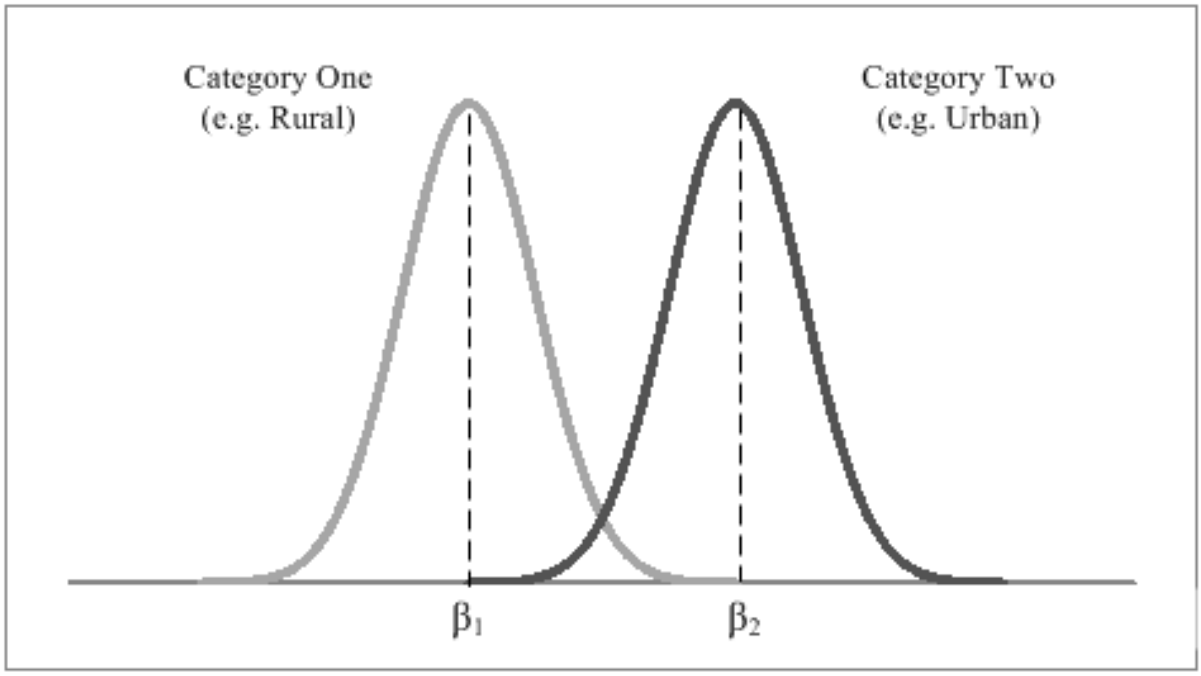
Number of Clusters per Site (m)	Number of Sites (J)					
	4	6	8	10	12	20
4	--	--	--	--	--	--
6	--	--	--	--	--	0.20 (0.92)
8	--	--	--	--	--	0.18 (0.77)
10	--	--	--	--	--	0.17 (0.68)
12	--	--	--	--	0.19 (0.90)	0.16 (0.61)
20	--	--	0.20 (0.94)	0.18 (0.82)	0.17 (0.72)	0.14 (0.49)

NOTES: Values in the table are for two-tail significance = 0.05, power = 80 percent, no level-one covariates, a single level-two covariate, $R^2_{CC} = 0.74$, $\bar{T} = 0.5$, , $\rho_{CS} = 0.07$, $\rho_{CC} = 0.10$, constant *n=200* within each cluster, constant *J* within each site, and $\tau_* = 0.10$ (or $\tau^2_* = 0.01$), and $\pi = 0.60$. Values in parentheses are the R-square that corresponds to each minimum detectable effect size difference.



Note: $\beta$is the cross-site mean program effect size and τ is the cross-site standard deviation of program effect sizes.

Figure 1. A cross-site distribution of program effect sizes.



Note: $\beta_I$is the mean program effect size for category one and $\beta_{II}$is the mean program effect size for category two.

Figure 2. Cross-site effect size distributions for two categories of sites.

Appendix A

Developing an Expression for the Minimum Detectable Effect Size Standard Deviation

This appendix develops a simple way to assess the precision of an estimator of a cross-site standard deviation of program effect sizes ($\hat{\tau}_*$) produced by a randomized multisite trial or MST. Although others have written about this topic (e.g. Raudenbush & Liu, 2000; Hedges & Pigott, 2004; Konstantopoulos, 2008; and Spybrook, 2014, we provide here an expanded discussion that is intended to help deepen applied researchers' understanding of the issues involved and produce a more transparent way to deal with these issues. For excellent background reading, see chapters 11 – 13 of Hays, William L. (1973) *Statistics for the Social Sciences* 2nd edition, New York: Holt, Rinehart and Winston, Inc. and chapters 1,4, and 7 of Liu, X. , L. (2014) *Statistical Power Analysis for the Social and Behavioral Sciences,* New York: Routledge.

Setup

One of the most frequent uses of the F distribution is to test the statistical significance of a *difference* between two estimates of a variance. The simplest version of this test involves a comparison of variance estimates from two independent samples that are hypothesized to represent the same population. The F statistic, which is the basis for this test, is the *ratio* of the two variance estimates. The F distribution for a given test is determined by the number of degrees of freedom used to estimate the variance in its numerator and the number of degrees of freedom used to estimate the variance in its denominator.

The *null* hypothesis for this test is that *no* difference exists between the two variances being estimated. The *alternative* hypothesis may be that the numerator variance is larger than the denominator variance. To test these hypotheses, a critical value from the appropriate F distribution is selected to represent a specified level of statistical significance. If the observed F statistic does not exceed this critical value then one cannot reject the null hypothesis. If the observed F statistic exceeds the critical value one must reject the null hypothesis.

As described below, this approach can be used to test the statistical significance of observed cross-site variation in program effect or effect-size *estimates*. This discussion takes place in the context of the illustrative MST that we have used throughout the present paper: with J sites, n sample members per site and proportion $\bar{T}$ of sample members at each site randomized to treatment. For each site (j) the program effect of intent to treat (B_j) is assumed to be constant across individual sample members (i).[15] This effect is assumed to vary approximately normally across sites with a mean of β and a variance of $\tau^2$and thus a standard deviation of τ.

[15]This eliminates the possibility of heteroskedasticity between treatment and control group members. To deal with such heteroskedasticity one could estimate a separate level-one residual variance for treatment group members and

To estimate β and τ from data for our MST we use the following two-level model with fixed site-specific intercepts (α_j), random site-specific impact coefficients (B_j) and fixed coefficients (θ_k) for each individual-level baseline covariate (X_k).[16]

Level One: Individuals

$$Y_{ij} = \alpha_j + B_j T_{ij} + \sum_{k=1}^{K} \theta_k X_{kij} + e_{ij} \quad \text{(A.1)}$$

Level Two: Sites

$$\alpha_j = \alpha_j \quad \text{(A.2)}$$

$$B_j = \beta + b_j \quad \text{(A.3)}$$

where $e_{ij} \sim N\left(0, \sigma^2_{|\boldsymbol{X}\alpha_j}\right)$, $b_j \sim N(0, \tau^2)$, Y_{ij} is the observed value of the outcome for individual *i* from site *j*; T_{ij} equals 1 if individual *i* from site *j* was randomized to treatment and 0 otherwise; X_{kij} is the value of baseline covariate k for individual *i* from site *j*; θ_k is a constant coefficient for covariate k; α_j is the conditional population mean control group outcome for site *j*; B_j is the population mean treatment effect for site *j*; β is the cross-site mean treatment effect; e_{ij} is a random error that varies independently and identically across individuals within sites and experimental conditions, with a mean of zero and a variance of $\sigma^2_{|\boldsymbol{X}\alpha_j}$; and b_j is a random error that varies independently and identically across sites with a mean of zero and a variance of (τ^2).

Approach

This section describes how to estimate τ^2 (and thus τ) and assess whether this estimate is statistically significantly different from zero.

Estimation

The present approach to estimating τ^2 rests on the well-known fact that *total* cross-site variation in program effect estimates equals the sum of cross-site variation in *true* program effects plus cross-site variation in estimation *error*. In symbols:

$$Var^{Total}\left(\hat{B}_j\right) = Var^{True}\left(B_j\right) + Var^{Error}(\hat{B}_j)$$

$$= \tau^2 + Var^{Error}(\hat{B}_j) \quad \text{(A.4)}$$

control group members (see Bloom et al., revise and resubmit). However that is beyond the scope of the present paper.

[16] Bloom et al. (revise and resubmit) discuss this model and its properties.

Thus:

$$\tau^2 = Var^{Total}\left(\hat{B}_j\right) - Var^{Error}(\hat{B}_j) \quad \text{(A.5)}$$

One can estimate τ^2 by estimating the total variance and error variance for impact estimates obtained from the model represented by Equations A.1 – A.3 and taking their difference. In symbols

$$\hat{\tau}^2 = \widehat{Var}^{Total}\left(\hat{B}_j\right) - \widehat{Var}^{Error}(\hat{B}_j) \quad \text{(A.6)}$$

For our multisite trial

$$\widehat{Var}^{Total}\left(\hat{B}_j\right) = \frac{\sum_j(\hat{B}_j - \hat{\beta})^2}{J-1} \quad \text{(A.7)}$$

$$\widehat{Var}^{Error}\left(\hat{B}_j\right) = \frac{\hat{\sigma}^2_{|\boldsymbol{X}\alpha_j}}{\bar{T}(1-\bar{T})n} \quad \text{(A.8)}$$

where:

$$\hat{\sigma}^2_{|\boldsymbol{X}\alpha_j} = \frac{\sum_j \sum_i \hat{e}^2_{ij}}{J(n-2)-K} \quad \text{(A.9)}$$

$\hat{\beta}$ is the sample cross-site mean value of the $\hat{B}_j$ and K is the number of baseline covariates.

Significance Testing

We can use an F statistic to test whether $\hat{\tau}^2$ is significantly different from zero by comparing two *independent* estimates of $Var^{Total}(\hat{B}_j)$. One estimate is obtained from observed impact variation between sites. The other estimate is obtained from observed outcome variation across individuals within sites.

Under the null hypothesis, total observed cross-site variation in estimates of program effects is due solely to variation in site-level estimation error. In symbols:

$$Var^{Total}\left(\hat{B}_j\right) = Var^{Error}(\hat{B}_j) \quad \text{(A.10)}$$

Hence under the null hypothesis, the estimator for $Var^{Total}\left(\hat{B}_j\right)$ in Equation A.7 and the estimator for $Var^{Error}(\hat{B}_j)$ reflected by Equations A.8 and A.9 represent two different estimates of *the same variance*. Because the distributions of e_{ij} and b_j are assumed to be approximately normal and independent of each other, their ratio is an F statistic with J-1 numerator degrees of freedom and (J(n-2)-K) denominator degrees of freedom. In symbols then:

$$F_{[(J-1),(J(n-2)-K)]} = \frac{\widehat{Var}^{Total}(\hat{B}_j)}{\widehat{Var}^{Error}(\hat{B}_j)} \quad \text{(A.11)}$$

Substituting Equations A.7 – A.9 into Equation A.11 yields:

$$F_{[(J-1),(J(n-2)-K)]} = \frac{\frac{\sum_j(\hat{B}_j-\hat{\beta})^2}{J-1}}{\frac{\sum_j\sum_i e_{ij}^2/(J(n-2)-K)}{\bar{T}(1-\bar{T})n}} \quad \text{(A.12)}$$

Statistical Power of the Approach

This section presents an approach for assessing the statistical power of the preceding estimator. Note first, that for our MST, the cross-site variance of program effect estimation error equals the error variance of a program effect estimator for a single site, or:

$$Var^{Error}\left(\hat{B}_j\right) = \frac{\sigma^2_{|X\alpha_j}}{\bar{T}(1-\bar{T})n} \quad \text{(A.13)}$$

The total cross-site variance of program effect estimates is:

$$Var^{Total}\left(\hat{B}_j\right) = \tau^2 + \frac{\sigma^2_{|X\alpha_j}}{\bar{T}(1-\bar{T})n} \quad \text{(A.14)}$$

The ratio of the expected value of the numerator to the expected value of the denominator of the F statistic that we observe is thus:

$$E[F] = \frac{E[\widehat{Var}^{Total}(\hat{B}_j)]}{E[\widehat{Var}^{Error}(\hat{B}_j)]}$$

$$= \frac{\tau^2+\frac{\sigma^2_{|X\alpha_j}}{\bar{T}(1-\bar{T})n}}{\frac{\sigma^2_{|X\alpha_j}}{\bar{T}(1-\bar{T})n}}$$

$$= 1 + \frac{\tau^2}{\frac{\sigma^2_{|X\alpha_j}}{\bar{T}(1-\bar{T})n}} \quad \text{(A.15)}$$

Before proceeding further, it is useful to reformulate site-specific program *effects* (B_j) as standardized mean-difference *effect sizes* ($B_{*(j)} = \frac{B_j}{\sigma_C}$), where σ_C is the total (between and within site) population standard deviation of control group outcomes (Y_{ij}) in their natural units. In practice σ_C can be estimated as the standard deviation of the outcome for all control group members in a study sample.[17]

Recall that Equation 12 in the main text of the present paper states that:

$$\sigma^2_{|X\alpha_j} = (1-\rho_C)(1-R^2_{C(within)})\sigma^2_C \quad \text{(A.16)}$$

[17] Although different authors use different standard deviations to define their effect size metrics, the total control-group standard deviation is a popular option.

where ρ_C is the proportion of the total control group outcome variance that is *between* sites (the control group intra-class correlation) and $R^2_{C(within)}$ is the proportion of the within-site outcome variance for control group members that is explained by our baseline covariates (***X***).

Substituting Equation A.16 into Equation A.15 and simplifying yields:

$$E[F] = 1 + \frac{\tau^2}{\frac{(1-\rho_C)(1-R^2_{C(within)})\sigma^2_C}{\bar{T}(1-\bar{T})n}}$$

$$= 1 + \left(\frac{\tau^2}{\sigma^2_C}\right)\left(\frac{\bar{T}(1-\bar{T})n}{(1-\rho_C)(1-R^2_{C(within)})}\right) \quad \text{(A.17)}$$

As in the main text of the paper, we define $\tau_* \equiv \frac{\tau}{\sigma_C}$ to be the cross-site standard deviation of program effects sizes ($B_{*(j)}$). Consequently:

$$E[F] = 1 + \tau^2_*\left(\frac{\bar{T}(1-\bar{T})n}{(1-\rho_C)(1-R^2_{C(within)})}\right)$$

$$= 1 + \omega \quad \text{(A.18)}$$

where by definition:

$$\omega \equiv \tau^2_*\left(\frac{\bar{T}(1-\bar{T})n}{(1-\rho_C)(1-R^2_{C(within)})}\right) \quad \text{(A.19)}$$

Dividing F by $(1+\omega)$ produces an F statistic, $\frac{F}{1+\omega}$, with a central F distribution.[18]

To see this, note first that:[19]

$$\frac{F}{(1+\omega)} = \left[\frac{\widehat{Var}^{Total}(\hat{B}_{*(j)})}{\widehat{Var}^{Error}(\hat{B}_{*(j)})}\right]/(1+\omega)$$

$$= \frac{(\widehat{Var}^{Total}(\hat{B}_{*(j)}))/(1+\omega)}{\widehat{Var}^{Error}(\hat{B}_{*(j)})} \quad \text{(A.20)}$$

Note next that

$$E\left[\frac{\widehat{Var}^{Total}(\hat{B}_{*(j)})}{(1+\omega)}\right] = Var^{Error}(\hat{B}_{*(j)}) \quad \text{(A.21)}$$

and

[18] See pages 539 – 540 in Hayes (1973) for a related discussion.

[19] Here we substitute our standardized effect size parameter ($B_{*(j)}$) for its unstandardized counterpart (B_j). This is possible because the former equals the latter divided by a constant (σ_C) which cancels out of Equation A.20.

$$E\left[\widehat{Var}^{Error}(\hat{B}_{*(j)})\right] = Var^{Error}(\hat{B}_{*(j)}) \quad \text{(A.22)}$$

Equations A.19 – A.21 imply that the numerator and denominator of the quantify ($\frac{F}{1+\omega}$) represent independent estimates of the *same variance*. Hence ($\frac{F}{1+\omega}$) follows a central F distribution.

Determining Statistical Power

This section describes how to determine the statistical power of a hypothesis test about τ_*^2, and thus about τ_*.

Theory: Equations A.21 and A.22 provide a way to determine the statistical power of an MST to detect a non-zero τ_*^2 using a central F distribution for a transformed F statistic ($\frac{F}{1+\omega}$).

The first step is to determine the desired level of statistical significance for testing the null hypothesis that τ_*^2 (and thus τ_*) equals zero. This will determine the critical value of the associated F test given its degrees of freedom ((J-1) and (J(n-2) – K)). If the observed F value ($F_{observed}$) exceeds its critical value ($F_{critical}$), we reject the null hypothesis that τ_*^2 is zero. But if $F_{observed} \leq F_{critical}$, we cannot reject the null hypothesis. Note that under the null hypothesis, $F_{observed}$ has a central F distribution.

The power of this test for a given positive value of τ_*^2 is the *probability* that the test will correctly reject the null hypothesis and thereby detect the positive effect size variance. To determine this probability, recall that under the alternative hypothesis that τ_*^2 is positive, the observed value of the transformed F statistic ($\frac{F_{observed}}{1+\omega}$) has a central F distribution.[20] Thus we can compute the statistical power of our test as follows.

First consult a table for the appropriate central F distribution to determine the value of $F_{critical}$, given the number of numerator and denominator degrees of freedom for your F statistic and the level of statistical significance for your hypothesis test. Next compute the value of ($\frac{F_{critical}}{1+\omega}$). Then consult the F table again to determine the probability that $\frac{F_{observed}}{1+\omega} \geq \frac{F_{critical}}{1+\omega}$. This is the probability that $F_{observed} \geq F_{critical}$, given the *specific positive value* of τ_*^2 under your alternative hypothesis.

Example: Consider an MST with 80 sites, 60 sample members at each site and 60 percent of sample members at each site randomized to treatment. Then assume that K= 1, $\rho_C = 0.2$ and $R^2_{C(within)} = 0.25$. What is the statistical power of a hypothesis test to detect a cross-site effect-size variance (τ_*^2) of $0.02\sigma_C^2$, which implies a cross-site effect size standard deviation (τ_*) of 0.141)? Note first that:

[20] $F^*_{observed}$ also has a central F distribution under the null hypothesis.

$$\omega = \tau_*^2\left(\frac{\bar{T}(1-\bar{T})n}{(1-\rho_C)(1-R^2_{C(within)})}\right)$$

$$= 0.02\left(\frac{0.6(0.4)60}{(1-0.2)(1-0.25)}\right) = 0.48 \quad \text{(A.23)}$$

Thus $1 + \omega = 1.48$.

Note next that the number of degrees of freedom for the numerator of $F_{observed}$ equals (80-1) or 79 and the number of degrees of freedom for its denominator equals (80(60-2) – 1) or 4639. The corresponding value of $F_{critical}$ for the 0.05 level of statistical significance is 1.28. Therefore $(\frac{F_{critical}}{1+\omega})$ equals $(\frac{1.26}{1.48})$ or 0.86. Based on the appropriate central F distribution, the probability that by chance $(\frac{F_{observed}}{1+\omega})$ will exceed 0.86 is 0.80.[21] Hence the power of your test to detect a true effect size variance of 0.02 is 80 percent.

Determining a Minimum Detectable Effect Size Variance or Standard Deviation

This section describes how to determine a minimum detectable effect size variance (MDESV) or a minimum detectable effect-size standard deviation (MDESSD).

Theory: An MDESV is the smallest value of τ_*^2 for which our hypothesis test has acceptable statistical power (typically 80 percent) at a specified level of statistical significance (typically 0.05). Note that 80 percent power implies a 0.80 probability that by chance, $(\frac{F_{observed}}{1+\omega}) \geq (\frac{F_{critical}}{1+\omega})$. This implies that $(\frac{F_{critical}}{1+\omega})$ must equal the value of $(\frac{F_{observed}}{1+\omega})$ that lies at or below 80 percent of the appropriate central F distribution. Denoting this value as $F_{0.80}$ implies that we must find the value of τ_* for which:

$$\frac{F_{critical}}{(1+\omega)} = F_{0.80} \quad \text{(A.24)}$$

Rearranging terms in Equation A.24 yields:

$$1 + \omega = \frac{F_{critical}}{F_{0.80}} \quad \text{(A.25)}$$

and thus

$$\omega = \frac{F_{critical}}{F_{0.80}} - 1 \quad \text{(A.26)}$$

Equating the expression for ω in Equation A.26 with the expression for ω in Equation A.19 and re-arranging terms yields:

[21] Findings for the conventional F distribution were obtained using the FDIST and FINV functions in excel.

$$\tau_*^2\left(\frac{\bar{T}(1-\bar{T})n}{(1-\rho_C)(1-R^2_{C(within)})}\right) = \frac{F_{critical}}{F_{0.80}} - 1 \quad \text{(A.27)}$$

Given $J, n, \bar{T},\ \rho_C$ and $R^2_{C(within)}$, the value of τ_*^2 that satisfies Equation A.27 is the minimum detectable effect size variance (MDESV). Noting this fact and solving Equation A.27 for τ_*^2 therefore yields:

$$\tau_*^2 = \left(\frac{(1-\rho_C)(1-R^2_{C(within)})}{\bar{T}(1-\bar{T})n}\right)\left(\frac{F_{critical}}{F_{0.80}} - 1\right) = MDESV \quad \text{(A.28)}$$

Taking the square root of Equation A.28 yields:

$$\tau_* = \sqrt{\left(\frac{(1-\rho_C)\left(1-R^2_{C(within)}\right)}{\bar{T}(1-\bar{T})n}\right)\left(\frac{F_{critical}}{F_{0.80}} - 1\right)} = MDESSD \quad \text{(A.29)}$$

Equation A.29 provides a simple expression for the MDESSD in terms of features of our MST, impact estimation model or data that can be specified ($J, n, \bar{T}$ and K) or assumed (ρ_C and $R^2_{C(within)}$) plus two F values that can be determined for a central F distribution.

Example: What is the MDESSD for an MST with 150 sites (J), 10 sample members per site (n), 60 percent of the sample members at each site randomized to treatment ($\bar{T} = 0.6$), one baseline covariate ($K = 1$), a value of $\rho_C = 0.10$, a value of $R^2_{C(within)} = 0.22$ and a statistical significance test at the 0.05 level. Our MST provides (150 -1) or 149 degrees of freedom for the numerator of the F test and (150(10-2)-1) or 1,199 degrees of freedom for its denominator. Corresponding values of $F_{critical}$ at the 0.05 significance level and $F_{0.80}$ are 1.212 and 0.897, respectively. Substituting the preceding information into Equation A.29 yields:

$$MDESSD = \sqrt{\left(\frac{(1-0.1)(1-0.22)}{0.6(0.4)10}\right)\left(\frac{1.21}{0.90} - 1\right)} = 0.32 \quad \text{(A.30)}$$

Appendix B

The Minimum Detectable Effect Size Difference

This appendix develops several expressions that are helpful for understanding how the *minimum detectable difference* between mean program effect sizes for two subgroups of sites (*MDESD*) can be larger than the *maximum possible difference* between their mean effect sizes ($\Delta_{*(max)}$. For the discussion, we assume an MST with J sites, n sample members per site, proportion $\bar{T}$ of the sample members at each site randomized to treatment, a control group intraclass correlation of ρ_C and a covariate explanatory power of $R^2_{C(within)}$ within sites and experimental groups. In addition, we assume that proportion π of our program sites are in subgroup II and proportion $(1-\pi)$ are in subgroup I, the cross-site mean program effect sizes are $\beta_{*(I)}$ and $\beta_{*(II)}$ for subgroups I and II, and the difference between these mean effect sizes (Δ_*) equals $(\beta_{*(II)} - \beta_{*(I)})$.

The Minimum Detectable Effect Size Difference

To derive an expression for the MDESD, note first that:

$$MDESD = M_{J-2} se(\hat{\Delta}_*)$$

$$= M_{J-2} se(\hat{\beta}_{*(II)} - \hat{\beta}_{I(*)})$$

$$= M_{J-2}\sqrt{Var(\hat{\beta}_{*(II)}) + Var(\hat{\beta}_{*(I)})} \qquad \text{(B.1)}$$

Note next that the *residual* cross-impact variation *within* each of our two subgroups equals $(1-R^2_W)\tau^2_*$, where R^2_W is the proportion of total cross-site impact variation (τ^2_*) that is predicted by our binary site subgroup indicator, W. This fact, in conjunction with Equation 15 in the main text of the present paper, implies that:

$$Var(\hat{\beta}_{*(I)}) = (\frac{1}{(1-\pi)J})((1-R^2_W)\tau^2_* + \frac{(1-\rho_C)(1-R^2_{C(within)})}{n\bar{T}(1-\bar{T})}) \qquad \text{(B.2)}$$

and

$$Var(\hat{\beta}_{*(II)}) = (\frac{1}{\pi J})((1-R^2_W)\tau^2_* + \frac{(1-\rho_C)(1-R^2_{C(within)})}{n\bar{T}(1-\bar{T})}) \qquad \text{(B.3)}$$

Equations B.2 and B.3 imply that:

$$Var(\hat{\beta}_{*(I)}) + Var(\hat{\beta}_{*(II)}) = \left(\frac{1}{\pi(1-\pi)J}\right)\left((1 - R_W^2)\tau_*^2 + \frac{(1-\rho_C)\left(1-R^2_{C(within)}\right)}{n\bar{T}(1-\bar{T})}\right) \quad \text{(B.4)}$$

Therefore:

$$MDESD = M_{J-2}\sqrt{\left(\frac{1}{\pi(1-\pi)J}\right)\left((1 - R_W^2)\tau_*^2 + \frac{(1-\rho_C)\left(1-R^2_{C(within)}\right)}{n\bar{T}(1-\bar{T})}\right)} \quad \text{(B.5)}$$

The Maximum Possible Effect Size Difference

The first step toward understanding the maximum possible effect size difference ($\Delta_{*(max)}$) is to understand the relationship between the actual effect size difference ($\Delta_* = \beta_{*(II)} - \beta_{*(I)}$) and the total amount of cross-site effect size variation (τ_*^2). To see this, note first that τ_*^2 is the sum of two cross-site impact variance components: (1) that which is predicted by our site subgroup indicator and thus is between subgroups ($\tau^2_{*(between)}$) and (2) that which is not predicted by our subgroup indicator and thus is within subgroups ($\tau^2_{*(within)}$). One can therefore express R_W^2 as:

$$R_W^2 \equiv \frac{\tau^2_{*(between)}}{\tau_*^2} \quad \text{(B.6)}$$

which implies that:

$$\tau^2_{*(between)} = R_W^2\tau_*^2 \quad \text{(B.7)}$$

One can also express the amount of cross-site impact variation that is between site subgroups ($\tau^2_{*(between)}$) as a function of the *difference* between mean effect sizes for the two subgroups (Δ_*). To demonstrate this, set $\beta_{*(I)}$ equal to zero for simplicity, but without loss of generality. This implies that $\beta_{*(II)}$ must equal Δ_* which in turn, implies that

$$\begin{aligned}\beta_* &= (1-\pi)\beta_{*(I)} + \pi\beta_{*(II)} \\ &= (1-\pi)0 + \pi\Delta_* \\ &= \pi\Delta_* \end{aligned} \quad \text{(B.8)}$$

We can now express $\tau^2_{*(between)}$ in terms of Δ_*. First recall that proportion $(1-\pi)$ of our sites are in subgroup I and thus have a mean program effect size of zero. The contribution of

each of these sites to the between-subgroup program effect size variance is therefore $(0 - \pi\Delta_*)^2$. Likewise, proportion π of our sites are in subgroup II and thus have a mean effect size of Δ_*. The contribution of each of these sites to the between-subgroup program effect size variance is therefore $(\Delta_* - \pi\Delta_*)^2$. Combining these two facts in their appropriate proportion yields:

$$\begin{aligned}\tau^2_{*(between)} &= (1-\pi)(0-\pi\Delta_*)^2 + \pi(\Delta_* - \pi\Delta_*)^2 \\ &= (1-\pi)\pi^2\Delta_*^2 + \pi\Delta_*^2(1-\pi)^2 \\ &= \pi^2\Delta_*^2 - \pi^3\Delta_*^2 + \pi\Delta_*^2(1-2\pi+\pi^2) \\ &= \pi^2\Delta_*^2 - \pi^3\Delta_*^2 + \pi\Delta_*^2 - 2\pi^2\Delta_*^2 + \pi^3\Delta_*^2 \\ &= -\pi^2\Delta_*^2 + \pi\Delta_*^2 \\ &= \Delta_*^2\pi(1-\pi) \qquad \text{(B.9)}\end{aligned}$$

Substituting Equation B.9 in Equation B.7 yields:

$$\Delta_*^2\pi(1-\pi) = R_W^2\tau_*^2 \qquad \text{(B.10)}$$

Solving Equation B.10 for Δ_* yields

$$\Delta_* = \sqrt{\frac{R_W^2\tau_*^2}{\pi(1-\pi)}} \qquad \text{(B.11)}$$