Using Multi-site Experiments to Study Cross-site Variation in Effects of Program Assignment

Howard S. Bloom
Stephen W. Raudenbush
Michael Weiss
Kristin Porter

Working Draft
03-28-16

This draft represents work in progress for a project funded by the Spencer Foundation, the William T. Grant Foundation and the Institute for Education Sciences. All statements are solely the responsibility of the authors.

Abstract

We consider a fundamental question in evaluation research: "By how much do program effects vary across sites?" After reviewing past research on this question, we present a theoretical model of cross-site impact variation and a simple analytic model with random coefficients and fixed intercepts. The approach eliminates several biases that can arise in unbalanced designs within a multi-site randomized trial. We describe how the approach operates, identify key assumptions, and apply it to a large welfare-to-work trial. We illustrate ways to report and interpret cross-site impact findings, and we present new diagnostic procedures for assessing the confidence that one should have in such findings. To keep the paper manageable, we focus solely on experimental estimates of effects of program assignment (effects of intent to treat), although the basic ideas presented can be extended to analyses of rigorous quasi-experiments and estimates of effects of program participation (complier average causal effects).

Using Multi-site Experiments to Study Cross-Site Variation In Effects of Program Assignment

Introduction

Variation in program effects has important consequences for policy, practice and research. If this variation is large and unexplained, knowing the average impact will tell us little about how well the program worked in particular sites or for particular sub-groups of persons. Moreover, the average impact will not extrapolate well to settings, points in time, or persons beyond those represented in the study (Tipton, 2014). However, if program effects vary little across sites, sub-groups, or points in time, the average impact estimate will apply well, not only to particular sites and sub-groups in the study but also to similar sites, times, and sub-groups not included in the study.

Variation in program effects also has implications for targeting services to identifiable sub-populations. If the effect of a program varies predictably across clients or contexts, it might be desirable to target the program, which, if feasible, can increase its cost-effectiveness[1].

In addition, variation in program effects has implications for equity or fairness. For example, a reading curriculum that is most effective for "struggling" readers might reduce disparity in reading achievement, whereas a reading curriculum that is most effective for "gifted and talented" students might increase disparity.

Furthermore, in many cases it can be important to know the likely best-case or worst-case effects of a program. For example, when evaluating charter schools for disadvantaged children, it could be important to know how effective the most effective charter schools are – as evidence of what can be achieved under the right circumstances. It could also be useful to know the proportion of schools that are less effective than their local alternatives.

Nonetheless, to date researchers have focused mainly on *average* program effects, even though scholars have argued for decades that to fully understand programs, one must understand how their effects vary.[2]

When in the past, researchers have attempted to go beyond just studying average program effects they have usually compared estimates of average effects for *subgroups* of individuals defined by characteristics such as their gender, race/ethnicity, age or a measure of risk status. For example, there is a long-standing theoretical and empirical

[1] Targeting is common practice in medicine (e.g. Horwitz et al., 1996 and Rothwell, 2005) and for some government programs (e.g. Berger, Black and Smith, 2001; Eberts, O'Leary and Wander, 2002; Haager et al., 2007 and Ehren et al., 2010).

[2] For example see Bryk and Raudenbush (1988); Heckman, Smith and Clements (1997); Friedlander and Robins (1997); Heckman (2001); Raudenbush and Liu (2000); Abadie, Angrist and Imbens (2002); Bitler, and Hoynes (2006); Bitler, Hoynes and Domina (2014); and Djebbari and Smith (2008).

debate in the literature on welfare-to-work programs about whether: (1) participants who are most at risk economically benefit the most (perhaps because they have the greatest margin for improvement), (2) participants who are least at risk economically benefit the most (perhaps because they are best able to utilize program services); or (3) participants who are between these two extremes benefit the most (perhaps because they have the best mix of room for and ability to achieve improvement -- Gueron and Pauly, 1991; Friedlander, 1993; Michalopoulos and Schwartz, 2000).

Likewise, in education research there is often interest in the potential effects of interventions for students who are especially at risk of dropping out of school (e.g. Kemple, Snipes and Bloom, 2001). Although this type of subgroup analysis is fraught with opportunities for distortion (e.g. through ex-post selection of subgroups based on observed findings and claims about subgroup differences that are not based on tests of statistical significance and erroneous conclusions about confidence due to multiple hypothesis testing), it can provide valuable information when conducted properly, particularly when it is based on a clearly-specified *a priori theory*.[3]

A second, less frequently-used approach to studying variation in program effects is *quantile regression analysis* (e.g. Friedlander and Robins, 1997; Heckman, Smith and Clements, 1997; Abadie, Angrist and Imbens, 2002; Bitler and Hoynes, 2006; and Bitler, Hoynes and Domina, 2014).[4] This approach can determine the effect of a program on the distribution of individual outcomes by estimating how the program changes the value of the outcome at each quantile (a percentile, a quartile, etc.). For example, we could compare the 25th, 50th and 75th percentile values of the test-score distribution for a randomized treatment group with their control group counterparts. In this way, we could determine for example, whether a reading program increases or decreases the disparity in reading achievement. However this method cannot estimate the distribution of site-specific program effects without imposing untestable assumptions.[5]

A third existing source of information about variation in program effects is *meta-analysis.* This approach is used widely to summarize findings from an existing body of research. Two international organizations – the Cochrane Collaboration for synthesizing research in health science and the Campbell Collaboration for synthesizing research in

---

[3] A 2009 conference on subgroup analysis was convened by the Agency for Children and Families and the Office of the Assistant Secretary for Planning and Evaluation of the U.S. Department of Health and Human Services, the Institute for Education Sciences of the U.S. Department of Education, the Division of Violence Prevention in the Centers for Disease Control and Prevention, the National Institute of Drug Abuse and the National Institute of Mental Health. Papers from this conference (including one by an author of the present paper, Bloom and Michalopoulos, 2011) were published in a special issue of *Prevention Science* 2011.

[4] Koenker (2005) and Buchinsky (1998) provide detailed discussions of quantile regression analysis.

[5] One such assumption is "rank preservation," which implies that a program can change the distance but not the rank order between individual outcomes. When this condition holds, the nth quantile score for control group members in a randomized trial is an unbiased estimate of the nth quantile counterfactual score for treatment group members. The difference between the nth quantile scores for treatment and control group members under this condition is thus an unbiased estimate of the program effect for persons who without the program would have scored at this quantile.

education, crime and justice and social welfare – exist to promote and advance the approach. In addition, textbooks (e.g. Lipsey and Wilson, 2001; Rosenthal, 1991; and Hedges and Olkin, 1985) and handbooks (e.g. Cooper and Hedges, 1994) exist for training researchers in the approach. Meta-analysis provides statistical and graphical ways to represent variation in program effects. Its primary limitations stem from its reliance on summary reports of past studies, as meta-analysts rarely have access to the person-level data collected by primary studies. Moreover, the outcome measures used by different studies, the interventions they test, the research designs they use, the estimation methods they use, and the types of information they report frequently differ across studies. This makes it difficult to distinguish variation in program effects from differences in study findings due to their methodological differences. Nonetheless, much has been learned from the approach.

We explore a fourth approach to studying impact variation: using data from *multi-site randomized trials* to study how program effects vary across sites. Only a few studies to date have used multi-site trials for this purpose (e.g. Bloom and Weiland, 2015; Bloom, Hill and Riccio, 2003; Konstantopoulos, 2011; Raudenbush, Reardon and Nomi, 2012; Lake et al., 2012; Angrist, Pathak and Walters, 2013, May et. al., 2013 and Walters, 2015). However with the large and growing number of multi-site trials, the ability to conduct such analyses is increasing rapidly. For example, during the past decade, the Institute of Education Sciences (IES) conducted roughly 175 large-scale randomized trials, most of which were multi-site trials (Spybrook, 2013). In addition, there is a growing number of such trials funded by federal initiatives to replicate evidence-based programs at large scale. Examples include the White House Social Innovation Fund, initiatives by the U.S. Department of Health and Human Services to replicate evidence-based home-visiting programs, pregnancy prevention interventions and fatherhood initiatives, plus programs supported by the Investment in Innovation Fund and First in the World program of the U.S. Department of Education.

To help researchers learn as much as possible from this large and growing body of evidence, we introduce a promising approach for using existing statistical methods to study a *cross-site distribution* of site mean program effects. To stay within the scope of a single paper we focus on effects of program assignment (intent to treat) and leave consideration of the effects of program participation (complier average causal effects *aka* local average treatment effects or) for future research.

The next section presents a theoretical model of the cross-site distribution of program assignment effects and a simple two-level hierarchical model for estimating the mean and variance of this distribution. Subsequent sections introduce key issues in detecting and quantifying cross-site impact variation. The penultimate section presents an empirical example to illustrate the approaches discussed and the final section provides some concluding thoughts.

Statistical Models

We begin with a theoretical model of a cross-site distribution of *site mean* program assignment effects and then propose a statistical model for estimating the mean and variance or standard deviation of this distribution. For this purpose, we focus on a population of sites, each equally important, and consider a study sample to represent a random sample of such sites. It is also possible to examine cross-site impact variation for a population of *program eligible persons*. In this case, the importance of each site is proportional to its number of program eligible persons. This type of analysis requires related estimation methods that are discussed elsewhere (e.g. see Raudenbush and Bloom, 2015, and Raudenbush 2015.)

Theoretical Model

Consider the following theoretical model for a multi-site trial that randomly assigns individuals within each site to a program (the treatment) or to a control group:

$$Y_{ij} = A_j + B_j T_{ij} + e_{ij} \qquad [1]$$

where:

Y_{ij} = the outcome for individual *i* from site *j*,
T_{ij} = 1 if individual *i* from site *j* is assigned to the program and 0 otherwise,
A_j = the mean outcome at site *j* if all of its program-eligible population members were assigned to the control condition,
B_j = the mean program effect for the population of program-eligible persons at site *j*,
e_{ij} = a random error that varies independently across individuals with a mean of zero and a variance that can differ between treatment and control group members and across sites.[6]

When studying cross-site impact variation, it is useful to consider study sites as a probability sample from some population of sites, regardless of whether the sites were chosen as a probability sample from a well-defined population or as a convenience sample from an implied population that cannot be fully described. This is because a primary goal of studying impact variation is to generalize to a population that is broader than the sample that was observed.

The National Head Start Impact Study (Puma et al., 2010) was based on a probability sample of over-subscribed sites from a national program during a specific program year (2002 – 2003). Consequently, this study was designed to generalize to a well-defined population. However, most evaluation studies are based on a convenience

[6] The treatment and control group variance difference allows for program effects that can vary across individuals within sites (see Raudenbush and Bloom, 2015).

sample of sites that are chosen to represent as broad a range of conditions as is possible. But even in a convenience sample, sites are usually chosen not because they *comprise* a population of interest, but rather because they *represent* a broader population of sites that might have participated in the study or might consider adopting the program being tested. Hence, the ultimate goal of such studies is to generalize findings beyond the sites observed, even though the target of generalization is not well-defined.

Of course, there can be cases in which there is no immediate interest in generalizing beyond study sites. For example, an intervention might be tested in both of the only two high schools in a specific town and the local school board might just want to know how well this intervention worked *in those two schools*. Here, the two schools (study sites) represent the population of interest and would be regarded as "fixed" instead of "random" in the experimental design literature (e.g. Kirk, 1982). Discussion of such fixed-site designs is outside the scope of the present paper.

A multi-site trial has a number of population parameters of potential interest. Often, the parameter of greatest interest is the population average program effect ($E(B) = \beta$), which for the present paper is defined as the mean of the mean program effects for all sites in the theoretical population. In this case, each site is equally important and thus has equal weight in the definition (but not necessarily in the estimate) of distributional parameters. A second parameter of interest, which is the central focus of the present paper, is the population *cross-site* variance of program effects ($Var(B) = \tau_B^2$) or its more readily interpretable counterpart, the population cross-site standard deviation (τ_B). Other parameters of potential interest are the population average control-group outcome ($E(A) = \alpha$), the population cross-site variance of control-group outcomes ($Var(A) = \tau_A^2$) and the population cross-site covariance between control-group outcomes and program effects ($Cov(A, B) = \tau_{AB}$). With this in mind, our theoretical model can be written as a two-level hierarchical linear model (HLM) in which level one is represented by Equation 1 above and level two is represented by Equations 2 and 3 below.

$$A_j = \alpha + a_j \quad [2]$$

and

$$B_j = \beta + b_j \quad [3]$$

where site-specific random effects a_j and b_j have cross-site means of zero, a cross-site covariance τ_{ab} and cross-site variances τ_a^2 and τ_b^2, respectively. Note that:

$$Cov(A, B) = Cov(a, b) = \tau_{AB} = \tau_{ab} \quad [4]$$

$$Var(A) = Var(a) = \tau_A^2 = \tau_a^2 \quad [5]$$

$$Var(B) = Var(b) = \tau_B^2 = \tau_b^2 \quad [6]$$

Combining Equations 1, 2 and 3 yields the following "mixed-model" formulation of our theoretical model:

$$Y_{ij} = \alpha + \beta T_{ij} + a_j + b_j T_{ij} + e_{ij} \quad [7]$$

Estimation Model

Of the conditions required for consistent estimates of our theoretical model, perhaps the most important is that the site-specific fraction of persons assigned to the treatment ($\bar{T}$) should be uncorrelated with site-level random effects (a and b).[7] If $\bar{T}$ varies across sites and is correlated with unobserved site characteristics (and thus related to a or b), standard methods can produce inconsistent estimates our model parameters.

One simple way to eliminate this problem is to *site-center* the variables in level one of our theoretical model (Equation 1). The mean value of these site-centered variables is zero for all sites and thus cannot be correlated to any site differences. To see how this works, note that according to our theoretical model, the sample mean outcome for a given site ($\bar{Y}_j$) is:

$$\bar{Y}_j = A_j + B_j \bar{T}_j + \bar{e}_j \quad [8]$$

where

$\bar{T}_j$ = the mean value of the treatment assignment indicator for the sample from site j,
$\bar{e}_j$ = the mean value of the individual-level error term for the sample from site j,

By site-centering Equation 1 we eliminate its site-specific intercepts (A_j) and obtain:

$$Y_{ij} - \bar{Y}_j = B_j\left(T_{ij} - \bar{T}_j\right) + e_{ij} - \bar{e}_j \quad [9]$$

Consequently our level-two model simplifies to:[8]

$$B_j = \beta + b_j \quad [10]$$

where $E(b) = 0$ and $Var(b) = \tau_b^2 = \tau_B^2$

[7] Standard identification assumptions for Equation 7 are that: (i) T must be independent of the individual-level random error (e); (ii) T must be independent of the individual-level effects a and b, (iii) for two participants i and i' from site j, e_{ij} must be independent of $e_{i'j}$ and uncorrelated with a_j and b_j ; (iv) the analyst must correctly specify the variance structure of e (for example e might be assumed to have a constant variance σ^2 or a variance that depends on T (that is $Var(e_{ij}) = T_{ij}\sigma_1^2 + (1 - T_{ij})\sigma_0^2$); and (v) the analyst must correctly specify the covariance structure of a and b. Of these, (i) is guaranteed by randomization within each site, failure of (iii) and (iv) can be overcome by estimating robust standard errors if the number of sites is not too small and (v) is not problematic in the case that we consider.

[8] Note that using a site-centered model like Equation 9 implies that we *are not attempting to estimate* three parameters, α, τ_A^2 or τ_{AB}. Estimating these parameters involves issues that are beyond the scope of the present paper (Raudenbush, 2015).

When estimating a site-centered hierarchical model like Equations 9 and 10 we must account for the loss of one degree of freedom per site produced by site-centering the dependent variable. If site-centering is a defined option for the software being used (e.g. it is for HLM), the software will automatically account for this loss of degrees of freedom. One simple way to achieve the same result using software that does not have a site-centering option, is to estimate the following two-level hierarchical linear model with *fixed* site-specific intercepts (A_j) and *random* site-specific program assignment effects (B_j).

Individual Level

$$Y_{ij} = A_j + B_j T_{ij} + e_{ij} \quad [11]$$

Site Level

$$A_j = A_j \quad [12]$$

$$B_j = \beta + b_j \quad [13]$$

This two-level *fixed (site-specific) intercept, random site coefficient model* is the basis for the discussion which follows. In practice we can add baseline covariates to Equation 11 in order to increase the precision of its estimated parameters.[9] Doing so does not change other basic properties of the model.

Estimation and Inference

To understand how the preceding model works and to gain insights into its strengths and weaknesses, it is useful to examine some key features of the maximum likelihood method that is the basis for estimating the model. The foundation of this analysis is a model of impact estimation error for each site based on individual outcome variation within sites and a sampling model of impact variation *between* sites.

Within-site Model

With random assignment of sample members to a treatment group or a control group at each site, one can obtain an unbiased ordinary least squares (OLS) estimate ($\hat{B}_j^{OLS}$) of the mean program effect for that site (B_j) from the difference between the mean outcome for its treatment group ($\bar{Y}_{1j}$) and the mean outcome for its control group ($\bar{Y}_{0j}$), or:

$$\hat{B}_j^{OLS} = \bar{Y}_{1j} - \bar{Y}_{0j} \quad [14]$$

[9] Note that adding baseline covariates to Equation 11 is equivalent to site-centering them.

We thus can regard a multi-site trial as a type of planned "meta-analysis" in which the OLS-estimated program effect ($\hat{B}_j^{OLS}$) for each site is an estimate of the site's "true" program effect (B_j) plus its random estimation error (r_j). Thus:

$$\hat{B}_j^{OLS} = B_j + r_j \qquad [15]$$

Note that $(r) = 0$ and $Var(r) = V_j$, where V_j is the site-specific estimation error variance of $\hat{B}_j^{OLS}$ (the square of its standard error).

For a treatment and control group difference in mean outcomes:

$$V_j = \frac{\sigma_{1j}^2}{n_{1j}} + \frac{\sigma_{0j}^2}{n_{0j}} = \frac{\sigma_{1j}^2}{n_j \cdot \bar{T}_j} + \frac{\sigma_{0j}^2}{n_j(1-\bar{T}_j)} \qquad [16]$$

where

σ_{1j}^2 = the within-site variance of outcomes for treatment group members from site j,
σ_{0j}^2 = the within-site variance of outcomes for control group members from site j,
n_j = the number of sample members from site j,
$n_{1j} = n_j\bar{T}_j$ = the number of treatment group members from site j
$n_{0j} = n_j(1 - \bar{T}_j)$ = the number of control group members from site j.

The typical OLS default is a within-site outcome variance that is the same for treatment and control group members and the same for all sites ($\sigma_{1j}^2 = \sigma_{0j}^2 = \sigma^2$ for all j). However, if program effects vary across individuals in a site, the outcome variance for its treatment group members will most likely differ from that for its control group members ($\sigma_{1j}^2 \neq \sigma_{0j}^2$). We refer to this condition as *T/C heteroskedasticity* and demonstrate later that if this condition is not accounted for, it can bias estimates of τ_B^2.

But what if the variance of individual outcomes also varies across sites? As discussed below, it is often not necessary to account for this *cross-site heteroskedasticity*. Indeed, if one were to try to do so by estimating separate individual-level outcome variances for each site, this could have the perverse effect of *overstating* the magnitude and statistical significance of $\hat{\tau}^2$. This bias, which can be substantial, is most extreme for studies with *many small sites* (see Appendix A).

We therefore propose an estimation model with a single individual-level outcome variance for all treatment group members from all sites ($\sigma_{1j}^2 = \sigma_1^2$ for all j) and a single individual-level outcome variance for all control group members from all sites ($\sigma_{0j}^2 = \sigma_0^2$ for all j). Consequently

$$V_j = \frac{\sigma_1^2}{n_j \cdot \bar{T}_j} + \frac{\sigma_0^2}{n_j(1-\bar{T}_j)} \qquad [17]$$

This option is currently available in software for estimating multi-level models like HLM, SAS PROC MIXED and R.

The Combined Between- and Within-site Model

Recall that true mean program effects vary across sites as:

$$B_j = \beta + b_j \qquad [18]$$

where

$$E(b) = 0$$

$$Var(b) = \tau_b^2 = \tau_B^2.$$

Combining Equations 18 and 15 yields:

$$\hat{B}_j^{OLS} = \beta + b_j + r_j \qquad [19]$$

Where $E\left(\hat{B}_j^{OLS}\right) = \beta$ and

$$Var\left(\hat{B}_j^{OLS}\right) = Var(\beta) + Var(b) + Var(r)$$

$$= 0 + \tau_B^2 + V_j$$

$$= \tau_B^2 + V_j \qquad [20]$$

Equation 20 indicates that an OLS mean program effect estimate ($\hat{B}_j^{OLS}$) for a given site j contains two sources of variation: (1) the cross-site variance of true mean program effects (τ_B^2): and (2) the estimation error variance of the impact estimate for that site (V_j). As demonstrated later, it is essential to use a method that distinguishes between these two sources of variation in order to estimate τ_B^2 (e.g. see Hedges and Pigott, 2001).

Estimating a Population Mean of the Distribution of Site-mean Program Effects

With random sampling of sites from a population of sites, $\hat{B}_j^{OLS}$ for a given site is an unbiased (but imprecise) estimator of the population mean program effect (b). Equation 20 indicates that this estimator has a total variance of ($\tau_B^2 + V_j$). If V_j and τ_B^2 were known or estimated accurately, the "best" estimator of the population mean

program effect is an average of the site specific OLS estimates, *weighted by their precision*, as follows: [10]

$$\hat{\beta} = \frac{\sum_{j=1}^{J}(\tau_B^2+V_j)^{-1}\hat{B}_j^{OLS}}{\sum_{j=1}^{J}(\tau_B^2+V_j)^{-1}} \quad [21]$$

Equation 21 down-weights program-effect estimates from sites with large estimation error variances (V_j) produced by small samples and/or values of $\bar{T}_j$ that are far from 0.5. Equation 21 up-weights program-effect estimates from sites with small estimation error variances (V_j) produced by large samples and/or values of $\bar{T}_j$ that are near 0.5. This tendency is partly offset by τ_B^2 which tends to equalize weights across sites. Other things being equal, the more V_j varies across sites, the more site weights differ and the more program effects vary across sites, the less site weights differ. Because this estimator requires knowledge of τ_B^2 and V_j, we must use consistent estimators of them to implement the estimator.

Testing for Cross-site Impact Variation

Under the null hypothesis that $\tau_B^2 = 0$, the site weight in Equation 21 simplifies to $\frac{V_j^{-1}}{\sum_{j=1}^{J} V_j^{-1}}$ and we can estimate β with the following fixed-effect estimator:

$$\hat{B}^{FIXED} = \frac{\sum_{j=1}^{J} V_j^{-1}\hat{B}_j^{OLS}}{\sum_{j=1}^{J} V_j^{-1}} \quad [22]$$

We can then compute a *Q*-statistic, which is widely used in meta-analysis to test the null hypothesis (Hedges and Olkin, 1985), where:

$$Q = \sum_{j=1}^{J} \frac{(\hat{B}_j^{OLS} - \hat{\beta}^{FIXED})^2}{V_j} \quad [23]$$

Under the null hypothesis, if our estimates, $\hat{B}_j^{OLS}$, are approximately normally distributed, the Q-statistic will approximate a central Chi-square distribution with *J*-1 degrees of freedom. The Q-statistic thus provides a statistical significance test for detecting cross-site impact variation if we replace V_j for each site with its sample-based estimate ($\hat{V}_j$).

Estimating a Cross-site Impact Variance

To use Equation 21 to estimate β requires estimating a value for τ_B^2, which is itself a parameter of interest. As a first step toward an estimator for τ_B^2 recall that:

[10] If b and r are normally distributed, then Equation 21 is the unique, minimum variance unbiased estimator of β, achieving the minimum variance bound $Var(\hat{\beta}) = [\sum_{j=1}^{J}(\tau_B^2 + V_j)^{-1}]^{-1}$. Without normality, Equation 21 is the best linear unbiased estimator under the Gauss-Markov theorem.

$$Var\left(\hat{B}_j^{OLS}\right) \equiv E[(\hat{B}_j^{OLS} - \beta)^2] = \tau_B^2 + V_j \quad [24]$$

Re-arranging terms in Equation 24 suggests how we could "back out" a limited estimator ($\hat{\tau}^2_{B(LIMITED)}$) of τ_B^2 that uses information for a single site.

$$\hat{\tau}^2_{B(LIMITED)} = (\hat{B}_j^{OLS} - \beta)^2 - V_j \quad [25]$$

We can then pool these estimates across sites to obtain:

$$\hat{\tau}^2_{B(POOLED)} = \sum_{j=1}^{J} \frac{\left[(\hat{B}_j^{OLS} - \beta)^2 - V_j\right]}{J} \quad [26]$$

Equation 26 is a "method-of-moments" estimator in that it substitutes sample moments $(\hat{B}_j^{OLS} - \beta)^2$ for the expected values of these moments $E[(\hat{B}_j - \beta)^2]$ and solves for an estimator ($\hat{\tau}^2_{B(POOLED)}$). Although this approach is intuitively appealing, it raises two issues.

First is that we must know the value of b in order to estimate $\hat{\tau}^2_{B(POOLED)}$. In this case, it seems natural to substitute a sample-based estimate of β from Equation 21. However, Equation 21 requires knowledge of τ_B^2. This kind of chicken-and-egg problem implies the need for an iterative procedure.

A second issue is how to weight estimates from different sites in order to account for differences in their precision. For example, we might want to down-weight estimates from sites with small unbalanced samples and up-weight estimates from sites with large balanced samples, such that our final estimator ($\hat{\tau}_B^2$) is:

$$\tau^2_{B(FINAL)} = \sum_{j=1}^{J} \frac{w_j[(\hat{B}_j^{OLS} - \beta)^2 - V_j]}{\sum_{j=1}^{J} w_j} \quad [27]$$

As in meta-analysis, the optimal weight (w_j) for Equation 27 is $(\tau_B^2 + V_j)^{-2}$ under the assumption that program effects are normally distributed across sites (see Raudenbush, 1994 and Raudenbush and Bryk, 2002). Maximum likelihood analysis thus uses an iterative procedure which alternates between estimating Equation 21 and Equation 27.

Consequences of Heterogeneous Individual-level Outcome Variances

Suppose that the individual-level outcome variance for treatment group members (σ_1^2) differs from that for control group members (σ_0^2). This situation, which we refer to as T/C heteroskedasticity, can be caused by individual variation in program effects (see Raudenbush and Bloom, 2015). Also assume for the moment that σ_1^2 and σ_0^2 are constant across sites. Thus:

$$V_j = \frac{\sigma_1^2}{n_{1j}} + \frac{\sigma_0^2}{n_{0j}} \quad [28]$$

where n_{1j} and n_{0j} are the number of treatment and control group members at site j.

Equation 28 accounts for the influence of T/C heteroskedasticity on V_j. However, the typical default for a difference of means or an OLS regression is to estimate a single pooled individual-level error variance, which produces the following expression for "apparent" estimation error (V_j^A) at site j (see Appendix A):

$$V_j^A = \frac{\sigma_1^2}{n_{0j}} + \frac{\sigma_0^2}{n_{1j}} \quad [29]$$

Note that Equation 29 for the apparent estimation error variance (V_j^A) divides the treatment group outcome variance by the control group sample size and divides the control group outcome variance by the treatment group sample size, which is the reverse of Equation 28 for the true estimation error variance (V_j). Consequently, V_j^A is a biased estimating of V_j unless: (1) the treatment and control group samples are the same size ($n_{1j} = n_{0j}$) or (2) there is no T/C heteroskedasticity ($\sigma_1^2 = \sigma_0^2$). Thus:

$$V_j^A = V_j + Bias(V_j^A) \quad [30]$$

where, as demonstrated in Appendix A:

$$Bias(V_j^A) = \frac{2(\sigma_1^2 - \sigma_0^2)(\bar{T}_j - 0.5)}{n_j \bar{T}_j (1 - \bar{T}_j)} \quad [31]$$

Using V_j^A to represent site j's contribution to an estimate of τ_B^2 will thus produce a bias equal to -1 times $Bias(V_j^A)$. Consequently:

$$Bias\left(\tau_{Bj}^2\right) = -Bias(V_j^A) = \frac{2(\sigma_1^2 - \sigma_0^2)(0.5 - \bar{T}_j)}{n_j \bar{T}_j (1 - \bar{T}_j)} \quad [32]$$

Equation 32 indicates that the magnitude of this bias (which can be positive or negative) depends on the degree of T/C heteroskedasticity ($\sigma_1^2 - \sigma_0^2$); the degree of T/C sample imbalance ($(0.5 - \bar{T}_j)$ or ($\bar{T}_j(1 - \bar{T}_j)$); and site sample sizes (n_j). Fortunately, *we can avoid this bias* by specifying a separate individual-level outcome variance for treatment group members and control group member using existing software like HLM, SAS *PROC MIXED* or R. And no harm is done if $\sigma_1^2 = \sigma_0^2$ and we specify separate T/C outcome variances.

Not only does specifying a separate treatment group and control group individual-level outcome variance have a methodological advantage, it also can provide useful substantive information. For example, a treatment and control group difference in this variance is evidence that program impacts vary across individuals within sites

(Raudenbush and Bloom, 2015). Furthermore, if the treatment group variance is smaller than the control group variance, this is evidence that the program reduces disparities in the outcome of interest (Raudenbush and Bloom, 2015). To assess the statistical significance of the difference between treatment and control group estimates of these two variances one can use a simple F test of their ratio.

Now consider the issue of *cross-site heteroskedasticity*. Specifically, how should a researcher proceed if individual-level outcome variances differ among sites? To address this issue, one's first instinct might be to estimate separate treatment-group and control-group outcome variances *for each site*. However Appendix A demonstrates that doing so will tend to understate these outcome variances, which will cause us to understate V_j for our sites, which in turn will cause us to overstate τ_B^2, often by a lot. Simulation findings in Appendix Table A.1 demonstrate that this bias increases rapidly as site samples decrease and the number of sites increases. Because of the potential for large bias, we strongly recommend not estimating individual-level outcome variances by site unless site-specific sample sizes are very large.

Fortunately, Appendix A demonstrates that we can ignore cross-site heteroskedasticity and pool estimates of σ_1^2 and σ_0^2 across sites, as long as site sample sizes and individual-level outcome variances are approximately uncorrelated. In addition, Appendix Equation A.18 provides a way to estimate the magnitude of bias from this correlation. If this estimate is unacceptably large for a given dataset, we recommend stratifying sites by their sample sizes and estimating treatment and control group outcome variances *for each stratum*. In this way, a researcher can ensure small correlations between stratum sample sizes and individual-level outcome variances. When using this approach, we recommend that each site stratum contain at least 50 treatment group members and 50 control group member in order to reduce the potential for bias due to small stratum sample size.

Estimating Site-specific Mean Program Effects

For some purposes, researchers might want to identify and study the operation of sites with the most beneficial or least beneficial effects. Also, local policy makers might want to know the mean effect for *their* particular site. Thus it is sometimes important to produce site-specific estimates of program effects.

Classical statistics, based on unbiased estimation, confronts researchers with a forced choice between two estimates of the program effect for a specific site. The first option is the OLS estimator for that site ($\hat{B}_j^{OLS}$). However, the error variance (V_j) for this estimator can be very large if site samples are very small, as in the National Head Start Impact Study (Puma et. al., 2010) where the harmonic mean site sample size was 13 children (Bloom and Weiland, 2015).

However, if we knew that all sites had the same program effect, we could impute a site-specific value for B_j by setting it equal to the best existing estimate of the cross-site mean effect ($\hat{\beta}$); no one could do better than that! But what should we do if we must

allow for the possibility that program effects vary across sites? Must we be satisfied with $\hat{B}_j^{OLS}$?

Aversion to such a forced choice might lead a researcher toward a composite estimator that is superior to either the site-specific estimator or the cross-site mean estimator alone, which is what Bayes theorem can provide. Specifically, if we knew the values of τ_B^2, β and V_j and could estimate B_j from sample data, Bayes theorem tells us that the "best" estimate[11] of the mean program effect for a given site is its *posterior mean,* $\hat{B}_j^{PM}$, (Raudenbush and Bryk, 2002), where

$$\hat{B}_j^{PM} \equiv E\left(B_j \middle| Y, \beta, \tau_B^2, V_j\right) = \lambda_j \hat{B}_j^{OLS} + (1 - \lambda_j)\beta \qquad [33]$$

Equation 33 represents a weighted composite of the OLS estimate, $\hat{B}_j^{OLS}$, and the known value of the grand mean program effect, β. The weight accorded $\hat{B}_j^{OLS}$ is its reliability, λ_j and the weight accorded β is $(1 - \lambda_j)$, where

$$\lambda_j = \frac{\tau_B^2}{\tau_B^2 + V_j} \ . \qquad [34]$$

Holding τ_B^2 constant, reliability increases as V_j decreases. Holding V_j constant, reliability decreases as τ_B^2 decreases. In the extreme, when there is no cross-site impact variation ($\tau_B^2 = 0$), the reliability of a site-specific program effect estimate is zero ($\lambda_j = 0$) and all weight is placed on the cross-site mean effect (β).

In practice, we do not know the values of τ_B^2, β or V_j . But we can use consistent estimators of these parameters to compute an empirical Bayes impact estimator ($\hat{B}_j^{EB}$):

$$\hat{B}_j^{EB} = \hat{\lambda}_j \hat{B}_j^{OLS} + (1 - \hat{\lambda}_j)\hat{\beta} \qquad [35]$$

Furthermore, it is possible to quantify uncertainty about empirical Bayes estimates by computing their posterior "credibility intervals" (essentially confidence intervals) using the posterior standard error of $\hat{B}_j^{EB}$ (Raudenbush and Bryk, 2002). If the values of τ_B^2, β and V_j were known, this standard error would equal $\tau_B^2\sqrt{1 - \lambda_j}$. Therefore posterior credibility intervals are narrow when site-level reliability is high and thus $\hat{B}_j^{OLS}$ is a good estimate of B_j; or when site-specific impacts are nearly homogenous, in which case $\beta$is a good approximation of B_j. When in practice, we estimate τ_B^2, β and V_j with uncertainty, the expression for the posterior standard error becomes more complex and the interval becomes larger to reflect this uncertainty. Software is available for

[11] In this case, we define the best estimate to be the value of $\hat{B}_j^{PM}$ which minimizes the expected sum of squared errors of estimation, ($E[\sum_{j=1}^{J}(\hat{B}_j^{PM} - B_j)^2]$). Bayes theorem tells us that this optimal value is the posterior mean defined by Equation 33 which follows.

Bayesian methods that address this concern (e.g. Spiegelhalter, Thomas, Best and Lunn, 2003 and Gelman, Hill and Yajima, 2012).

Reporting a Cross-Site Impact Distribution

Although empirical Bayes estimation can be useful for assessing site-specific program effects, a histogram of empirical Bayes estimates for a sample of sites will tend to *understate* true cross-site impact variation. At the opposite extreme, a histogram of site-specific OLS impact estimates will tend to *overstate* true cross-site impact variation – often by a lot. This is because:

$$Var(\hat{B}_j^{EB}) < \tau_B^2 < Var(\hat{B}_j^{OLS}) \qquad [36]$$

To better represent a cross-site distribution of program effects, one can constrain empirical Bayes estimates in a way that ensures their variance for a given sample equals the estimated value of τ_B^2 for that sample. This idea was introduced by Louis (1984) and a simple approach for making its operational is described in Appendix B and illustrated in our empirical example below.

A Caveat

Although the present approach to studying a cross-site distribution of program assignment effects is promising for many applications, it has a limitation that could be important in some cases. This limitation involves the precision weights used to combine site contributions to parameter estimates.

Specifically, the present approach will provide consistent estimates when *site-specific precision weights are uncorrelated with site program effects* (Raudenbush and Bloom, 2015 and Raudenbush, 2015). But if for some reason, site weights and program effects are correlated, the present method will not produce consistent estimates. For example, if sites with larger-than-average program effects have larger-than average weights, the present method will tend to over-state the cross-site mean effect (β). Recall that the site weight for estimating a random-coefficient mean effect is $(\hat{\tau}_B^2 + \hat{V}_j)^{-1}$, where $\hat{V}_j$ depends on the site sample size (n_j) and treatment allocation ($\bar{T}_{ij}$). Consequently, the issue boils down to the strength of the cross-site covariance between true program effects (B_j) and n_j or $\bar{T}_j$. Note that this problem applies with even greater force to a standard fixed-effect estimator of mean program effects (e.g. an OLS regression with a single treatment assignment indicator) because it weights each site's impact estimate inversely proportionally to $\hat{V}_j$.

One way to eliminate this potential inconsistency is to use a random-coefficients model that weights each site's impact estimates *equally* (see Raudenbush and Bloom, 2015 and Raudenbush, 2015). However, if site sample sizes or treatment allocations vary substantially – which they often do – weighting sites equally can reduce precision appreciably. Thus we are faced with a tradeoff between potential inconsistency from the present method and a potential loss of precision from a method that weights sites equally.

Because little is currently known about this tradeoff in practice, we and our colleagues are conducting a study of the tradeoffs, using simulation studies and re-analyses of a substantial number of multi-site trials in education research and related fields.

Empirical Example: Variation in Welfare-to-Work Program Effects

This section uses the preceding ideas to study cross-site variation in the effects of welfare-to-work programs by pooling data from three multi-site trials conducted by MDRC over more than a decade: The Greater Avenues for Independence (GAIN) project conducted in 22 local welfare offices from six California counties (Riccio and Friedlander, 1992); Project Independence conducted in 10 local welfare offices from nine Florida counties (Kemple and Haimson, 1994); and the National Evaluation of Welfare-to-Work Strategies conducted in 27 local welfare offices from seven states (Hamilton, 2002).

Background

Because these new programs were mandatory, all treatment group members were exposed to them, even if only to their threat of sanctions (loss of welfare payments) for non-participation. Furthermore, no control group members experienced the new programs, although some might have received related services elsewhere. Compliance with random assignment was therefore complete and the average effect of program assignment was thus equal to the average effect of program participation.

The outcome measure for the present analysis is sample members' total earnings during their first two years after random assignment, reported in constant dollars.[12] Data for this outcome were obtained from quarterly administrative records of the State Unemployment Insurance agency for each local program. The pooled cross-study sample contains 59 sites (local welfare offices) with a total of 69,399 individuals randomized by site to a new mandatory welfare-work-program or a control group.

Findings for the present analysis were obtained by using HLM to estimate the two-level site-centered model represented by Equations 9 and 10 with a separate individual-level outcome variance for treatment group members and control group members. As noted, this is equivalent to estimating the two-level fixed-site-intercept, random-coefficient model represented by Equations 11 – 13, with a separate individual-level outcome variance for treatment group members and control group members.

Findings

The estimated cross-site mean program effect ($\hat{\beta}$) was \$875 and the estimated cross-site variance of program effects ($\hat{\tau}_B^2$) was $(742)^2$, both of which are statistically

[12] Our findings are based on data from Bloom, Hill and Riccio (2003), which are reported in 1996 dollars. This metric was maintained to ensure comparability with the original results.

significant at well beyond the 0.001 level.[13]The estimated variance of individual-level outcomes for treatment group members ($\hat{\sigma}_1^2$) was $(10{,}068)^2$ and that for control group members ($\hat{\sigma}_0^2$) was $(9{,}643)^2$. Their ratio of 1.09, which is statistically significantly different from one at beyond the 0.001 level, which indicates that program effects varied across individuals within sites.[14]

To further explore the cross-site impact distribution and to illustrate an important point about representing such a distribution, Figure 1 presents three histograms. The top histogram summarizes our site-specific OLS impact estimates ($\hat{B}_j^{OLS}$).[15] Note that it overstates cross-site impact variation because cross-site variation in $\hat{B}_j^{OLS}$ reflects true impact variation (τ_B^2) plus site-level estimation error (V_j).

The bottom histogram summarizes our site-specific empirical Bayes estimates ($\hat{B}_j^{EB}$), which "shrink" each OLS estimate toward the estimated cross-site mean ($\hat{\beta}$). Note that the empirical Bayes estimates vary by much less than the OLS estimates. This is because an empirical Bayes estimator "nets out" cross-site variation due to site-level estimation error (V_j).

However, even though empirical Bayes estimators have the smallest mean squared error for predicting a specific parameter value, like the mean program effect for a given site (Lindley and Smith, 1972), these estimators are biased toward the overall mean (Raudenbush and Bryk, 2002). Consequently, the cross-site variance of empirical Bayes estimates tends to *understate* the cross-site variance of true mean program effects. Thus for studying a cross-site impact distribution, empirical Bayes estimates "over-shrink" their OLS counterparts.[16] Consequently, the sample variance of empirical Bayes estimates for the present analysis ($\hat{Var}(\hat{B}_j^{EB})$) is only 91 percent of the estimated variance of true mean program effects ($\hat{\tau}_B^2$), where:

$$\hat{Var}(\hat{B}_j^{EB}) = \left(\frac{\sum_j (\hat{B}_j^{EB} - \hat{\beta})^2}{J}\right) \qquad [38]$$

To compensate for this over-shrinkage, the middle histogram in Figure 1 adjusts empirical Bayes estimates in a way that *constrains* their sample variance to equal the model-based estimate of the variance of true program effects ($\hat{\tau}_B^2$). Appendix B derives this adjustment, which "stretches" the distance between each empirical Bayes estimate and the estimated cross-site mean by a proportion, γ, where:

$$\gamma = \frac{\hat{Var}(\hat{B}_j^{EB})}{\hat{\tau}_B^2} \qquad [39]$$

---

[13] The statistical significance of $\hat{\beta}$ was based on its t-statistic and the statistical significance of $\hat{\tau}_B^2$ was based on its Q-statistic.

[14] The statistical significance of this ratio was determined by an F-statistic.

[15] These estimates were obtained from a pooled-sample model with fixed site-specific intercepts, fixed site-specific impact coefficients and a separate outcome variance for treatment and control group members

[16] Raudenbush and Bryk (2002) page 88 discuss these issues for a two-level hierarchical model.

Each constrained empirical Bayes estimate $\hat{B}_j^{CEB}$ is obtained as follows:

$$\hat{B}_j^{CEB} = \hat{\beta} + \frac{1}{\sqrt{\gamma}}(\hat{B}_j^{EB} - \hat{\beta}) \qquad [40]$$

This adjustment is similar to that in the literature (e.g. Louis, 1984 and Rao, 2003).

Now compare the three histograms. As can be seen, they all have approximately the same cross-site mean, which ranges from $875 to $897. However, they reflect widely differing amounts of cross-site variation, with OLS estimates having a standard deviation of $1,209, empirical Bayes estimates having a standard deviation of $708 and constrained empirical Bayes estimates having a standard deviation of $742.

Because constrained empirical Bayes estimates reflect cross-site variation most accurately, they are the best guide for interpreting findings about a cross-site impact *distribution*. These findings suggest that only about 6 of the 59 local programs examined have negative effects (i.e. were less effective than existing alternatives).

Diagnostics

Figures 2 and 3 provide some useful diagnostic for further examining the preceding findings. Figure 2 presents a "caterpillar plot" of our site-specific empirical Bayes impact estimates. This plot, which was produced in HLM, is a simple way to illustrate what is known and not known about site-specific impacts.[17] Sites are represented in order from lowest to highest estimated impact with a square representing each empirical Bayes estimate and two vertical lines around each square representing its 95 percent posterior credibility interval (a type of confidence interval). The more cross-site impact variation there is, the steeper the slope of the empirical Bayes estimates will be; he more precise these estimates are, the narrower the confidence band around the slope will be. Given the present large site samples, Figure 2 has a narrow band around a clearly discernible slope. This indicates that we can distinguish among site impacts.

A second tool for assessing site-specific impact estimates is the "profile likelihood" plot in Figure 3 (Murphy and Vander der Vaart, 2000). This plot, which was produced in HLM, illustrates how the empirical Bayes estimates ($\hat{B}_j^{EB}$) for our 59 sites vary as a function of alternative values for $\hat{\tau}_B^2$. This is accomplished by plotting alternative possible empirical Bayes estimates for each site on the vertical axis as a function of alternative values for $\hat{\tau}_B^2$ on the horizontal axis. The resulting alternative empirical Bayes estimates for each site are smoothed by a line that expands from a single point for $\hat{\tau}_B^2 = 0$ to a broad band as $\hat{\tau}_B^2$ increases. This pattern illustrates a fundamental property of empirical Bayes estimates, that other things being equal, they vary more across sites as $\hat{\tau}_B^2$ increases. This is because as $\hat{\tau}_B^2$ increases the weight placed on $\hat{B}_j^{OLS}$

[17] Other existing software can also produce caterpillar plots.

increases relative to the weight placed on $\hat{\beta}$. Thus when $\hat{\tau}_B^2$ equals zero all $\hat{B}_j^{EB}$ converge to $\hat{\beta}$ and as $\hat{\tau}_B^2$ increases these values diverge toward $\hat{B}_j^{OLS}$.

What makes a profile likelihood plot particularly useful is the superimposed plot of the profile likelihood function evaluated for each possible value of $\hat{\tau}_B^2$ (Rubin, 1981). This is the inverted U-shaped curve in the figure. The vertical axis of this curve provides a measure of the relative plausibility of each possible value of $\hat{\tau}_B^2$ given the existing data. The steeper and narrower this likelihood profile is, the less uncertainty there is about the true value of τ_B^2 and in turn, the less uncertainty there is about each site-specific empirical Bayes estimate. The profile likelihood for the present example is symmetric and peaked at a point that is substantially different from zero. This indicates that we can be confident that program effects vary appreciably across sites. This conclusion is consistent with the fact that the 95 percent confidence interval produced by HLM for our estimate of τ_B^2 ranges from $(525)^2$ to $(1{,}048)^2$, which implies a 95 percent confidence interval for τ_B of \$525 to \$1,048.[18]

This relatively high degree of confidence about the presence of substantial cross-site impact variation reflects the very large sample for the present analysis (69,399 randomized individuals from 59 sites). Other situations will present greater uncertainty and the preceding diagnostic tools can help researchers to assess this uncertainty.

A profile likelihood plot also can help us to identify sites with especially large or small impacts. In Figure 3 there are several sites with impacts near \$1700, which is about twice the cross-site average; and there are several sites with impacts near zero. And over the plausible range of values for τ_B^2 (i.e., values for which the likelihood is substantial), there is little variation in the determination of these sites.

Lastly, consider the implications for our welfare-to-work example of two potential methodological issues discussed earlier. Note first that cross-site heteroskedasticity does not appear to be a problem. Pooling estimates of σ_1^2 and σ_0^2 across sites produced only a seven percent bias in our estimate of τ_B^2 and thus a four percent bias in our estimate of τ_B (see Appendix A). Consequently, it was not necessary to stratify sites and estimate stratum-specific values for σ_1^2 and σ_0^2.

Second, there does not appear to be a problematic bias from correlation between site HLM precision weights and site mean program effects. Evidence of this is that our cross-site mean impact estimate (\$875) is very similar to an equally-weighted mean estimate (\$906). Furthermore, our estimate is more precise than its equally-weighted

[18] Now-standard software for hierarchical linear models uses the Fisher information matrix to estimate standard errors for estimated variances. However these estimated standard errors are not generally useful for computing confidence intervals for variances because they are bounded by zero. To obtain such confidence intervals, we instead use a logarithmic transformation of the variance and the associated transformation of the information matrix. We then exponentiate the upper and lower limits of this interval to obtain an appropriate asymmetric confidence interval for the variance.

alternative, with an estimated standard error of 139 versus 156.[19] Consequently, an equally-weighted estimator would need a sample that is 26 percent larger than that for our present estimator in order to achieve the same precision.[20]

Concluding Thoughts

The primary contributions of the present paper are: (1) presentation of a promising statistical method for using data from multi-site randomized trials to detect and quantify cross-site variation in program assignment effects, (2) a detailed intuitive discussion of the statistical issues that underlie the method, and (3) a real-world empirical example of how the method works and the types of information that it can provide.

The method we propose distinguishes between variation in site-level *estimates* of program effects (which reflects true effect variation *plus* variation due to estimation error) and variation in *true site-level program effects.* Application of this method can be a valuable component of multi-site evaluations of ongoing pubic programs like Head Start (Puma et. al, 2010) or Job Corps (Schochet et. al, 2008); multi-site studies of large-scale government interventions like New York City's Small High Schools of Choice (Bloom and Unterman, 2014) or Massachusetts' charter schools (Angrist et. al, 2013); and multi-site studies of demonstration projects like the Enhanced Reading Opportunities demonstration for high school students (Somers et. al, 2010) or the enhanced coaching project for college students (Bettinger and Baker, 2014).

One issue to consider when *designing* future multi-site trials to study cross-site impact variation is the number and size of site samples needed for adequate statistical power. Although this issue is beyond the scope of the present paper, analyses reported elsewhere (Bloom and Spybrook, 2013; Bloom, Raudenbush and Reardon, 2014; and Raudenbush and Liu, 2000) provide some guidance. These analyses indicate for example, that a fully-balanced trial with 20 sites and 100 sample members per site (2,000 total sample members) or with 50 sites and 50 sample members per site (2,500 total sample members) has a minimum detectable cross-site effect-size standard deviation (MDESSD) of about 0.16σ.[21]

To help interpret this finding note that the estimated cross-site standard deviation of effect sizes is 0.08σ for our welfare-to-work example.[22] Note also that estimates of the

---

[19] The estimated standard error ($\widehat{se}(\hat{\beta}^{EW})$) of the equally-weighted mean impact estimate ($\hat{\beta}^{EW}$) was computed as $\widehat{se}(\hat{\beta}^{EW}) = \sqrt{\frac{\sum_{j=1}^{J}(B_j^{OLS}-\hat{\beta}^{EW})^2/(J-1)}{J}}$

[20] To see this, note that the error variance of an estimator (its standard error squared) is inversely proportional to its sample size and $\frac{(\widehat{se}(\hat{\beta}^{EW}))^2}{(\widehat{se}(\hat{\beta}^{HLM}))^2} = \frac{(156)^2}{(139)^2} = 1.26$.

[21] This minimum detectable effect size standard deviation assumes: (1) a two-sided test of statistical significance at the 0.05 level, (2) 80 percent statistical power and (3) a joint individual-level R^2 value of 0.5 for individual-level covariates and site indicators).

[22] This was computed by dividing the estimated cross-site standard deviation of welfare-to-work program effects ($742) by the control-group standard deviation of individual outcomes ($8961).

cross-site standard deviation of Head Start program assignment effect sizes for four cognitive outcomes and two socio-emotional outcomes reported by Bloom and Weiland (2014) range from 0.07σ to 0.25σ. These findings bracket the preceding MDESSD of 0.16σ, which suggests that cross-site variation of that magnitude might be a reasonable expectation for some important situations. If so, then study samples with 2,000 or more persons and 20 or more sites might be adequately powered to detect cross-site impact variation of a realistic magnitude.[23]

One issue to consider when *analyzing* future multi-site trials to study cross-site impact variation is whether an "omnibus" statistical test of the existence of such variation should be used as a "gateway" criterion for deciding whether to try to predict it.[24] We recommend that such an omnibus test *not be used* for this purpose when considering an a priori theory-based hypothesis about a site-level impact predictor. This is because an omnibus test can have less statistical power than a "focused" test about a difference between mean program effects for specific subgroups of sites (e.g. rural versus urban). Hedges and Pigott (2001) note this power difference in the context of meta-analysis, Rosenthal and Rosnow (1985) note it in the context of experimental design, and Appendix C provides insights into the factors that influence this difference in power. On the other hand, if an omnibus test fails to detect cross-site impact variation, it should serve as an important *caveat* for ex post exploratory hypothesis tests about site-level impact predictors.

Lastly, we note the potential limitation of the present method that was identified earlier – the fact that it could provide inconsistent or biased estimates if there is an appreciable correlation between site precision weights and true program effects (Raudenbush and Bloom, 2015 and Raudenbush, 2015). In that case, a researcher might want to consider methods that weight sites equally. However, as discussed, these alternative methods will have less precision than the present method because they weight sites with precise impact estimates the same as sites with imprecise estimates. Consequently, it is important to examine the bias/precision tradeoff between these methods. Although this trade-off was favorable for the present method in our welfare-to-work example, it remains to be seen what the tradeoff will look like for a broader range of programs, participant populations and local settings. This issue is currently being explored by the present authors and their colleagues using data from numerous multi-site trials in education and related areas as part of a project funded by the Spencer Foundation and the William T. Grant Foundation.

As these and other statistical issues are addressed by future research, it will be equally important to develop realistic but practical conceptual frameworks for studying the predictors of program effects. Likewise it will be essential for the next generation of multi-site trials to collect high-quality data on these predictors. We have confidence that

[23] Note that the MDESSD for a given study depends not only its number of sites and their average sample size, but also on the variation in site sample sizes and the proportion of sample members randomized to treatment or control status (see Bloom and Spybrook; 2013).

[24] For a conceptual discussion of potential predictors of cross-site impact variation see Weiss, Bloom and Brock (2014); for an empirical study of these predictors see Bloom, Hill and Riccio (2003).

together, these new statistical methods, conceptual frameworks and high-quality data can produce the knowledge that is needed by policy makers, practitioners and researchers to understand when, how and why programs do or do not work.

Figure 1

Cross-site Distributions of Estimated Effects of 59 Welfare-to-Work Programs

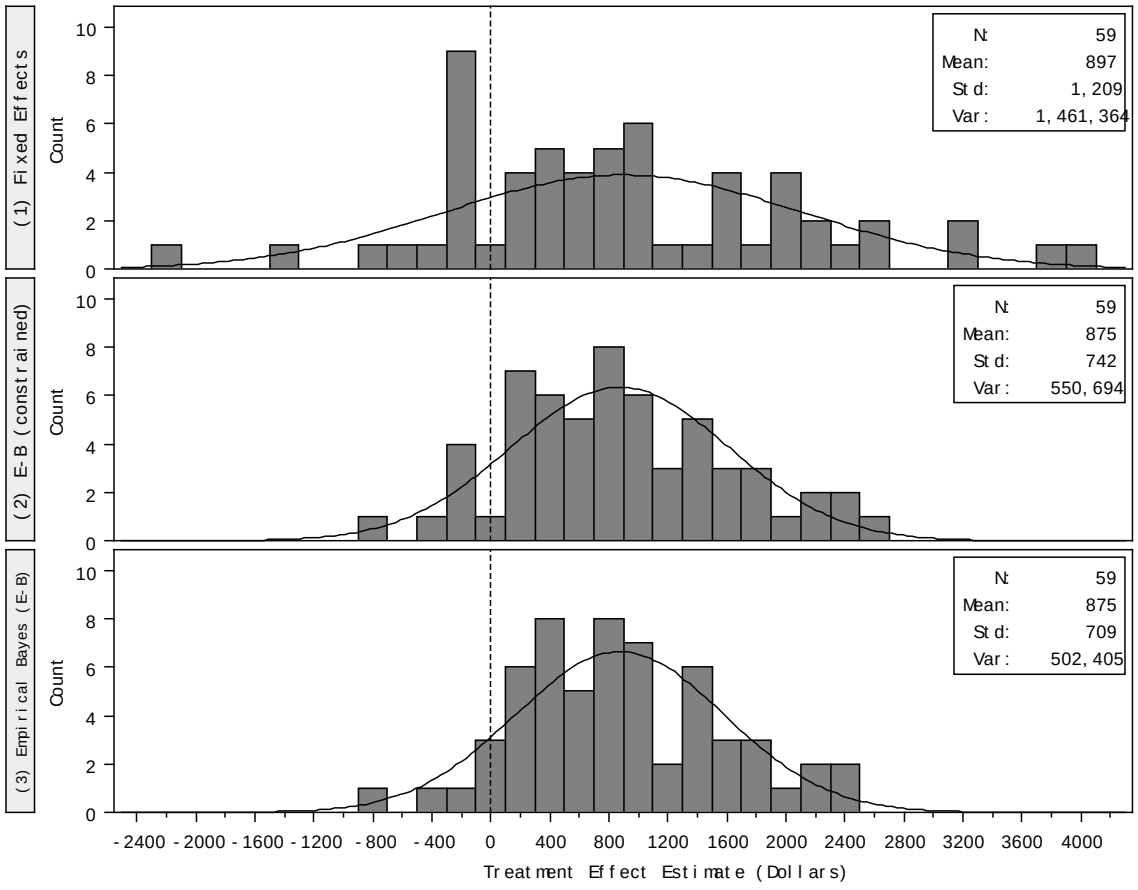


Figure 2

Caterpillar Plot of Empirical Bayes Estimates of the Effects of 59 Welfare-to-Work Programs

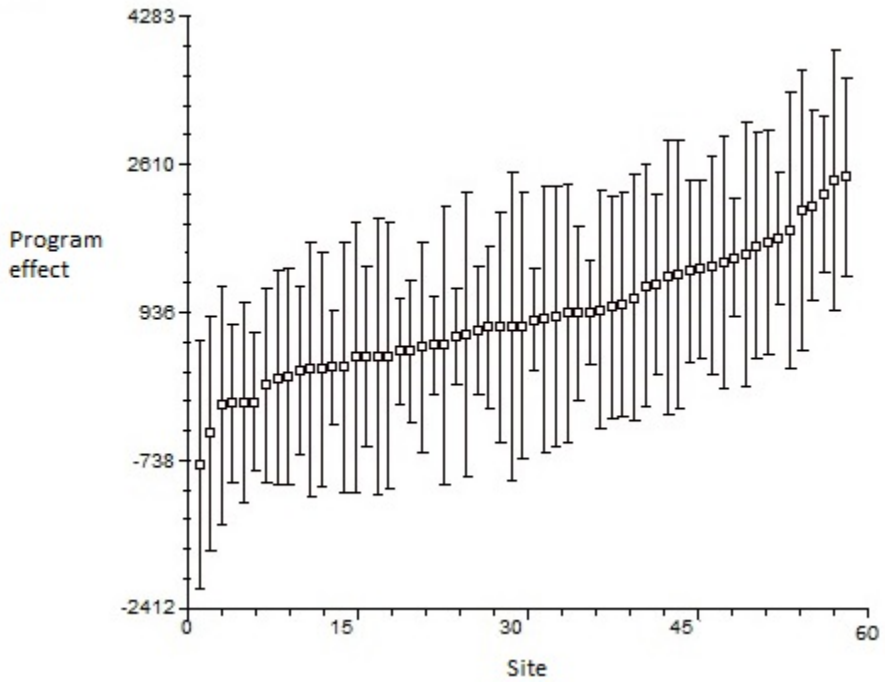
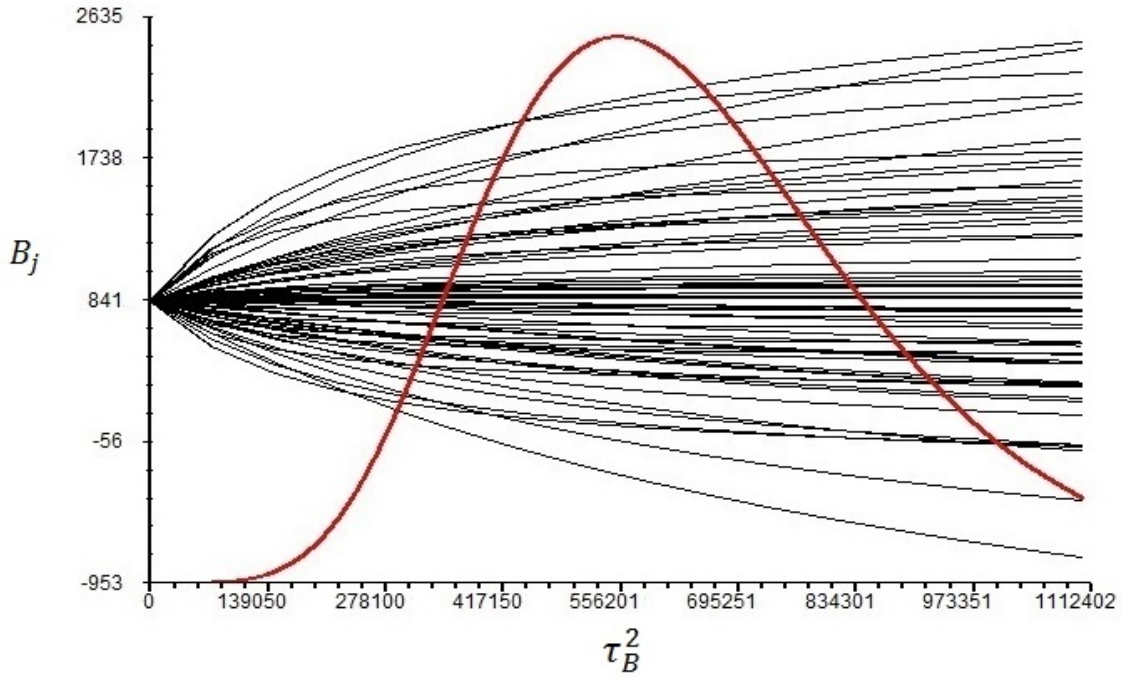


Figure 3

Profile Likelihood Graph of Empirical Bayes Estimates of the Effects of 59 Welfare-to-Work Programs



References

Abadie, A., Angrist, J.D., & Imbens G. (2002). Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earnings. *Econometrica*, 70, 91-117.

Angrist, J.D., Pathak, P., & Walters C.R. (2013). Explaining charter school Effectiveness. Working paper. *American Economic Journal: Applied Economics*.

Berger, M., Black, D., & Smith J. (2001). Evaluating profiling as a means of allocating government services. In M. Lechner and F. Pfeiffer (Eds.), *Econometric Evaluation of Active Labor Market Policies, Physica* (59 – 84).

Bettinger, E. P., & Baker, R. (2014). The Effects of Student Coaching An Evaluation of a Randomized Experiment in Student Advising *Educational Evaluation and Policy Analysis., 36*(1), 3-19.

Bitler, M. P. & Hoynes, H.W. (2006). What mean impacts miss: Distributional effects of welfare reform experiments. *American Economic Review*, 96, 988 – 1012.

Bitler, M.P. H.W. Hoynes and T. Domina (2014) "Experimental Evidence on Distributional Effects of Head Start," NBER Working Paper 20434, August.

Bloom, H.S. (1984). Accounting for no-shows in experimental evaluation designs. *Evaluation Review* 8(2), 225-246.

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., 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 Managemen,t* 22(4), 551-575.

Bloom, H.S. (2003). Using "short" interrupted yime-series analysis to measure the impacts of whole-school reforms: With applications to a study of accelerated schools. *Evaluation Review*, 27(1), 3-49.

Bloom, H.S. & Michalopoulos, C. (2011). When is the story in the subgroups? Strategies for interpreting and reporting intervention effects on subgroups. *Prevention Science*.

Bloom, H.S. and J. Spybrook (2013) "Statistical Power/Precision for Multi-site Trials," Presentation at the University of Chicago to the Workshop on *Learning from*

Variation in Program Effects funded by the William T. Grant Foundation, October 7.

Bloom, H.S., S.W. Raudenbush and S.F. Reardon (2014) "Detecting and Quantifying Variation in Effects of Program Assignment," Presentation at a Workshop in Washington DC on *Using Cross-site Variation in Program Effects to Study What Works for Whom and Under What Conditions* as part of the spring conference of the Society for Research on Educational Effectiveness, March 8.

Bloom, H.S. and C. Weiland (2015) "Quantifying Variation in Head Start Effects on Young Children's Cognitive and Socio-Emotional Skills Using Data from the National Head Start Impact Study," New York: MDRC, March.

Brachet, T. (2007). Documentation for computing clustered standard errors for two-stage least squares in SAS. Retrieved [insert retrieval data] from http://works.bepress.com/tbrachet/2/.

Bryk, A.S. & Raudenbush, S.W. (1988). Heterogeneity of variance in experimental studies: A challenge to conventional interpretations. *Psychological Bulletin*, 104 (3), 396-404.

Buchinsky, M. (1998). Recent advances in quantile regression models: A practical guide for empirical research. *The Journal of Human Resources*, 33(1), 88-126.

Cooper, H. & Hedges, L.V. (Eds.). (1994). *The Handbook of Research Synthesis*. New York: Russell Sage Foundation.

Djebbari, H. & Smith, J. (2008). Heterogeneous impacts in PROGRESA. *Journal of Econometrics*, 145, 64-80.

Eberts, R. and O'Leary, C. (2002). Targeting employment services. Kalamazoo, MI: Upjohn Institute for Employment Research.

Ehren, B.J., Deshler, D.D., & Graner, P.S. (2010). Using the content literacy continuum as a framework for implementing RTI in secondary schools. *Theory into Practice,* 49(4), 315-322.

Fisher, R.A. (1935). *The Design of Experiments*. London: Oliver and Boyd.

Friedlander, D. (1993). Subgroup impacts of large-scale welfare employment programs. *The Review of Economics and Statistics*, 75(1), 138-143.

Friedlander, D. & Robins, P.K. (1997). The distributional impacts of social programs. *Evaluation Review*, 21(5), 531-553.

Gelman, A.,Hill, J. and Yajima, M. (2012). Why we (usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness*, 5(2): 189-211.

Gueron, J.M. & Pauly, E. (1991). *From Welfare to Wor*k**.** New York: Russell Sage Foundation.

Haager, D., Klingner, J., & Vaughn S. (2007). *Evidence-Based Reading Practices for Response to Intervention*. Baltimore, MD: Paul H. Brookes Publishing Co., Inc.

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 and U.S. Department of Education, Office of the Under Secretary and Office of Vocational and Adult Education.

Heckman, J.J. (2001). Micro data, heterogeneity, and the evaluation of public policy: Nobel lecture. *Journal of Political Economy*, 109(4), 673-748.

Heckman, J.J. (2005). The scientific model of causality. *Sociological Methodology* 35(1), 1-97.

Heckman, J.J., Schmierer, D.A., & Urzua S.S. (2009). Testing the correlated random coefficient model. NBER Working Paper #15463, Cambridge, MA: National Bureau of Economic Research.

Hedges, L.V. & Olkin, I. (1985). *Statistical Methods for Meta Analysis*. San Diego, CA: Academic Press, Inc.

Hedges, L.V. and T.D. Pigott (2001) "The Power of Statistical Tests in Meta-Analysis," *Psychological Methods*, 6(3): 203 – 217.

Holland, P. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 81(396), 945 – 960.

Horwitz, R.I., Singer, B.H. Makuch, R.W., & Viscoli C.M. (1996). Can treatment that is helpful on average be harmful to some patients? A study of the conflicting information needs of clinical inquiry and drug regulation. *Journal of Clinical Epidemiology*, 49(4), 395-400.

Kemple, J.J., Snipes, J.C., & Bloom, H.S. (2001). A regression-based strategy for defining subgroups in a social experiment. New York: MDRC.

Kemple, J. & Haimson, J. (1994). Florida's Project Independence: Program implementation, participation patterns and first-year impacts. New York: MDRC.

Kirk, R.E. (1982). *Experimental Design*. Hoboken, NJ: John Wiley & Sons.

Koenker, R. (2005). *Quantile Regression*. New York: Cambridge University Press.

Konstantopoulos, S. (2001). How consistent are class size effects? *Evaluation Review*, 35(1), 71-92.

Lake, R., Bowen, M., Demeritt, A., McCullough, M., Haimson, J., & Gill B. (2012). Learning from charter school management organizations: Strategies for student behavior and teacher coaching. Princeton, NJ: Mathematica Policy Research.

Lindley, D. V. & Smith A. F. M. (1972). Bayes estimates for the linear model. *Journal of the Royal Statistical Society*, Series B, 34, 1- 41.

Lipsey, M.W. & Wilson, D.B. (2001). *Practical Meta-Analysis*. Thousand Oaks, CA: Sage Publications.

Louis, T. A. (1984). Estimating a population of parameter values using Bayes and Empirical Bayes methods. *Journal of the American Statistical Association*, 79, 393-398.

May, H., Gray, A., Gillespie, J.N., Sirinides, P., Sam, C., Goldsworthy, H., Amijo, M., & Tagnotta N. (2013). Evaluation of the i3 scale-up of eading Recovery: Year one report, 2011-12. Consortium for Policy Research in Education, August.

Michalopoulos, C. & Schwartz, C. (2000). What works best for whom: Impacts of 20 welfare-to-work programs by subgroup. Washington, DC: : U.S. Department of Health and Human Services, Office of the Assistant Secretary for Planning and Evaluation and Administration for Children and Families, and U.S. Department of Education.

Murphy, S.A. & Van der Vaart A.W. (2000). On Profile Likelihood. *Journal of the American Statistical Association*, 95, 449 – 465.

Neyman, J. (1923). Statistical Problems in Agricultural Experiments. *Journal of the Royal Statistical Society*, 2, 107 -180.

Puma, M., Bell, S., Cook, R., & Heid, C. (2010). Head Start impact study final report. Washington, DC: Prepared for the Office of Planning, Research and Evaluation of the Administration for Children and Families of the U.S. Department of Health and Human Services, January.

Quandt, R. (1972). Methods of estimating switching regressions. *Journal of the American Statistical Association*, 67,306 – 310.

Rao, J.N.K. (2003). *Small Area Estimation*. Hoboken, NJ: John Wiley, Section 9.6.1.

Raudenbush, S.W. (2015). *Consistent Estimation of Means and Covariance Components in Multi-site Randomized Trials*. Occasional paper, University of Chicago Committee on Education.

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

Raudenbush, S.W. & Bryk, A.S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods*, 2nd Edition. Thousand Oaks, CA: Sage Publications.

Raudenbush, S.W. and H.S. Bloom (2015) "Learning About and From a Distribution of Program Impacts Using Multisite Trials," *American Journal of Evaluation*, DOI: 10.1177/1098214015600515.

Reardon, S. F., and Raudenbush S. W. (2013). Under what assumptions do site-by-treatment instruments identify average causal effects? *Sociological Methods and Research*. Vol. 42, 2:pp 143-163.

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

Rosenthal, R. (1991). *Meta-Analytic Procedures for Social Research*. Newbury Park, CA: Sage Publications.

Rosenthal, R., & Rosnow, R. L. (1985). *Contrast analysis: Focused comparisons in the analysis of variance*. CUP Archive.

Rothwell, P.M. (2005). Subgroup analysis in randomised control trials: Importance, indications and interpretation. *The Lancet*, 365, 176 – 86.

Roy, A. D. (1951). Some thoughts on the distribution of earnings. *Oxford Economic Papers*, 3, 135 – 146.

Rubin, D. (1974). Estimating causal effects of rreatments in randomized and nonrandomized studies. *Journal of Education Psychology*, 66, 688 – 701.

Rubin, D. (1978). Bayesian inference for causal effects: The role of randomization. *Annals of Statistics*, 6(1), 34 – 58.

Rubin, D. (1981). Estimation in parallel randomized experiments. J*ournal of Educational Statistics*, 6 (4), 377 – 401.

Seltzer, M. H., Wong, W.W. & Bryk, A.S. (1995). Bayesian analysis in applications of hierarchical models: Issues and methods. *Journal of Educational and Behavioral Statistics,* 21(2), 131-167.

Schochet, P.Z., Burghardt, J., & McConnell S. (2008).”Does Job Corps Work? Impact Findings from the National Job Corps Study,” American Economic Review, 98(5): 1864 – 1886.

Somers, Marie-Andrée, William Corrin, Susan Sepanik, Terry Salinger, Jesse Levin, and Courtney Zmach. 2010. *The Enhanced Reading Opportunities Study Final Report: The Impact of Supplemental Literacy Courses for Struggling Ninth-Grade Readers* (NCEE 2010-4021). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance

Spybrook, J. (2013). Detecting intervention effects across context: An examination of the precision of cluster randomized trials. *The Journal of Experimental Education*.

Tipton, E**.** (2014) How generalizable is your experiment? Comparing a sample and population through a generalizability index. *Journal of Educational and Behavioral Statistics, 39*(6): 478 – 501.

Walters, C. R. (2015) “Inputs in the production of early childhood human capital: evidence from Head Start,” *American Economic Journal: Applied Economics*, 7(4): 76 – 102.

Weiss, M. J., H. S. Bloom, and T. Brock (2014) "A Conceptual Framework for Studying the Sources of Variation in Program Effects." *Journal of Policy Analysis and Management* 33, 3, 778-808.

Appendix A

Dealing With Heterogeneous Outcome Variances

This appendix demonstrates that:

- Conclusion #1: When it exists, *it is necessary* to account for a treatment and control group difference in individual-level outcome variances (T/C heteroskedasticity) in order to obtain an unbiased estimator of the cross-site variance of program effects (τ_B^2). When this issue does not exist, taking this step, which is simple to do using existing software like HLM, SAS PROC MIXED or R, does not hurt. Furthermore, documenting a T/C difference in individual-level outcome variances can provide an important substantive finding.

- Conclusion #2: When they exist, *it may not necessary* to account for cross-site differences in individual-level outcome variances (cross-site heteroskedasticity) in order to obtain an unbiased estimator of the cross-site variance of program effects (τ_B^2). In addition, there is a straight-forward empirical test of the magnitude of the bias produced by ignoring these differences.

- Conclusion #3: Trying to account for cross-site heteroskedasticity by estimating site-specific individual-level outcome variances can cause one to *overstate* the cross-site variance of program effects (τ_B^2), potentially by a lot. This is especially the case for trials with site samples of 50 persons or less and large numbers of sites.

- Recommendation: One should use Equation A.18 below to estimate the bias from ignoring cross-site heteroskedasticity for a given dataset. If this bias appears to be substantial relative to the magnitude of $\hat{\tau}_B^2$ one should stratify sites by sample size and estimate a separate individual-level outcome variance for each stratum's treatment and control group. If not, then one can pool individual-level variance estimates across sites.

Conclusion #1: We must account for T/C heteroskedasticity when it exists and doing so does not hurt when this issue does not exist. Furthermore, treatment and control group differences in individual-level outcome variances can be substantively important.

Recall that our aim is to estimate t_B^2 as the following method-of-moments estimator:

$$t_B^2 = \frac{\sum_{j=1}^{J} w_j\left[(\hat{B}_j - b)^2 - V_j\right]}{\sum_{j=1}^{J} w_j} \qquad \text{A.1}$$

where $B_j = \mu_j - \mu_0 \; \hat{B}_j = \bar{Y}_{1j} - \bar{Y}_{0j}$, μ_{1j} and μ_{0j} are the population mean outcomes for treatment and control cases at site j, $\bar{Y}_{1j}$ and $\bar{Y}_{0j}$ are corresponding sample mean outcomes and V_j is the estimation error variance of $\hat{B}_j$. (For simplicity we assume that β is known, whereas in practice it must be estimated.) In order for this estimator of τ_B^2 to be unbiased our estimator of V_j for each site must be unbiased.

Note first, that when our treatment and control groups have different individual-level outcome variances:

$$V_j \equiv Var\left(\hat{B}_j | B_j\right) = \frac{\sigma_1^2}{n_j \bar{T}_j} + \frac{\sigma_0^2}{n_j(1-\bar{T}_j)} \qquad \text{A.2}$$

where σ_1^2, σ_0^2 are the individual-level treatment and control group outcome variances, (assumed for this part of the discussion to be constant across sites), n_j is the sample size for site j, and $\bar{T}_j$ is the fraction of site j's sample assigned to treatment.

Suppose one falsely assumed that $\sigma_1^2 = \sigma_0^2$ and thus used the following pooled estimate for a single individual-level outcome variance, σ^2.

$$\hat{s}^2 = \sum_{j=1}^{J} \sum_{i=1}^{n_j} \hat{e}_{ij}^2 / (N - 2J) \qquad \text{A.3}$$

where $\hat{e}_{ij} = Y_{ij} - \hat{\mu}_0 - \hat{B}_j T_{ij}$ and $N = \sum_{j=1}^{J} n_j$. The expected value of this pooled estimator is:

$$\begin{aligned} E(\hat{\sigma}^2) &= \frac{(N\bar{T}-J)\sigma_1^2 + [N(1-\bar{T})-J]\sigma_0^2}{N-2J} \\ &= \frac{(\bar{n}\bar{T}-1)\sigma_1^2 + [\bar{n}(1-\bar{T})-1]\sigma_0^2}{\bar{n}-2} \\ &\approx \bar{T}\sigma_1^2 + (1-\bar{T})\sigma_0^2 \end{aligned} \qquad \text{A.4}$$

where the individual-level outcome variance is specified to be the same for treatment and control group members and is the following function, $\sigma^2 = \bar{T}_j\sigma_1^2 + (1-\bar{T}_j)\sigma_0^2$ for a given site j. In this case, we are falsely assuming that V_j for site j equals a homoskedastic "apparent" variance V_j^A for that site, where

$$V_j^A = \frac{\sigma^2}{n_j \bar{T}_j (1-T_j)} \qquad \text{A.5}$$

The bias from using V_j^A to estimate V_j is:

$$Bias\left(V_j^A\right) = V_j^A - V_j = \frac{2\left(\sigma_1^2 - \sigma_0^2\right)(\bar{T}_j - 1/2)}{n_j \bar{T}_j (1-\bar{T}_j)} \qquad \text{A.6}$$

A proof of this result is as follows:

$$\begin{aligned}
V_j^A - V_j &= \frac{s^2}{n_j \bar{T}_j (1 - \bar{T}_j)} - \left(\frac{s_1^2}{n_j \bar{T}_j} + \frac{s_0^2}{n_j (1 - \bar{T}_j)}\right) \\
&= \frac{\bar{T}_j s_1^2 + (1 - \bar{T}_j) s_0^2}{n_j \bar{T}_{\cdot j} (1 - \bar{T}_j)} - \left(\frac{s_1^2}{n_j \bar{T}_{\cdot j}} + \frac{s_0^2}{n_j (1 - \bar{T}_j)}\right) \\
&= \frac{s_1^2}{n_j (1 - \bar{T}_j)} - \frac{s_1^2}{n_j \bar{T}_j} + \frac{s_0^2}{n_j \bar{T}_j} - \frac{s_0^2}{n_j (1 - \bar{T}_j)} \\
&= \frac{s_1^2}{n_j \bar{T}_j (1 - \bar{T}_j)} \left(2\bar{T}_j - 1\right) + \frac{s_0^2}{n_j \bar{T}_j (1 - \bar{T}_j)} (1 - 2\bar{T}_j) \\
&= \frac{2(s_1^2 - s_0^2)\left(\bar{T}_j - \frac{1}{2}\right)}{n_j \bar{T}_j (1 - \bar{T}_j)}.
\end{aligned} \qquad \text{A.7}$$

Let us apply this result to the problem of estimating t_B^2 using Equation A.1. The efficacy of this estimator is based on the fact that $E(\hat{B}_j - b)^2 = t_B^2 + V_j$. Hence, $E(\hat{t}_B^2) = t_B^2$. If, however, we use the "apparent" sampling variance (V_j^A) instead of the true sampling variance V_j, our estimator ($\tau_B^{2(A)}$) will be

$$\hat{\tau}_B^{2(A)} = \frac{\sum_{j=1}^J w_j [(\hat{B}_j - \beta)^2 - V_j^A]}{\sum_{j=1}^J w_j} \qquad \text{A.8}$$

which has expectation

$$\begin{aligned}
E\left(\tau_B^{2(A)}\right) &= \frac{\sum_{j=1}^J w_j E[(\hat{B}_j - \beta)^2 - V_j^A]}{\sum_{j=1}^J w_j} \\
&= \frac{\sum_{j=1}^J w_j [(\tau_B^2 + V_j) - (V_j + Bias\left(V_j^A\right)]}{\sum_{j=1}^J w_j} \\
&= \tau_B^2 - \frac{\sum_{j=1}^J w_j Bias(V_j^A)}{\sum_{j=1}^J w_j}
\end{aligned} \qquad \text{A.9}$$

Hence, the bias induced by using V_j^A in place of V_j is the negative of the average bias induced by estimating V_j^A rather than V_j.

<u>Conclusion #2</u>: We do not need to account for cross-site heteroskedasticity unless: (1) this heteroskedasticity is substantial; (2) site sample sizes (n_j) and sample treatment allocations ($\bar{T}_j$) vary substantially; and (3) site sample sizes and treatment allocations covary substantially with level one variances. As described below, there is an empirical test of the likely magnitude of this potential bias.

To understand the basis for this conclusion, consider the basic OLS estimate of the mean program effect for a given site j ($\hat{B}_j^{OLS}$), where:

$$\hat{B}_j^{OLS} = \bar{Y}_{1j} - \bar{Y}_{0j} = \beta + b_j + e_j, \qquad b_j \sim (0, \tau_B^2) \qquad e_j \sim (0, V_j) \tag{A.10}$$

where:

$$V_j = E(\hat{B}_j^{OLS} - B_j)^2 = \frac{\sigma_{1j}^2}{n_{1j}} + \frac{\sigma_{0j}^2}{n_{0j}} \tag{A.11}$$

And we denote the treatment and control group sample sizes for site j as n_{1j} and n_{0j}.

Now suppose that we incorrectly assume cross-site homogeneity of individual-level variances within treatments, or that:

$$\sigma_{1j}^2 = \sigma_1^2 \text{ and } \sigma_{0j}^2 = \sigma_0^2 \qquad \text{for all } j \tag{A.12}$$

when in fact, these variances differ across sites. So how for example, does the incorrectness of our assumption (A.12) affect the following simplified *unweighted* method-of-moments estimator?[1]

$$\hat{\tau}_B^2 = \sum_{j=1}^{J} \frac{(\hat{B}_j^{OLS} - \hat{\beta})^2}{j} - \bar{V} \tag{A.13}$$

where $\bar{V} = \sum_{j=1}^{J} \frac{\hat{V}_j}{J}$

To address this question, first consider the following site-specific estimates of individual-level outcome variances for treatment and control group members:

$$\hat{s}_{1j}^2 = \frac{\sum_{i=1}^{n_j} T_{ij} \hat{e}_{ij}^2}{n_{1j} - 1} \qquad \hat{s}_{0j}^2 = \frac{\sum_{i=1}^{n_j} (1 - T_{ij}) \hat{e}_{ij}^2}{n_{0j} - 1} \tag{A.14}$$

[1] To simplify the discussion here, we are illustrating the basic principles involved for a fully balanced design, where cross-site weighting is not necessary.

where $T_{ij} = 1$ if individual *i* from site *j* was assigned to treatment and 0 otherwise; and e_{ij} is the error term for individual i from site j.

Under the false cross-site homoskedastic assumption (A.12), we will obtain one pooled estimate of the individual-level outcome variance for each experimental group:

$$\hat{s}_1^2 = \frac{\sum_{j=1}^{J}\sum_{i=1}^{n_j} T_{ij}\hat{e}_{ij}^2}{\sum_{j=1}^{J} n_{1j} - J} = \frac{\sum_{j=1}^{J}(n_{1j} - 1)\hat{s}_{1j}^2}{\sum_{j=1}^{J} n_{1j} - J} = \sum_{j=1}^{J} w_{\text{hom}1j}\hat{s}_{1j}^2;$$

$$\hat{s}_0^2 = \frac{\sum_{j=1}^{J}\sum_{i=1}^{n_j} (1 - T_{ij})\hat{e}_{ij}^2}{\sum_{j=1}^{J} n_{0j} - J} = \frac{\sum_{j=1}^{J}(n_{0j} - 1)\hat{s}_{0j}^2}{\left(\sum_{j=1}^{J} n_{0j} - J\right)} = \sum_{j=1}^{J} w_{\text{hom}0j}\hat{s}_{0j}^2 \qquad \text{A.15}$$

where

$$w_{\text{hom}kj} = \frac{n_{kj} - 1}{\sum_{j=1}^{J}(n_{kj} - 1)}$$

for *k*=1 or 0. Using these two pooled variance estimates will give us the following estimate of the average estimation error variance ($\bar{V}_{hom}$), which is needed for the method of moments estimator in Equation A.13.

$$\bar{V}_{\text{hom}} = \frac{1}{J}\sum_{j=1}^{J}\left(\frac{\hat{s}_1^2}{n_{1j}} + \frac{\hat{s}_0^2}{n_{0j}}\right) = \frac{\hat{s}_1^2}{\tilde{n}_1} + \frac{\hat{s}_0^2}{\tilde{n}_0} = \frac{\sum_{j=1}^{J} w_{\text{hom}1j}\hat{s}_{1j}^2}{\tilde{n}_1} + \frac{\sum_{j=1}^{J} w_{\text{hom}0j}\hat{s}_{0j}^2}{\tilde{n}_0}. \qquad \text{A.16}$$

where

$\tilde{n}_1 = J / \sum_{j=1}^{J} n_{1j}^{-1}$, $\tilde{n}_0 = J / \sum_{j=1}^{J} n_{0j}^{-1}$ are the harmonic mean sample sizes for each experimental group.

In contrast, if we make the cross-site heteroskedastic assumption that each site has a unique variance for each experimental group, our estimate of the sampling variance will be

$$\bar{V}_{het} = \frac{1}{J}\sum_{j=1}^{J}\left(\frac{\hat{s}_{1j}^2}{n_{1j}} + \frac{\hat{s}_{0j}^2}{n_{0j}}\right) = \frac{\sum_{j=1}^{J} w_{het1j}\hat{s}_{1j}^2}{\tilde{n}_1} + \frac{\sum_{j=1}^{J} w_{het0j}\hat{s}_{0j}^2}{\tilde{n}_0}. \qquad \text{A.17}$$

where

$$w_{hetkj} = \frac{n_{jk}^{-1}}{\sum_{j=1}^{J} n_{jk}^{-1}} \qquad \text{for } k=0,1.$$

We can now derive the bias of estimation of $\bar{V}$ (and therefore of estimation of t_B^2) when the model is mis-specified so that we are using $\bar{V}_{\text{hom}}$ rather than $\bar{V}_{het}$ in the method of moments estimator. The bias will be:

$$\begin{aligned}
E(\bar{V}_{\text{hom}} - \bar{V}_{het}) &= \frac{\sum_{j=1}^{J} w_{\text{hom}1j}s_{1j}^2 - \sum_{j=1}^{J} w_{het1j}s_{1j}^2}{\tilde{n}_1} + \frac{\sum_{j=1}^{J} w_{\text{hom}0j}s_{1j}^2 - \sum_{j=1}^{J} w_{het0j}s_{0j}^2}{\tilde{n}_0} = \\
&= \frac{\sum_{j=1}^{J} (w_{\text{hom}1j} - w_{het1j})s_{1j}^2}{\tilde{n}_1} + \frac{\sum_{j=1}^{J} (w_{\text{hom}0j} - w_{het0j})s_{0j}^2}{\tilde{n}_0} \\
&= \frac{\bar{s}_1^2 \sum_{j=1}^{J} (w_{\text{hom}1j} - w_{het1j})}{\tilde{n}_1} + \frac{\sum_{j=1}^{J} (w_{\text{hom}j} - w_{het1j})(\bar{s}_{1j}^2 - \bar{s}_1^2)}{\tilde{n}_1} \\
&\quad + \frac{\bar{s}_0^2 \sum_{j=1}^{J} (w_{\text{hom}0j} - w_{het0j})}{\tilde{n}_0} + \frac{\sum_{j=1}^{J} (w_{\text{hom}0j} - w_{het0j})(\bar{s}_{0j}^2 - \bar{s}_0^2)}{\tilde{n}_0} \\
&= \frac{\sum_{j=1}^{J} (w_{\text{hom}1j} - w_{het1j})(s_{1j}^2 - \bar{s}_1^2)}{\tilde{n}_1} + \frac{\sum_{j=1}^{J} (w_{\text{hom}0j} - w_{het0j})(\bar{s}_{0j}^2 - \bar{s}_0^2)}{\tilde{n}_0}
\end{aligned} \qquad \text{A.18}$$

where

$\bar{s}_1^2$ and $\bar{s}_0^2$ = the average variances within the treatment and control groups, respectively.

Equation A.18 demonstrates that if the difference between our homoskedastic and heteroskedastic weights ($w_{homkj} - w_{hetkj}$) is uncorrelated with the level one variance s_{kj}^2 for each experimental group k=0,1, the bias will be null even if the variance for each experimental group differs substantially across sites. Moreover, if the sample sizes within experimental groups are similar across sites, the bias will be null regardless of this correlation.

In other words, we do not need to account for cross-site heteroskedasticity as long as site treatment group sizes are approximately independent of their individual-level

variances and site control group sizes are approximately independent of their individual-level variances, which seems plausible for many situations and can be tested empirically by estimating this correlation and estimating the likely bias using Equation A.18.

For example, in the work-welfare program evaluation that was used as an empirical example for the present paper, the estimated cross-site correlation between the treatment group size (n_{1j}) and outcome variance (σ_{1j}^2) is only + 0.17 and the estimated cross-site correlation between the control group size (n_{0j}) and outcome variance (σ_{0j}^2) is only – 0.12. These estimates are probably fairly precise given the large site samples for the example. Based on Equation A.18 the resulting bias in the estimate of τ_B^2 using our proposed approach (which pools estimates of individual-level outcome variances across sites) is only *seven percent* of its magnitude. Hence, the bias in the cross-site standard deviation of program effects is only *four percent* of its magnitude.

In conclusion, ignoring heteroskedasticity across sites within treatment groups can lead to robust estimates of τ_B^2. Bias in these estimates will be minimal when any of the following conditions are met: (a) level-one variances for each experimental group do not vary appreciably across sites; (b) sample sizes for each experimental group do not vary appreciably across sites; or (c) variances and sample sizes are weakly correlated. However given these conditions the bias increases with the number of sites. Fortunately, one can estimate the resulting bias using Equation A.18.

<u>Conclusion #3</u>: Trying to account for cross-site heteroskedasticity can often make things far worse than they would have been if this issue had been ignored.

This section demonstrates that estimating site-specific individual-level outcome variances to account for cross-site heteroskedasticity can cause one to *overstate* the magnitude of τ_B^2 and overstate the statistical significance of this estimate, potentially by a lot. To see this, recall that for a given site, j, the estimated sampling variance, $\hat{V}_j$, of the OLS-estimated mean program effect, $\hat{B}_j^{OLS}$, is

$$\hat{V}_j = \frac{\hat{\sigma}_{1j}^2}{n_{1j}} + \frac{\hat{\sigma}_{0j}^2}{n_{0j}} \qquad \text{A.19}$$

Thus for site j, the estimator, $\hat{V}_j$, will understate the true sampling variance, V_j, when the estimators $\hat{\sigma}_{1j}^2$ and $\hat{\sigma}_{0j}^2$ understate the true values of σ_{1j}^2 and σ_{0j}^2.

Consider how this affects the m[th]-iteration maximum likelihood estimator $\hat{\tau}_{B(m)}^2$, where:

$$\hat{\tau}_{B(m)}^2 = \sum_{j=1}^{J} \frac{(\hat{\tau}_{B(m-1)}^2 + \hat{V}_j)^{-2}((\hat{B}_j^{OLS} - \hat{\beta})^2 - \hat{V}_j)}{\sum_{j=1}^{J}(\hat{\tau}_{B(m-1)}^2 + \hat{V}_j)^{-2}} \qquad \text{A.20}$$

When $\hat{\sigma}_{1j}^2$ and $\hat{\sigma}_{0j}^2$ cause $\hat{V}_j$ to understate V_j, this has two compounding effects on site *j*'s contribution to $\hat{\tau}_{B(m)}^2$. One effect is a tendency to *overstate* the deviation of site *j*'s program effect from the grand mean program effect. In other words, when $\hat{V}_j$ is too small $(\hat{B}_j^{OLS} - \hat{\beta})^2 - \hat{V}_j$ will tend to be too large. The second effect is to *over-weight* site *j*'s contribution to $\hat{\tau}_{B(m)}^2$. In other words, when $\hat{V}_j$ is too small, the weight for site *j* $(\hat{\tau}_{B(m-1)}^2 + \hat{V}_j)^{-2}$ will be too large. On balance then, site *j*'s contribution to $\hat{\tau}_{B(m)}^2$ will over-weight an over-estimate.

When $\hat{\sigma}_{1j}^2$ and $\hat{\sigma}_{0j}^2$ cause $\hat{V}_j$ to overstate V_j, this also has two compounding effects on site *j*'s contribution to $\hat{\tau}_{B(m)}^2$. One effect is a tendency to *understate* the deviation of site *j*'s program effect from the grand mean program effect. In other words, when $\hat{V}_j$ is too large $(\hat{B}_j^{OLS} - \hat{\beta})^2 - \hat{V}_j$ will tend to be too small. The second effect is to *under-weight* site *j*'s contribution to $\hat{\tau}_{B(m)}^2$. In other words, when $\hat{V}_j$ is too large, the weight for site *j* $(\hat{\tau}_{B(m-1)}^2 + \hat{V}_j)^{-2}$ will be too small. This will cause site *j*'s contribution to $\hat{\tau}_{B(m)}^2$ to under-weight an under-estimate.

Over-weighting over-estimates for some sites and under-weighting under-estimates for other sites will upwardly bias $\hat{\tau}_{B(m)}^2$. This problem is most serious in studies with small site samples because $\hat{V}_j$ is large for these sites and dominates their weights and estimated deviations from the grand mean. In addition, estimation error for $\hat{\sigma}_{1j}^2$ and $\hat{\sigma}_{01}^2$ is greatest for small sites.

Now consider the implications of estimation error in $\hat{V}_j$ for the Q statistic, which is used to test the statistical significance of $\hat{\tau}_B^2$:[2]

$$Q \equiv \sum_{j=1}^{J} \frac{(\hat{B}_j^{OLS} - \hat{\beta})^2}{\hat{V}_j} \qquad \text{A.21}$$

If V_j were known for each site, the Q statistic would be Chi-Square distributed with J-1 degrees of freedom. But in practice, V_j must be estimated with error, which as described below will *upwardly bias* the Q statistic. To see this, consider site *j*'s contribution to Q.

$$Q_j \equiv \frac{(\hat{B}_j^{OLS} - \hat{\beta})^2}{\hat{V}_j} = \left(\frac{1}{\hat{V}_j}\right)(\hat{B}_j^{OLS} - \hat{\beta})^2 \qquad \text{A.19}$$

Because $\hat{V}_j$ is independent of $(\hat{B}_j^{OLS} - \hat{\beta})^2$ [3]

[2] Recall that because the Q statistic is based on the null hypothesis that $\tau_B^2 = 0$, the weight for each site used to compute $\hat{\beta}$ is $\hat{V}_j$.

[3] This derivation assumes an increasingly large number of sites ($J \to \infty$) such that the influence of $\hat{B}_j^{OLS}$ on $\hat{\beta}$ becomes vanishingly small.

$$E(Q_j) = E\left(\frac{1}{\hat{V}_n}\right) \cdot E(\hat{B}_j^{OLS} - \hat{\beta})^2 \qquad \text{A.20}$$

Note that from the properties of harmonic and arithmetic means

$$E\left(\frac{1}{\hat{V}_j}\right) > \frac{1}{E(\hat{V}_j)} = \frac{1}{V_j} \qquad \text{A.21}$$

Thus

$$E(Q_j) = E\left(\frac{1}{\hat{V}_j}\right) \cdot E(\hat{B}_j^{OLS} - \hat{\beta})^2 > (\frac{1}{V_j}) \cdot E(\hat{B}_j^{OLS} - \hat{\beta})^2 \qquad \text{A.22}$$

Therefore the expected value of the contribution of site j *to the Q statistic is larger than it should be in order for the Q statistic to be Chi-Square distributed*.

To assess the likely severity of this problem in a simple case with a fully-balanced multi-site randomized trial (with experimental group sizes that are constant across sites and experimental conditions) we conducted a series of simulations. For a given sample structure (combination of J and n), data were generated as follows: (1) individual outcomes without treatment were generated from a normal distribution with a mean of zero and a standard deviation of one; (2) the program effect was set equal to zero for all sites and sample members; and (3) data for 1,000 simulated samples were generated and analyzed.

Q statistics were then estimated for the 1,000 simulated samples for each sample structure using the following two approaches:

- A Split-sample approach (assuming cross-site heteroskedasticity): By estimating site-specific program effects (B_j) and their estimation error variances (V_j) from *separate* data for each site.

- A Pooled-sample approach (assuming cross-site homoskedasticity): By estimating site-specific program effects (B_j) from a pooled-sample OLS model with fixed site-specific intercepts and fixed site-specific program effects and then estimating the error variance for each site (V_j) as the square of the estimated standard error for its program effect estimate. These estimated standard errors are implicitly based on a within-site estimated outcome variance ($\hat{\sigma}^2$) that is pooled across all sites. To simplify the simulations, reduce their extremely long run times, and be consist with the fact that program effects were uniformly zero, we did not estimate a separate treatment and control group outcome variance.

Table A.1 reports the proportion of Q values that were statistically significant at the 0.05 level from the 1,000 simulated replications for each estimation approach and sample structure.

Table A.1

Percentage of Simulated Q Values that Were Statistically Significant At the 0.05 Level for Split-Sample and Pooled-Sample Estimates Of Individual-Level Outcome Variances

Sample Members per Site (with $n_T = n_C$)	Number of Sites				
	10	20	50	100	200
	Split Sample Approach				
10	0.168	0.229	0.397	0.541	0.760
20	0.092	0.115	0.150	0.233	0.348
50	0.068	0.070	0.080	0.110	0.122
100	0.057	0.056	0.057	0.065	0.083
200	0.048	0.052	0.070	0.059	0.062
500	0.045	0.044	0.050	0.051	0.049
	Pooled Sample Approach				
10	0.065	0.056	0.065	0.074	0.062
20	0.060	0.057	0.067	0.059	0.055
50	0.053	0.046	0.056	0.060	0.048
100	0.052	0.049	0.050	0.064	0.049
200	0.051	0.047	0.046	0.066	0.047
500	0.052	0.064	0.051	0.049	0.048

* 95 % confidence interval for estimates are $\pm$ 0.014 for true p = 0.05 to $\pm$ 0.019 for true p = 0.10 and are somewhat larger for larger values of true p.

As can be seen, with site samples of 50 or fewer persons, the proportion of Q values that appear to be statistically significant at the 0.05 level for the split sample approach is much greater than 0.05 and this problem worsens as the number of sites increases. In contrast, there is little or no such problem for the pooled-sample approach. With site samples of 100 or larger this problem does not arise until the number of sites becomes extremely large. Nonetheless, to be on the safe side it is probably best to use the pooled sample approach for any multi-site trial with a *harmonic mean* site sample size less than 100.

Appendix B

A Constrained Empirical Bayes Method for Estimating Site-Specific Mean Program Effects in Order to Represent a Cross-Site Distribution of Program Effects

This appendix derives an adjustment that corrects for the fact that the sample variance of empirical Bayes estimates of site mean program effects tends to *understate* the true cross-site variance of these effects. To begin, note that the sample variance of empirical Bayes estimates ($\hat{B}_j^{EB}$) around their estimated grand mean ($\hat{\beta}$) for J blocks is:

$$\hat{Var}(\hat{B}_j^{EB}) \equiv \frac{\sum_{j=1}^{J}(\hat{B}_j^{EB}-\hat{\beta})^2}{J} \qquad \text{B.1}$$

Then recall that estimating Equations 11 – 13 in the present paper produces an unbiased estimator of the cross-site variance of true mean program effects $\hat{\tau}_B^2$. The problem is that

$$\hat{Var}(\hat{B}_j^{EB}) < \hat{\tau}_B^2 \qquad \text{B.2}$$

or stated another way:

$$\hat{Var}(\hat{B}_j^{EB}) = \gamma \cdot \hat{\tau}_B^2 \qquad \text{B.3}$$

where

$$\gamma \equiv \frac{\hat{Var}(\hat{B}_j^{EB})}{\hat{\tau}_B^2} \qquad \text{B.4}$$

and

$$0 < \gamma < 1.$$

This implies that:

$$\hat{\tau}_B^2 = \frac{1}{\gamma}\hat{Var}(\hat{B}_j^{EB}) \qquad \text{B.5}$$

Define a constrained empirical Bayes estimator ($\hat{B}_j^{CEB}$) with a sample variance:

$$\hat{Var}(\hat{B}_j^{CEB}) \equiv \frac{\sum_{j=1}^{J}(\hat{B}_j^{CEB}-\hat{\beta})^2}{J} \qquad \text{B.6}$$

Then specify that this sample variance should equal the model-based estimate of the true variance:

$$\hat{Var}(\hat{B}_j^{CEB}) = \hat{\tau}_B^2 \qquad \text{B.7}$$

Substituting Equations B.5 and B.1 into Equation B.7 yields;

$$\begin{aligned}\hat{Var}(\hat{B}_j^{CEB}) &= \frac{1}{\gamma}\hat{Var}(\hat{B}_j^{EB}) \\ &= \frac{1}{J}\cdot\frac{1}{\gamma}\sum_j(\hat{B}_j^{EB}-\hat{\beta})^2 \\ &= \frac{1}{J}\cdot\sum_j(\frac{1}{\sqrt{\gamma}}(\hat{B}_j^{EB}-\hat{\beta}))^2 \qquad \text{B.8}\end{aligned}$$

Equation B.8 indicates that multiplying the deviation of each empirical Bayes estimate from its grand mean by $\frac{1}{\sqrt{\gamma}}$ (which "stretches" these deviations) produces a sample variance that equals the estimated variance of true program effects. The resulting constrained empirical Bayes estimator for a given site *j* is

$$\hat{B}_j^{CEB} = \hat{\beta} + \frac{1}{\sqrt{\gamma}}(\hat{B}_j^{EB} - \hat{\beta}) \qquad \text{B.9}$$

This approach is similar to that suggested by Louis (1984).

Appendix C

The Statistical Power of Omnibus versus Focused Tests Of Cross-site Variation in Program Effects

This appendix demonstrates that an "omnibus" test of the statistical significance of estimated cross-site variation in mean program assignment effects can have less power than a "focused" test of the difference between mean program effects for two subgroups of these sites. Illustrative numerical findings presented below provide a reasonable guide for expectations about the relative statistical power of focused versus omnibus tests under plausible conditions in practice.

Setup

Consider a study that has *J* sites with n sample members each, half of whom are randomized to a treatment group and half of whom are randomized to a control group ($\bar{T}_j = \bar{T} = 0.5$). Assume for simplicity that the program effect for each site (B_j) is constant across all individuals at that site ($B_{ij} = B_j$ for all i) and the variance of individual potential outcomes (σ^2) is constant across sites. Thus $\sigma^2_{1j} = \sigma^2_{0j} = \sigma^2$ for all j. Lastly, designate the cross-site variance of program effects as τ^2_B.

Our *focused* hypothesis test is about the existence of a difference between mean treatment effects for two types of sites (e.g. urban versus non-urban site). Our *omnibus* hypothesis test is about the existence of cross-site variation in program effects. The following individual-level estimation model applies to both tests.

Individual-level model for both tests

$$Y_{ij} = A_j + B_j T_{ij} + e_{ij} \qquad \text{C.1}$$

The following site-level estimation models apply to each test.

Site-level model for omnibus test

$$B_j = \beta + b_j \qquad \text{C.2}$$

Site-level model for focused test

$$B_j = \beta + \varphi \cdot W_j + r_j \qquad \text{C.3}$$

where:

Y_{ij} = the outcome for sample member *i* from site *j*,
T_{ij} = one if sample member *i* from site *j* was assigned to treatment and zero otherwise,
A_j = the mean effect of assignment to the control group for all population

members from site j,

B_j = the effect of assignment to treatment for all population members from site j,

e_{ij} = a random error that varies across all population members with a mean of zero and a variance of σ^2,

β = the grand mean population effect of assignment to treatment,

W_j = one for type A sites and zero for type B sites,

φ = the difference between mean treatment effects for type A and B sites,

b_j = a random error that varies independently and identically across sites with a mean of zero and a variance of τ_b^2,

r_j = a random error that varies independently and identically across sites with a mean of zero, a variance of τ_r^2.

The unconditional cross-site variance of mean program effects τ_B^2 is equal to τ_b^2 and we designate the proportion of this variance that is "explained" by the binary site characteristic (W_j) as R_Δ^2. Thus:

$$\tau_r^2 = (1 - R_\Delta^2)\tau_B^2 \qquad \text{C.4}$$

Our null hypotheses of interest are:

- Omnibus Hypothesis: There is no *variation* in mean treatment effects across sites ($\tau_B^2 = 0$).

- Focused Hypothesis: There is no *difference* between mean treatment effects for the two subgroups of sites ($R_\Delta^2 = 0$).

We proceed by: (1) deriving an expression for each test's F statistic; (2) deriving an expression for the relative magnitudes of the expected values of these F statistics; and (3) comparing the relative magnitudes of the expected and critical F values.

The F Statistics

An F statistic can be used to test for a statistically significant difference between two independent estimates of a common variance. One estimate becomes the numerator of the F statistic and the other becomes its denominator. An F test for a *difference* between two variance estimates is based on whether their *ratio* differs significantly from one.

The Omnibus F Statistic

An F statistic for the omnibus test (F_O) can be defined as the ratio of two independent estimates of the variance of individual outcomes (σ^2).[4] One estimate ($\hat{\sigma}_{BS}^2$) is based on differences *between sites* in their estimated mean program effects. The other

[4] Recall that the present simplified example assumes that $\sigma_{1j}^2 = \sigma_{0j}^2 = \sigma^2$ for all j.

estimate ($\hat{\sigma}^2_{WS}$) is based on variation of individual outcomes *within sites* and experimental groups. Hence:

$$F_O = \frac{\hat{\sigma}^2_{BS}}{\hat{\sigma}^2_{WS}} \qquad \text{C.5}$$

where the degrees of freedom for the numerator and denominator of F_0 equal the degrees of freedom for their respective variance estimators.

The within-site variance estimator is obtained from the following sum of squared residual outcomes

$$\hat{\sigma}^2_{WS} = \sum_{j=1}^{J} \frac{\sum_{i=1}^{n/2}(Y_{1ij}-\bar{Y}_{1j})^2+\sum_{i=\frac{n}{2}+1}^{n}(Y_{0ij}-\bar{Y}_{0j})^2}{(n-2)J} \qquad \text{C.6}$$

where:

Y_{1ij} = the outcome for treatment-group member *i* from site *j*,
Y_{0ij} = the outcome for control-group member *i* from site *j*,
$\bar{Y}_{1j}$ = the mean treatment-group outcome for site *j*,
$\bar{Y}_{0j}$ = the mean control-group outcome for site *j*,

To obtain an expression for $\hat{\sigma}^2_{BS}$ note first that the cross-site variance of OLS-estimated mean program effects, $Var(\hat{B}_j^{OLS})$ equals

$$Var\left(\hat{B}_j^{OLS}\right) = \tau_B^2 + \bar{V}_j \qquad \text{C.7}$$

where

τ_B^2 = the cross-site variance of true mean program effects,

$\bar{V}_j$ = the mean value of the *error variances* for site-specific OLS program effect estimates.

Because for the present discussion each site has the same sample size (n), the same proportion of sample members assigned to treatment ($\bar{T} = 0.5$) and the same individual outcome variance (σ^2), V_j is the same for all sites. Thus $\bar{V}_j = V_j$, where:

$$V_j = \frac{\sigma^2}{n_{1j}} + \frac{\sigma^2}{n_{0j}} = \frac{\sigma^2}{\bar{T}\cdot n} + \frac{\sigma^2}{(1-\bar{T})n} = \frac{\sigma^2}{\bar{T}(1-\bar{T})n}$$

$$= \frac{\sigma^2}{0.5(0.5)n} = \frac{4\sigma^2}{n} \qquad \text{C.8}$$

Substituting Equation C.8 into Equation C.7 yields:

$$Var\left(\hat{B}_j^{OLS}\right) = \tau_B^2 + \frac{4\sigma^2}{n} \quad \text{C.9}$$

Re-arranging terms in Equation C.9 yields:

$$\sigma^2 = \left(\frac{n}{4}\right)\left(Var\left(\hat{B}_j^{OLS}\right) - \tau_B^2\right) \quad \text{C.10}$$

Under the omnibus null hypothesis that $\tau_B^2 = 0$, Equation C.10 becomes:

$$\sigma^2 = \frac{n}{4} Var(\hat{B}_j^{OLS}) \quad \text{C.11}$$

From the following estimate of $Var\left(\hat{B}_j^{OLS}\right)$[5]

$$\widehat{Var}\left(\hat{B}_j^{OLS}\right) = \left(\frac{1}{J-1}\right)\sum_{j=1}^{J}(\hat{B}_j^{OLS} - \hat{\beta})^2 \quad \text{C.12}$$

we can obtain the between site estimate ($\hat{\sigma}_{BS}^2$) of σ^2:

$$\hat{\sigma}_{BS}^2 = \frac{n}{4(J-1)}\sum_{j=1}^{J}(\hat{B}_j^{OLS} - \hat{\beta})^2 \quad \text{C.13}$$

Consequently:

$$F_O = \frac{\hat{\sigma}_{BS}^2}{\hat{\sigma}_{WS}^2}$$

$$= \frac{\frac{n}{4(J-1)}\sum_{j=1}^{J}(\hat{B}_j^{OLS} - \hat{\beta})^2}{\sum_{j=1}^{J}\frac{\sum_{i=1}^{n/2}(Y_{1ij} - \bar{Y}_{1j})^2 + \sum_{i=\frac{n}{2}+1}^{n}(Y_{0ij} - \bar{Y}_{0j})^2}{(n-2)J}} \quad \text{C.14}$$

The numerator of this F statistic has *(J*-1*)* degrees of freedom and the denominator has (n-2)*J* degrees of freedom.

The Focused F Statistic

Note that:

$$F_F = \frac{\hat{\sigma}_{BSG}^2}{\hat{\sigma}_{WSG}^2} \quad \text{C.15}$$

[5] Under the omnibus null hypothesis that $\tau_B^2 = 0$, $\hat{\beta}$ is the weighted mean of the $\hat{B}_j^{OLS}$, with site weights equal to their value of $\hat{V}_j$. Because $\hat{V}_j$ is constant across sites in the present example, $\hat{\beta}$ is equivalent to the un-weighted mean of the $\hat{B}_j^{OLS}$.

where $\hat{\sigma}^2_{BSG}$ is the estimated value of σ^2 based on variation in impacts between our two subgroups of sites and $\hat{\sigma}^2_{WSG}$ is the estimated value of σ^2 based on variation in outcomes within our two subgroups of sites.

The within-subgroup variance estimate ($\hat{\sigma}^2_{WSG}$) for the focused test is the same as the within-site variance estimate ($\hat{\sigma}^2_{WS}$) for the omnibus test so that:

$$\hat{\sigma}^2_{WSG} = \sum_{j=1}^{J} \frac{\sum_{i=1}^{n/2}(Y_{1ij}-\bar{Y}_{1j})^2+\sum_{i=\frac{n}{2}+1}^{n}(Y_{0ij}-\bar{Y}_{0j})^2}{(n-2)J} \quad \text{C.16}$$

To obtain an expression for the estimated between subgroup variance ($\hat{\sigma}^2_{BSG}$) note that each subgroup has J/2 sites. Hence, the estimated mean program effect for a specific subgroup of sites ($\hat{B}_{sg}$) is

$$\hat{B}_{sg} = \sum_{j=1}^{J/2} \hat{B}_j^{OLS} / (\frac{J}{2}) \quad \text{C.17}$$

Because the estimation error variance (V_j) for a single site is $\frac{4\sigma^2}{n}$

$$Var(\hat{B}_{sg}) = (\frac{4\sigma^2}{n})/(J/2)$$

$$= \frac{8\sigma^2}{Jn} \quad \text{C.18}$$

Re-arranging terms in Equation C.18 yields:

$$\sigma^2 = \frac{Jn}{8} Var(\hat{B}_{sg}) \quad \text{C.19}$$

An unbiased estimate of $Var(\hat{B}_{sg})$ with *one* degree of freedom is:

$$\widehat{Var}(\hat{B}_{sg}) = \sum_{sg=1}^{2}(\hat{B}_{sg} - \hat{\beta})^2/1$$

$$= \sum_{sg=1}^{2}(\hat{B}_{sg} - \hat{\beta})^2 \quad \text{C.20}$$

Thus an unbiased between-subgroup estimate of σ^2 is:

$$\hat{\sigma}^2_{BSG} = \frac{Jn}{8}\sum_{sg=1}^{2}(\hat{B}_{sg} - \hat{\beta})^2 \quad \text{C.21}$$

Consequently:

$$F_F = \frac{\hat{\sigma}^2_{BSG}}{\hat{\sigma}^2_{WSG}}$$

$$= \frac{\frac{Jn}{8}\sum_{sg=1}^{2}(\hat{B}_{sg}-\hat{\beta})^2}{\sum_{j=1}^{J}\frac{\sum_{i=1}^{n/2}(Y_{1ij}-\bar{Y}_{1j})^2+\sum_{i=\frac{n}{2}+1}^{n}(Y_{0ij}-\bar{Y}_{0j})^2}{(n-2)J}} \qquad \text{C.22}$$

The numerator of this F statistic has one degree of freedom and the denominator has (n-2)J degrees of freedom.

Because F_F and F_O have the same denominators, we can use their numerators to help assess their relative statistical power. With this in mind, dividing Equation C.22 by Equation C.14 yields:

$$\frac{F_F}{F_O} = \frac{\frac{Jn}{8}\sum_{sg=1}^{2}(\hat{B}_{sg}-\hat{\beta})^2}{\frac{n}{4(J-1)}\sum_{j=1}^{J}(\hat{B}_j^{OLS}-\hat{\beta})^2} \qquad \text{C.23}$$

Expected Values of the Numerators When Program Effects Vary Across Sites

The next step is to derive expressions for the expected values of the numerators of the omnibus and focused F statistics when treatment effects vary across blocks ($\tau_B^2 > 0$).

The Omnibus Numerator (NUM_O)

$$E[NUM_O] = E[\frac{n}{4(J-1)}\sum_{j=1}^{J}(\hat{B}_j^{OLS}-\hat{\beta})^2]$$

$$= \frac{n}{4}E[\sum_{j=1}^{J}(\hat{B}_j^{OLS}-\hat{\beta})^2/(J-1)]$$

$$= \frac{n}{4}(\tau_B^2 + \frac{4\sigma^2}{n})$$

$$= \frac{n\cdot\tau_B^2}{4} + \sigma^2 \qquad \text{C.24}$$

The Focused Numerator (NUM_F)

$$E[NUM_F] = E[\frac{Jn}{8}\sum_{sg=1}^{2}(\hat{B}_{sg}-\hat{\beta})^2]$$

$$= \frac{Jn}{8}Var(\hat{B}_{sg}) \qquad \text{C.25}$$

where $VAR(\hat{B}_{sg})$ has three components:

- The true cross-subgroup variance of mean program effects ($R_\Delta^2\tau_B^2$),
- A function of the true cross-site variance of mean program effects that lies within each subgroup $(1-R_\Delta^2)(\tau_B^2)/(J/2)$ and
- A function of the site-level error variance of estimated mean program effects $(V_j/(J/2)$ or $(\frac{4\sigma^2}{n})/(J/2)$.

Substituting these components into Equation C.25 yields:

$$E[NUM_F] = \frac{Jn}{8}[R_\Delta^2\tau_B^2 + \frac{(1-R_\Delta^2)(\tau_B^2)}{\frac{J}{2}} + \left(\frac{\frac{4\sigma^2}{n}}{\frac{J}{2}}\right)]$$

$$= \frac{JnR_\Delta^2\tau_B^2}{8} + \frac{n(1-R_\Delta^2)\tau_B^2}{4} + \sigma^2 \qquad \text{C.26}$$

Thus the ratio of expected values for the two F statistics is:

$$\frac{E[NUM_F]}{E[NUM_O]} = \frac{\frac{JnR_\Delta^2\tau_B^2}{8} + \frac{n(1-R_\Delta^2)\tau_B^2}{4} + \sigma^2}{\frac{n\tau_B^2}{4} + \sigma^2} \qquad \text{C.27}$$

When R_Δ^2 equals zero (subgroup designation explains *none* of the true cross-block variation in program effects), the two numerators are the same and therefore:

$$\frac{E[NUM_F]}{E[NUM_O]} = \frac{\frac{n\cdot\tau_B^2}{4} + \sigma^2}{\frac{n\cdot\tau_B^2}{4} + \sigma^2} = 1 \qquad \text{C.28}$$

When R_Δ^2 equals one (subgroup designation explains *all* of the true cross-block variation in program effects) then $[NUM_F] > E[NUM_O]$ when $(\frac{J}{8} > \frac{1}{4})$ and thus there are two or more sites per subgroup. In that case:

$$\frac{E[NUM_F]}{E[NUM_O]} = \frac{\frac{Jn\tau_B^2}{8} + \sigma^2}{\frac{n\tau_B^2}{4} + \sigma^2} > 1 \qquad \text{C.29}$$

Table C.1 reports values of $\frac{E[NUM_F]}{E[NUM_O]}$ for specified values of J and R_Δ^2 given $\sigma^2 = 1$, $\tau_B = 0.16$ and $n = 100$. Setting $\sigma^2 = 1$ implies that $\tau_B$is in units of standardized mean-difference effect-sizes. Thus τ_B is the standard deviation of true program effect-sizes across blocks. The value of this parameter is assumed to be 0.16, which as noted in the paper, is the minimum detectable effect-size standard deviation (MDESSD) for a fully-balanced multi-site trial with 20 sites and 100 individually-randomized sample members per site (assuming an individual-level covariate R-square of 0.5).

Given the specified parameters, note that whenever $R_\Delta^2 > 0$, the expected value of F_F is greater than that of F_O. However this does not mean that the focused hypothesis test is necessarily more powerful than the omnibus test. In order make this determination we must also take into account the difference in the numbers of degrees of freedom for their denominators. One way to do this is by comparing their critical F values for a given level of statistical significance.

Table C.1

Ratio of Focused to Omnibus Expected F Values

R^2_Δ	Number of Blocks (J)				
	10	20	40	80	160
0	1.00	1.00	1.00	1.00	1.00
0.1	1.32	1.72	2.52	4.12	7.32
0.2	1.64	2.44	4.04	7.24	13.64
0.3	1.96	3.16	5.56	10.36	19.96
0.4	2.28	3.88	7.08	13.48	26.28
0.5	2.60	4.60	8.60	16.60	32.60
0.6	2.92	5.32	10.12	19.72	38.92
0.7	3.24	6.04	11.64	22.84	45.24
0.8	3.56	6.76	13.16	25.96	51.56
0.9	3.88	7.48	14.68	29.08	57.88
1	4.20	8.20	16.20	32.20	64.20

NOTE: Values in the table assume that $\sigma^2 = 1, n = 100$ and $\tau_B = 0.16$.

Comparing Critical F values

To interpret the findings in Table C.1 they must be compared to the ratio of critical F values for the two tests. This ratio reflects the fact that the focused test has only one degree of freedom in its numerator while the omnibus test has *J*-1 degrees of freedom. The focused test thus has larger critical F values because of its greater uncertainty. Table C.2 reports the ratio of these critical F values assuming a very large number of denominator degrees of freedom (20,000).[6]

Table C.2

Ratio of Focused to Omnibus Critical F Values

Number of Blocks (J)				
10	20	40	80	160
2.04	2.45	2.75	3.02	3.22

NOTE: Critical values are for the 0.05 level of statistical significance with J-1 degrees of freedom in the numerator for the omnibus test, one degree of freedom in the numerator for the focused test and a very large number of degrees of freedom (20,000) in the denominators for both tests.

Putting the Pieces Together

To assess the relative statistical power of the focused versus omnibus hypothesis tests, we must compare results from Table C.1 about the relative magnitudes of their expected F values to results from Table C.2 about the relative magnitudes of their critical

[6] These findings are good approximations for F statistics with more than about 100 denominator degrees of freedom.

F values. When the former is larger than the latter, the focused test will tend to have more power. When the former is smaller than the latter, the omnibus test will have more power.

Table C.3 puts these pieces together by reporting the *ratio of the ratio* in Table C.1 to its counterpart in Table C.2. Table C.3 shades cases where the expected F value advantage of the focused test outweighs the critical F value advantage of the omnibus test. In the present hypothetical example this occurs in the overwhelming majority of the cases. Further examination of Equation C.27 indicates that the advantage of the focused test will increase as the cross-site standard deviation of program effect sizes (τ_B) increases and will decrease as the site sample size (n) decreases. Nonetheless, a cross-site effect size standard deviation of 0.16 is consistent with the limited empirical evidence that currently exists (a value of 0.08 for the welfare-to-work example in the present paper and values ranging from 0.07 to 0.25 for Head Start effects reported by Bloom and Weiland, 2014). In addition, a study with 20 or more sites having 100 sample members per site for a total of 2,000 or more sample members is comparable to many past multi-site trials and within the range of feasibility for future trials. Thus findings in Table C.3 may provide a reasonable guide for expectations about the relative statistical power of focused versus omnibus tests under plausible conditions.

However as a guide to practice, the mere fact that it is *possible* for a focused test to detect systematic cross-site variation in program effects that is missed by an omnibus test is enough to support our recommendation that an omnibus test not be used as a "gateway" criterion for deciding whether to test a theory-based a priori, hypothesis about a site-level moderator.

Table C.3

Ratio of the Focused to Omnibus Ratio of Expected F values
To the Ratio of the Focused to Omnibus Critical F Values

R^2_Δ	Number of Sites (J)				
	10	20	40	80	160
0	0.49	0.41	0.36	0.33	0.31
0.1	0.65	0.70	0.91	1.37	2.27
0.2	0.80	1.00	1.47	2.40	4.23
0.3	0.96	1.29	2.02	3.44	6.19
0.4	1.12	1.59	2.57	4.47	8.15
0.5	1.27	1.88	3.12	5.51	10.11
0.6	1.43	2.18	3.67	6.54	12.07
0.7	1.59	2.47	4.23	7.58	14.03
0.8	1.74	2.76	4.78	8.61	15.99
0.9	1.90	3.06	5.33	9.64	17.95
1	2.06	3.35	5.88	10.68	19.91

NOTE: Values in the table assume that $\sigma^2 = 1, \tau_B = 0.16$ and $n = 100$