Routledge
Taylor & Francis Group

Journal of Research on Educational Effectiveness

ISSN: 1934-5747 (Print) 1934-5739 (Online) Journal homepage: www.tandfonline.com/journals/uree20

Using Multisite Experiments to Study Cross-Site Variation in Treatment Effects: A Hybrid Approach With Fixed Intercepts and a Random Treatment Coefficient

Howard S. Bloom, Stephen W. Raudenbush, Michael J. Weiss & Kristin Porter

To cite this article: Howard S. Bloom, Stephen W. Raudenbush, Michael J. Weiss & Kristin Porter (2017) Using Multisite Experiments to Study Cross-Site Variation in Treatment Effects: A Hybrid Approach With Fixed Intercepts and a Random Treatment Coefficient, Journal of Research on Educational Effectiveness, 10:4, 817-842, DOI: 10.1080/19345747.2016.1264518

To link to this article: https://doi.org/10.1080/19345747.2016.1264518

JOURNAL OF RESEARCH ON EDUCATIONAL EFFECTIVENESS
2017, VOL. 10, NO. 4, 817–842
https://doi.org/10.1080/19345747.2016.1264518

METHODOLOGICAL STUDIES

Using Multisite Experiments to Study Cross-Site Variation in Treatment Effects: A Hybrid Approach With Fixed Intercepts and a Random Treatment Coefficient

Howard S. Bloom[a], Stephen W. Raudenbush[b], Michael J. Weiss[a], and Kristin Porter[c]

ABSTRACT

The present article considers a fundamental question in evaluation research: "By how much do program effects vary across sites?" The article first presents a theoretical model of cross-site impact variation and a related estimation model with a random treatment coefficient and fixed site-specific intercepts. This approach eliminates several biases that can arise from unbalanced sample designs for multisite randomized trials. The article then describes how the approach operates, explores its assumptions, and applies the approach to data from three large welfare-to-work trials. The article also illustrates how to report cross-site impact findings and presents diagnostics for assessing these findings. To keep the article manageable, it focuses on experimental estimates of effects of program assignment (effects of intent to treat), although the ideas presented can be extended to analyses of multisite quasi-experiments and experimental estimates of effects of program participation (complier average causal effects).

KEYWORDS

multisite trials
impact variation

Cross-site variation in program effects has important consequences for policy, practice, and research. For example, if this variation is large and unexplained, knowing the average impact of a program will tell us little about how well it worked in particular settings. Moreover, with substantial cross-site impact variation, an observed average program effect (or impact) may not extrapolate well to settings beyond those studied (Tipton, 2014). However, if program effects vary little across sites in a study, an average impact estimate will provide a reasonable approximation for each of those sites and may provide a reasonable approximation for similar sites that were not in the study.[1]

Cross-site variation in program effects also can have implications for targeting services to identifiable subpopulations of sites. In addition, the pattern of cross-site variation in program effects can have implications for equity or fairness. For example, a reading curriculum that is

CONTACT Howard S. Bloom howard.bloom@mdrc.org MDRC, 16 East 34th Street, 19th Floor, New York, NY 10016, USA.
[a]MDRC, New York, New York, USA
[b]University of Chicago, Chicago, Illinois, USA
[c]MDRC, Oakland, California, USA

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/uree.

Supplemental data for this article can be accessed at https://doi.org/10.1080/19345747.2016.1264518.

[1]Strictly speaking, this conclusion is only valid for studies based on a random sample of sites. However, the conclusion seems plausible for studies based on a diverse nonrandom sample of sites.

most effective for sites with large concentrations of struggling readers might help to reduce cross-site disparity in reading achievement, whereas a reading curriculum that is most effective for sites with large concentrations of high-achieving students might increase this 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 were—as evidence of what can be achieved under the right conditions. It could also be useful to know the proportion of charter schools that had negative impacts (i.e., were less effective than their local alternatives). Nonetheless, to date, researchers have focused mainly on average program effects.

To help fill this void, we propose an approach that uses multisite randomized trials to detect and quantify cross-site impact variation. Only a few studies to date have used multisite trials for this purpose (e.g., Angrist, Pathak, & Walters, 2013; Bloom, Hill, & Riccio, 2003; Bloom & Weiland, 2015; Konstantopoulos, 2011; Lake et al., 2012; May et al., 2013; Raudenbush, Reardon, & Nomi, 2012; Walters, 2015). However, with the current large and growing number of multisite trials, our ability to conduct such analyses is increasing. For example, Weiss et al. (2017) recently reanalyzed data from 16 multisite trials in education and job training using the methods we present. Furthermore, there is likely to be an increasing number of multisite evaluations funded by federal initiatives to scale-up evidence-based programs.[2]

To help researchers capitalize on this new body of evidence, we introduce a promising approach for studying a *cross-site distribution* of site mean program effects. To stay within the scope of a single article we focus on effects of program assignment (intent to treat or ITT) and leave consideration of the effects of program participation (complier average causal effects [CACE] *aka* local average treatment effects [LATE]) for future research. Our approach may be regarded as a hybrid that builds on two well-known methods: random-effects meta-analysis and two-level hierarchical linear models with randomly varying intercepts and slopes.

A multisite trial can be regarded as a "planned meta-analysis," that is, a fleet of experiments, each conducted in a single site. Given the similarity of the multisite trial to meta-analysis, one might suggest using now-standard meta-analytic procedures to study cross-site impact variation. The core idea in random effects meta-analysis (cf. Hedges & Olkin, 2014; Raudenbush & Bryk, 1985; Rubin, 1981) is to use an impact estimate and its estimated standard error from each study to estimate the mean and standard deviation of impacts across studies. Although this approach can be applied to multisite trials, we depart from it in order to correctly estimate the standard errors of site-specific impact estimates. If those standard errors are misspecified, for example, if we assume the same individual-level outcome variance for our treatment and control groups when in fact those variances differ, we can obtain biased estimates of cross-site impact variation. Or, as explained later, if we estimate a separate individual-level outcome variance for each site in our study (which is analogous to the meta-analytic approach for individual studies) and samples for those sites are small, we could greatly overestimate cross-site impact variation. In contrast, meta-analysts typically must rely on published summary statistics for each study to obtain their estimated standard

[2]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.

errors, which limits one's ability to account for the structure of error variation within each study (or, in our case, site). Fortunately, multisite trials can provide access to the raw data for each site, so one can be more flexible in this regard.

A popular alternative to the meta-analytic approach is a two-level hierarchical linear model with random intercepts and a randomly varying treatment coefficient (cf. Raudenbush & Bryk, 2002, chapter 4). Indeed, our theoretical model is of this type. Using this approach, we can estimate heterogeneous within-site variance structures. In this way we can obtain unbiased estimates of the average treatment effect from a multisite trial under fairly weak assumptions. However, obtaining good estimates of impact variation using this approach requires strong assumptions about the randomly varying intercept, which are relaxed by the approach we propose.

To address these limitations, we recommend a hybrid two-level model with fixed site-specific intercepts and a treatment coefficient that varies randomly across sites. Like the meta-analytic method, our proposed approach estimates a single random program effect for each site in a way that avoids some of the strong assumptions associated with the random intercept model. However, unlike the meta-analytic method, our proposed alternative is flexible in estimating the structure of outcome variation within sites. And the required assumptions about this outcome variation are weaker than those for meta-analytic and other conventional multilevel models.

Our article makes the following contributions. First, it presents a simple theoretical model of cross-site variation in ITT effects in order to clarify our parameters of interest, and describes our hybrid statistical model for estimating those parameters, with a brief discussion of its identification. Second, the article provides intuition about the hybrid approach in order to help researchers understand how, why, and when to use it, and provides guidance on how to implement the approach using conventional software. Third, the article considers problems that can arise from two forms of individual-level heteroskedasticity and describes how to assess the likely severity of these problems and address them when they exist. Fourth, the article describes how to graphically represent a cross-site impact distribution using adjusted empirical Bayes estimates that are based on our proposed model. Fifth, the article presents several graphical diagnostics to help researchers examine the uncertainty about their estimates of cross-site impact variation and site-specific impacts. Sixth, the article illustrates how to interpret findings from our proposed approach by applying it to a major empirical example, clarifying the key assumptions involved in doing so, and testing the sensitivity of our results to those assumptions.

Before proceeding, note that the present article is the first in a series of three related articles being published together in the present issue of this journal. The second article, "Assessing the Precision of Multisite Trials for Estimating the Parameters of a Cross-Site Population Distribution of Program Effects," by Howard S. Bloom and Jessaca Spybrook (2017), builds on the present analytic framework to describe how to design samples for multisite trials that provide the desired level of statistical precision for estimating three key parameters of a cross-site distribution of program effects: a cross-site mean effect, a cross-site standard deviation of effects, and a difference in cross-site mean effects for two sub-populations of sites. The third article in the series, "How Much Do the Effects of Education and Training Programs Vary Across Sites? Evidence From Past Multisite Randomized Trials," by Michael J. Weiss, Howard S. Bloom, Natalya Verbitsky Savitz, Himani Gupta, Alma Vigil, and Dan Cullinan (2017), uses the present analytic framework to estimate cross-site means and standard deviations for education and training program effects from data for 16 large-scale multisite randomized trials.

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 cross-site 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 sites from this population.[3]

Theoretical Model

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

$$Y_{ij} = A_j + B_j T_{ij} + e_{ij} \tag{1}$$

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 in sites with a mean of zero and a variance that can differ between treatment and control group members and across sites.[4]

When studying cross-site impact variation, it is useful to consider study sites as a probability sample from a super population of sites, regardless of whether the sites were chosen as a probability sample from a well-defined population or, as is the case for most program evaluations, the study is based on a convenience sample from an implied population that cannot be fully described.[5] 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 usually to generalize findings beyond the sites observed, even though the target of generalization is not well-defined.

A multisite trial has several population parameters of interest. First is the population average program effect ($E(B) = \beta$), which for the present article is defined as the mean of the

[3] 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 analysis requires related estimation methods (see Raudenbush & Bloom, 2015, and Raudenbush & Schwartz, 2017).

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

[5] Two important but rare exceptions are the National Head Start Impact Study (Puma, Bell, Cook, & Heid, 2010), which was based on a national probability sample of local Head Start centers, and the National Job Corps Study (Schochet, Burghardt, & McConnell, 2006), which was based on the experience of almost all eligible youths who applied to the program between November 1994 and December 1996.

site-specific mean program effects for all sites in the theoretical population. For this parameter, each site is considered to be equally important and thus each site's true mean impact has equal weight. A second parameter of interest, which is the central focus of the present article, is the population cross-site variance of program effects ($Var(B) = \tau_B^2$) or its more interpretable counterpart, the population cross-site standard deviation of program effects (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 ($Cov(A, B) = \tau_{AB}$).

Our theoretical model can be written as a two-level hierarchical linear model (HLM) in which level one is represented by Equation 1 and level two is represented by Equations 2 and 3.

$$A_j = \alpha + a_j \tag{2}$$

and

$$B_j = \beta + b_j \tag{3}$$

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

Estimation Model

One condition required for consistent estimates of our theoretical model is that the site-specific fraction of persons assigned to the treatment ($\overline{T}_j$) be uncorrelated with site-level random effects (a_j and b_j).[6] If $\overline{T}_j$ varies across sites, which often occurs in multisite RCTs,[7] and is correlated with unobserved site characteristics (and thus is related to a_j or b_j), standard methods can produce inconsistent parameter estimates.

One way to eliminate this problem is to site-center the variables in level one of our theoretical model (Equation 1). Because the mean value of such site-centered variables is zero for all sites, they cannot be correlated with characteristics of sites. When estimating a site-centered hierarchical model one must account for the loss of a degree of freedom per site produced by site-centering the dependent variable. If site-centering is a regular 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).

[6] Standard identification assumptions for Equations 1–3 are that: (i) T must be independent of the individual-level random error (e); (ii) T must be independent of the site-level effects a_j and b_j, (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; (v) the analyst must correctly specify the covariance structure of a_j and b_j, and (vi) site sample size (n_j) is not correlated with mean site impact (B_j). 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 (vi) is not problematic in the empirical example that we consider.

[7] All of the 16 multisite education or job-training RCTs reanalyzed by Weiss et al. 2017 had values for $\overline{T}_j$ that varied across sites.

Individual Level

$$Y_{ij} = \alpha_j + B_j T_{ij} + e_{ij} \tag{4}$$

Site Level

$$\alpha_j = \alpha_j \tag{5}$$

$$B_j = \beta + b_j \tag{6}$$

This model is the basis for the discussion that follows. In practice we can add individual-level baseline covariates to Equation 4 in order to increase the precision of its estimated parameters.[8] Doing so does not change the basic properties of the model. Note that in order to identify our parameters of interest for the present analysis, $E(B)$ and $Var(B)$, with a minimum of assumptions, we forego the ability to estimate other parameters of a cross-site impact distribution, namely, $E(A)$, $Var(A)$, and $Cov(A, B)$.

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 cross-site impact variation.

Within-Site Model

With random assignment of sample members to a treatment or control group at each site, we 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 outcomes for its treatment group ($\overline{Y}_{1j}$) and control group ($\overline{Y}_{0j}$), such that:

$$\hat{B}_j^{OLS} = \overline{Y}_{1j} - \overline{Y}_{0j} \tag{7}$$

Thus as noted earlier, we can regard a multisite 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 the site's random estimation error (r_j). Thus:

$$\hat{B}_j^{OLS} = B_j + r_j \tag{8}$$

Note that $E(r) = 0$ 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).

[8] Specifying fixed site-specific intercepts in the model is equivalent to site-centering the values of all variables in the model (see Greene, 2008).

For a treatment and control group difference of mean outcomes:

$$V_j = \frac{\sigma_{1j}^2}{n_{1j}} + \frac{\sigma_{0j}^2}{n_{0j}} = \frac{\sigma_{1j}^2}{n_j \overline{T}_j} + \frac{\sigma_{0j}^2}{n_j\left(1-\overline{T}_j\right)} \tag{9}$$

where

$\sigma_{1j}^2 =$ the within-site variance of outcomes for treatment group members from site j,
$\sigma_{0j}^2 =$ the within-site variance of outcomes for control group members from site j,
$n_j =$ the number of sample members from site j,
$n_{1j} = n_j \overline{T}_j =$ the number of treatment group members from site j,
$n_{0j} = n_j(1-\overline{T}_j) =$ 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 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 ($\sigma_{1j}^2 \neq \sigma_{0j}^2$, see Raudenbush & Bloom, 2015). We refer to this condition as "*T/C* heteroskedasticity" and demonstrate later that if the condition is not accounted for, it can bias an estimate of τ_B^2.

But what if the variance of individual outcomes also differs across sites? As discussed later, it is only necessary to account for such "cross-site heteroskedasticity" under certain conditions, and we present an approach for assessing those conditions. But if one tried to account for cross-site heteroskedasticity by estimating separate individual-level outcome variances for each site, this could have the perverse effect of overstating the magnitude and statistical significance of τ_B^2. This bias, which can be substantial, is most extreme for studies with many small sites (see online Appendix A).[9]

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,

$$V_j = \frac{\sigma_1^2}{n_j \overline{T}_j} + \frac{\sigma_0^2}{n_j\left(1-\overline{T}_j\right)} \tag{10}$$

This option is currently available in software for estimating multilevel models such as HLM, SAS PROC MIXED, MLwiN, and R; and online Appendix B illustrates how to use SAS to estimate our model with the option.

Combined Between-Site and Within-Site Model

Recall that true mean program effects vary across sites as:

$$B_j = \beta + b_j \tag{11}$$

where

[9] Online Appendix A demonstrates that data for small site samples tend to understate individual-level residual outcome variances (σ_1^2 and σ_0^2), which causes one to understate the error variance (V_j) of site-specific impact estimates, which in turn, causes one to overstate true cross-site impact variation (τ_b^2), especially for studies with many small sites.

$E(b) = 0$
$Var(b) = \tau_b^2 = \tau_B^2$.
Combining Equations 11 and 8 yields:

$$\hat{B}_j^{OLS} = \beta + b_j + r_j \tag{12}$$

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

$$Var\left(\hat{B}_j^{OLS}\right) = Var(\beta) + Var(b) + Var(r) = 0 + \tau_B^2 + V_j = \tau_B^2 + V_j \tag{13}$$

Equation 13 indicates that an OLS mean program effect estimate ($\hat{B}_j^{OLS}$) for a given sample site reflects two sources of variation: (a) the cross-site variance of true mean program effects $\left(\tau_B^2\right)$ and (b) 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 & Pigott, 2001).

Estimating the Mean of a Population Distribution of Site-Mean Impacts

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 (β), and this estimator has a total variance of $(\tau_B^2 + V_j)$. If V_j and τ_B^2 were known and if V_j were independent of B_j, 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]

$$\hat{\beta} = \sum_{j=1}^{J} \frac{(\tau_B^2 + V_j)^{-1} \hat{B}_j^{OLS}}{\sum_{j=1}^{J} (\tau_B^2 + V_j)^{-1}} \tag{14}$$

Equation 14 down-weights program-effect estimates from sites with large estimation error variances (V_j) produced by small samples and/or values of $\overline{T}_j$ that are far from 0.5. Equation 14 up-weights program-effect estimates from sites with small estimation error variances (V_j) produced by large samples and/or values of $\overline{T}_j$ that are near 0.5. This 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 the estimator represented by Equation 14 requires knowledge of τ_B^2 and V_j, we must use consistent estimators of them to implement Equation 14. An important assumption then is that our precision weights are uncorrelated with site-specific treatment effects. Later we illustrate a strategy for assessing the likely bias produced by violations of this assumption.

[10] If b and r are normally distributed, then Equation 14 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 14 is the best linear unbiased estimator under the Gauss-Markov theorem. These conditions hold if there is no correlation between site weights and site mean impacts.

Statistical Testing for Cross-Site Impact Variation

Under the null hypothesis that $\tau_B^2 = 0$, the site weight in Equation 14 simplifies to $\frac{V_j^{-1}}{\sum_{j=1}^{J} V_j^{-1}}$ and we can estimate β under the null hypothesis with the following fixed-effect estimator:

$$\hat{\beta}^{FIXED} = \sum_{j=1}^{J} \frac{V_j^{-1} \hat{B}_j^{OLS}}{\sum_{j=1}^{J} V_j^{-1}} \tag{15}$$

We can then compute a Q statistic, which is widely used in meta-analysis to test the null hypothesis of zero cross-study impact variation (Hedges & Olkin, 2014), where:

$$Q = \sum_{j=1}^{J} \frac{\left(\hat{B}_j^{OLS} - \hat{\beta}^{FIXED}\right)^2}{V_j} \tag{16}$$

Under the null hypothesis, if our estimate $(\frac{\hat{B}_j^{OLS} - \hat{\beta}^{FIXED}}{\sqrt{\hat{V}_j}})$ approximates a normal variate,[11] then 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 $(\hat{V}_j)$. To assess how well values for $(\frac{\hat{B}_j^{OLS} - \hat{\beta}^{FIXED}}{\sqrt{\hat{V}_j}})$ approximate a normal distribution, we can use a standard Q-Q plot (see http://support.sas.com/kb/47/246.html and Blom, 1958), as illustrated later for our empirical example.[12]

Estimating a Cross-Site Impact Variance

To use Equation 14 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:

$$Var\left(\hat{B}_j^{OLS}\right) \equiv E\left[\left(\hat{B}_j^{OLS} - \beta\right)^2\right] = \tau_B^2 + V_j \tag{17}$$

[11] Under the null hypothesis, dividing $(\hat{B}_j^{OLS})$ by $\sqrt{\hat{V}_j}$ ensures a constant cross-site variance.

[12] Although some scholars have expressed concern about using a chi-square distribution to make statistical inferences about population variances (e.g., Pearson, 1931 and Scheffe, 1959) because of potential bias from non-normally distributed sample points, this approach is widely used in meta-analysis to assess the statistical significance of cross-study impact variation (e.g., Hedges & Olkin, 2014). Furthermore, our experience with data from numerous multisite trials has been that inferences about the statistical significance of estimates of cross-site impact variation $(\hat{\tau}_B^2)$ based on a Q statistic are consistent with the magnitudes of estimates of τ_B^2 and the size of the samples involved (Weiss et al., 2017). This robustness probably reflects the large amount of random estimation error that exists in site-specific impact estimates, which tends to be normally distributed due to the Central Limit Theorem. It could also partly reflect true impacts that are approximately normally distributed across sites.

Rearranging terms in Equation 17 suggests how we could "back out" a limited estimator ($\hat{\tau}^2_{B(j,\,LIMITED)}$) of τ^2_B from information for a single site, j if we knew the value of β.

$$\hat{\tau}^2_{B(j,LIMITED)} = \left(\hat{B}_j^{OLS} - \beta\right)^2 - V_j \tag{18}$$

This reasoning suggests that we could pool site-specific estimates as follows:

$$\hat{\tau}^2_{B(POOLED)} = \sum_{j=1}^{J} \frac{\left[\left(\hat{B}_j^{OLS} - \beta\right)^2 - V_j\right]}{J} \tag{19}$$

Equation 19 is a "method-of-moments" estimator in that it substitutes sample moments $\left(\hat{B}_j^{OLS} - \beta\right)^2$ for the expected values of these moments $E\left[(\hat{B}_j - \beta)^2\right]$ and solves for an estimator ($\hat{\tau}^2_{B(POOLED)}$). Although this approach is intuitively appealing, it raises two issues.

First, we must know the value of β in order to compute ($\hat{\tau}^2_{B(POOLED)}$). In this case, it seems natural to substitute a sample-based estimate of β from Equation 14. However, Equation 14 requires knowledge of τ^2_B. This "chicken-and-egg" problem suggests 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:

$$\tau^2_{B(FINAL)} = \sum_{j=1}^{J} w_j \left[\left(\hat{B}_j^{OLS} - \beta\right)^2 - V_j\right] \tag{20}$$

As in random-effects meta-analysis, the optimal weight (w_j) for Equation 20 is $\frac{(\tau^2_B + V_j)^{-2}}{\sum_{j=1}^{J}(\tau^2_B + V_j)^{-2}}$ under the assumption that program effects are normally distributed across sites (see Hedges & Olkin, 2014; Raudenbush, 1994; Raudenbush & Bryk, 2002). Maximum likelihood analysis thus uses an iterative procedure that alternates between estimating Equation 14 and Equation 20.

Consequences of Heterogeneous Individual-Level Outcome Variances

This section addresses issues that can arise when individual-level residual outcome variances differ between treatment and control group members or across study sites.

Treatment and Control Group Variance Differences

Suppose that the individual-level outcome variance for treatment group members (σ^2_1) differs from that for control group members (σ^2_0). This T/C heteroskedasticity can be caused by individual variation in program effects (Raudenbush & Bloom, 2015). To simplify our discussion, assume for the moment that σ^2_1 and σ^2_0 are constant across sites. Then, as noted

earlier:

$$V_j = \frac{\sigma_1^2}{n_{1j}} + \frac{\sigma_0^2}{n_{0j}} \tag{21}$$

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

Equation 21 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 in expectation, produces the following expression for apparent estimation error (V_j^A) at site j (see online Appendix A):

$$V_j^A = \frac{\sigma_1^2}{n_{0j}} + \frac{\sigma_0^2}{n_{1j}} \tag{22}$$

Note that Equation 22 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 21 for the true estimation error variance (V_j). Consequently, V_j^A is a biased estimator of V_j unless: (a) the treatment and control group samples are the same size ($n_{1j} = n_{0j}$) or (b) there is no T/C heteroskedasticity ($\sigma_1^2 = \sigma_0^2$). Thus:

$$V_j^A = V_j + Bias\left(V_j^A\right) \tag{23}$$

where, as demonstrated in online Appendix A:

$$Bias\left(V_j^A\right) = \frac{2\left(\sigma_1^2 - \sigma_0^2\right)\left(\overline{T}_j - 0.5\right)}{n_j \overline{T}_j\left(1 - \overline{T}_j\right)} \tag{24}$$

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:

$$Bias\left(\tau_{Bj}^2\right) = -Bias\left(V_j^A\right) = \frac{2\left(\sigma_1^2 - \sigma_0^2\right)\left(0.5 - \overline{T}_j\right)}{n_j \overline{T}_j\left(1 - \overline{T}_j\right)} \tag{25}$$

Equation 25 indicates that the magnitude of this bias (which can be positive or negative) depends on: the degree of T/C heteroskedasticity $\left(\sigma_1^2 - \sigma_0^2\right)$; the degree of T/C sample imbalance $((0.5 - \overline{T}_j)$ or $(\overline{T}_j(1 - \overline{T}_j))$; 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 members using existing software such as HLM, SAS *PROC MIXED, MLwiN,* or R. And no harm is done if $\sigma_1^2 = \sigma_0^2$ and we specify separate T/C outcome variances. Online Appendix B provides illustrative SAS code for estimating our cross-site impact variation model with separate values for σ_1^2 and σ_0^2.

Not only does estimating separate values for σ_1^2 and σ_0^2 have a methodological advantage, but it also can provide useful substantive information. For example, Raudenbush and Bloom (2015) demonstrate how a treatment and control group difference in this variance is evidence that program impacts vary across individuals within sites. They also note that if the treatment group variance is smaller than the control group variance, this indicates that the program reduced disparities in the outcome of interest. To assess the statistical significance of the observed difference between estimates of σ_1^2 and σ_0^2 one can use a simple F test of their ratio.

Cross-Site Heteroskedasticity

Now consider what happens if individual-level outcome variances differ across study sites, which we refer to as cross-site heteroskedasticity. To address this issue, one's first instinct might be to estimate separate treatment- and control-group outcome variances for each site. However, online Appendix A demonstrates that doing so can overstate τ_B^2, often by a lot.

To see how this can occur, note that the estimated sampling variance, $\hat{V}_j$, for an OLS-estimated mean program effect, $\hat{B}_j^{OLS}$, for a given site, j, is

$$\hat{V}_j = \frac{\hat{\sigma}_{1j}^2}{n_{1j}} + \frac{\hat{\sigma}_{0j}^2}{n_{0j}} \tag{26}$$

Thus $\hat{V}_j$ will understate the true sampling variance for site j when $\hat{\sigma}_{1j}^2$ and $\hat{\sigma}_{0j}^2$ understate the true values of σ_{1j}^2 and σ_{0j}^2 (which can happen by chance for a sample) and $\hat{V}_j$ will overstate the true sampling variance for site j when $\hat{\sigma}_{1j}^2$ and $\hat{\sigma}_{0j}^2$ overstate the true values of σ_{1j}^2 and σ_{0j}^2 (which also can happen by chance for a sample).

Consider how these chance estimation errors systematically influence the mth-iteration maximum likelihood estimator of cross-site impact variation, $\hat{\tau}_{B(m)}^2$, where:[13]

$$\hat{\tau}_{B(m)}^2 = \sum_{j=1}^{J} \frac{\left(\tau_{B(m-1)}^2 + \hat{V}_j\right)^{-2}\left(\left(\hat{B}_j^{OLS} - \hat{\beta}\right)^2 - \hat{V}_j\right)}{\sum_{j=1}^{J}(\tau_{B(m-1)}^2 + \hat{V}_j)^{-2}} \tag{27}$$

For sites with $\hat{\sigma}_{1j}^2$ and $\hat{\sigma}_{0j}^2$ that are smaller than their true values, $\hat{V}_j$ will understate the true value of V_j. This will have two compounding effects on those sites' contributions to $\hat{\tau}_{B(m)}^2$. One effect is a tendency to *overstate* the true squared deviations of their program effects from the grand mean program effect. In other words, when $\hat{V}_j$ is too small $\left(\left(\hat{B}_j^{OLS} - \hat{\beta}\right)^2 - \hat{V}_j\right)$ will tend to be too large. The second effect is to *over-weight* the contribution of these sites to $\hat{\tau}_{B(m)}^2$. In other words, when $\hat{V}_j$ is too small, the weight for site j $\left(\hat{\tau}_{B(m-1)}^2 + \hat{V}_j\right)^{-2}$ will be too large. Hence these sites will tend to over-weight an overestimate of $\hat{\tau}_{B(m)}^2$. The opposite will tend to occur for sites where, by chance, $\hat{\sigma}_{1j}^2$ and $\hat{\sigma}_{0j}^2$ are larger than their true values. Thus $\hat{\tau}_{B(m)}^2$ will tend to over-weight sites with overestimated values for $\hat{\tau}_{B(\mathrm{j,\,LIMITED})}^2$ and under-weight sites with under-estimated values for $\hat{\tau}_{B(\mathrm{j,\,LIMITED})}^2$ τ_B^2, on average.

[13] Equation 27 is for full maximum likelihood. The corresponding expression for restricted maximum likelihood is slightly more complex.

Online Appendix A describes how the use of site-specific estimates of σ_1^2 and σ_0^2 also can cause one to overstate the statistical significance of $\hat{\tau}^2_{B(m)}$ and presents simulation results that demonstrate that this bias increases markedly as site sample size declines and the number of sites increases. These findings indicate that with total site samples of 20 or fewer persons, this inferential problem can be severe.

In addition, online Appendix A demonstrates that cross-site heteroskedasticity is only a problem for pooling estimates of σ_1^2 and σ_0^2 across sites if all the following conditions hold: (a) sample sizes vary substantially across sites, (b) residual variances vary substantially across sites, and (c) sample sizes and residual variance are correlated substantially across sites. In addition, later in this article, we describe and illustrate a sensitivity test of the robustness of estimates of τ_B^2 or τ_B that are based on full-sample estimates of σ_1^2 and σ_0^2 plus an alternative specification of residual variances to use if the sensitivity test suggests that full-sample estimates of σ_1^2 and σ_0^2 produce badly biased estimates of τ_B^2.

Estimating Site-Specific Mean Program Effects

For some purposes, researchers might want to study the operation of sites with the most beneficial or least beneficial effects. Also, local policymakers might want to know the mean program effect for their particular site. Thus it is sometimes important to produce site-specific estimates of program effects.[14]

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 large if site samples are small.

A second option would exist if we knew that all sites had the same program effect. In that case, 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 ($\hat{\beta}$). 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 (Lindley & Smith, 1972; see review by Morris, 1983). 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[15] of the mean program effect for a given site is its posterior mean, $\hat{B}_j^{PM}$, (Raudenbush & Bryk, 2002), where

$$\hat{B}_j^{PM} \equiv E\left(B_j | Y, \beta, \tau_B^2, V_j\right) = \lambda_j \hat{B}_j^{OLS} + \left(1 - \lambda_j\right)\beta \tag{28}$$

Equation 28 represents a weighted composite of the site-specific 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,

[14] This section is an example of small-area parameter estimation. For an extensive discussion of this more general topic, see Rao and Molina (2015).

[15] In this case, we define the best estimate to be the value of $\hat{B}_j^{PM}$ that minimizes the expected sum of squared errors of estimation, $(E[\sum_{j=1}^{J}(\hat{B}_j^{PM} - B_j)^2])$. Bayes theorem indicates that this optimal value is the posterior mean defined by Equation 28 which follows.

λ_j and the weight accorded β is $(1-\lambda_j)$, where

$$\lambda_j = \frac{\tau_B^2}{\tau_B^2 + V_j}. \tag{29}$$

Holding τ_B^2 constant, reliability increases as V_j decreases. Holding V_j constant, reliability increases as τ_B^2 increases. At one 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 grand mean effect (β). At the other extreme, when there is very little site-specific impact estimation error (which will only occur for sites with very large samples) and a great deal of true cross-site impact variation, the reliability of the site-specific impact estimate will be close to one and almost all of the weight will be placed on the site-specific impact estimate.

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 ($\hat{B}_j^{EB}$) for each site j, where:

$$\hat{B}_j^{EB} = \hat{\lambda}_j \, \hat{B}_j^{OLS} + \left(1 - \hat{\lambda}_j\right)\hat{\beta} \tag{30}$$

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 & 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 β 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 *fully* Bayesian methods that address this concern (e.g., Gelman, Hill, & Yajima, 2012; Spiegelhalter, Thomas, Best, & Lunn, 2003) but a discussion of this is beyond the scope of the present article.

Note last that Equation 30 "shrinks" unreliable site-specific estimates of program impacts toward the estimated grand mean, $\hat{\beta}$. The idea here is that the grand mean is a sensible predicted value of the impact in the absence of an a priori belief that particular sites or particular subsets of sites can be expected to produce impacts that are larger or smaller than average. However, if the analyst has prior hypotheses about sources of site-specific differences in effectiveness, they should be incorporated into the empirical Bayes or fully Bayes estimates of site-specific impacts. In that case, site-specific estimates will be "shrunk" toward different predicted values (see Raudenbush & Bryk, 2002, Chapter 3 for details on such "conditional shrinkage").

Reporting a Cross-Site Impact Distribution

Although empirical Bayes estimators (often referred to as "shrinkage estimators") have the smallest mean squared error for predicting a specific parameter value (Lindley & Smith, 1972), like a mean program effect for a specific site, these estimators are biased toward the overall mean of those parameter values (Raudenbush & Bryk, 2002). Thus in the case of a cross-site distribution of program effects, empirical Bayes estimators will tend to "over-shrink" site-

specific OLS impact estimates toward the cross-site grand mean impact. Hence, the cross-site variance of empirical Bayes estimates ($Var(\hat{B}_j^2)$) will tend to understate the cross-site variance of true mean program effects (τ_B^2). Consequently, for a given sample, the estimated cross-site variance of empirical Bayes estimates of site-level mean impacts, $\widehat{Var}(\hat{B}_j^{EB})$, where:

$$\widehat{Var}\left(\hat{B}_j^{EB}\right) = \left(\frac{\sum_j \left(\hat{B}_j^{EB} - \hat{\beta}\right)^2}{J-1} \right) \tag{31}$$

will tend to be smaller than the corresponding model-based estimate of the true cross-site impact variance ($\hat{\tau}_B^2$).

To better represent a cross-site distribution of program effects, one can adjust empirical Bayes estimates of the average treatment effect at each site in a way that makes $\widehat{Var}(\hat{B}_j^{EB})$ equal $\hat{\tau}_B^2$. Thus when reporting a cross-site distribution of these adjusted empirical Bayes estimates, the amount of impact variation reflected is consistent with $\hat{\tau}_B^2$, which is the best existing estimate of true cross-site variation. This idea was introduced by Louis (1984) and a simple approach for making it operational is developed in online Appendix C.[16] The adjustment presented in online Appendix C "stretches" the distance between each empirical Bayes estimate and the estimated cross-site mean by a constant proportion, γ, where:

$$\gamma = \frac{\widehat{Var}\left(\hat{B}_j^{EB}\right)}{\hat{\tau}_B^2} \tag{32}$$

Each adjusted empirical Bayes estimate $\hat{B}_j^{AEB}$ is thus computed as:

$$\hat{B}_j^{AEB} = \hat{\beta} + \frac{1}{\sqrt{\gamma}}\left(\hat{B}_j^{EB} - \hat{\beta}\right) \tag{33}$$

This adjustment is similar to that in the literature (e.g., Louis, 1984; Rao, 2015).

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. Specifically, the present approach will provide consistent estimates when site-specific precision weights are uncorrelated with site-specific program effects (Raudenbush & Bloom, 2015; Raudenbush & Schwartz, 2017). 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 overstate the cross-site mean effect (β).[17]

[16] Judkins and Liu (2000), among others, have discussed this issue.

[17] To see how this might occur, consider a hypothetical group of charter school lotteries where schools that are more effective than average are more popular than average and thus are more heavily oversubscribed than average. Assuming a roughly similar number of openings for each charter school, this would imply that sample sizes (and thus weights) are larger than average for charter schools that are more effective than average.

Recall that the site weight for estimating a random-coefficient mean effect is proportional to $\left(\hat{\tau}_B^2 + \hat{V}_j\right)^{-1}$, where $\hat{V}_j$ depends on the site sample size (n_j) and treatment allocation ($\overline{T}_{ij}$). Consequently, the issue boils down to the strength of the cross-site association between true program effects (B_j) and n_j or $\overline{T}_j$. 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 only 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 & Bloom, 2015; Raudenbush & Schwartz, 2017). 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 trade-off between potential inconsistency from the present method and a potential loss of precision from a consistent method that weights sites equally. Because little is currently known about this trade-off in practice, we and our colleagues are currently studying it through simulations and re-analyses of multisite trials.

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 large multisite trials conducted by MDRC: the Greater Avenues for Independence (GAIN) project conducted in 22 local welfare offices from six California counties (Riccio & Friedlander, 1992); Project Independence conducted in 10 local welfare offices from nine Florida counties (Kemple & Haimson, 1994); and the National Evaluation of Welfare-to-Work Strategies conducted in 27 local welfare offices from seven states (Hamilton, 2002). Programs at the sites in these studies comprised a varying mix of human development services which included, among other things, basic education geared toward obtaining a high school diploma or GED plus vocational training on the job or in a training center plus job-search assistance and placement services.

Background

Because these new programs were mandatory, all treatment group members in the analysis sample were exposed to them, even if only to their threat of sanctions (loss of welfare payments) for nonparticipation. Furthermore, no control group members experienced the new programs (although some might have received related services elsewhere). Compliance with random assignment was thus complete and the average effect of program assignment was equal to the average effect of program participation (Bloom et al., 2003).

The outcome measure for the present analysis is sample members' total earnings during their first two years after random assignment, reported in constant dollars.[18] 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

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

(local welfare offices) with a total of 69,399 individuals randomized within site to a new mandatory welfare-to-work program or to a control group.

Findings for the present analysis were obtained using restricted maximum likelihood in SAS to estimate a two-level model such as Equations 7–9, with a fixed intercept for each of the 59 sites in the analysis sample and a random impact coefficient for these sites plus separate individual-level outcome variances for treatment and control group members.

Findings

The estimated cross-site mean program effect ($\hat{\beta}$) was \$878 ($p \leq 0.0001$) and the estimated cross-site variance of program effects ($\hat{\tau}_B^2$) was $(771)^2$ ($p \leq 0.0001$).[19] The estimated residual variance of individual-level outcomes for treatment group members ($\hat{\sigma}_1^2$) was $(10{,}084)^2$ and that for control group members ($\hat{\sigma}_0^2$) was $(8{,}931)^2$. Their ratio of 1.27 ($p < 0.001$) indicates that program effects varied across individuals within sites.[20]

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 ($\hat{B}_j^{OLS}$).[21] Note that even with the very large site samples for the present analysis, OLS impact estimates dramatically overstate cross-site impact variation because their cross-site variation 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 as noted earlier, shrink each OLS estimate toward the estimated cross-site mean ($\hat{\beta}$). As can be seen, the empirical Bayes estimates vary by much less than their OLS counterparts. This is because an empirical Bayes estimator nets out cross-site variation in estimates due to site-level estimation error (V_j).

However, as noted above, empirical Bayes estimators also tend to understate true cross-site impact variation. Hence, the cross-site variance of our empirical Bayes estimates ($\widehat{Var}(\hat{B}_j^{EB})$) is only 52% of our estimated cross-site variance of true mean program effects ($\hat{\tau}_B^2$). To address this issue, the middle histogram in Figure 1 presents a histogram of adjusted empirical Bayes estimates that were discussed earlier and derived in online Appendix C.

Now compare the three histograms. Note first that they all have approximately the same cross-site mean, which ranges from \$878 to \$906. 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 \$558, and constrained empirical Bayes estimates having a standard deviation of \$771.

Because constrained empirical Bayes estimates reflect cross-site variation most accurately, they are the best guide for interpreting findings for the cross-site impact distribution. These findings suggest that only about 6 of the 59 local programs examined have negative effects (i.e., they were less effective than existing alternatives). In contrast, our OLS estimates indicate that 15 of the 59 local programs had negative effects, and the magnitudes of

[19] 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.
[20] The statistical significance of this ratio was determined by an *F* statistic.
[21] 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.

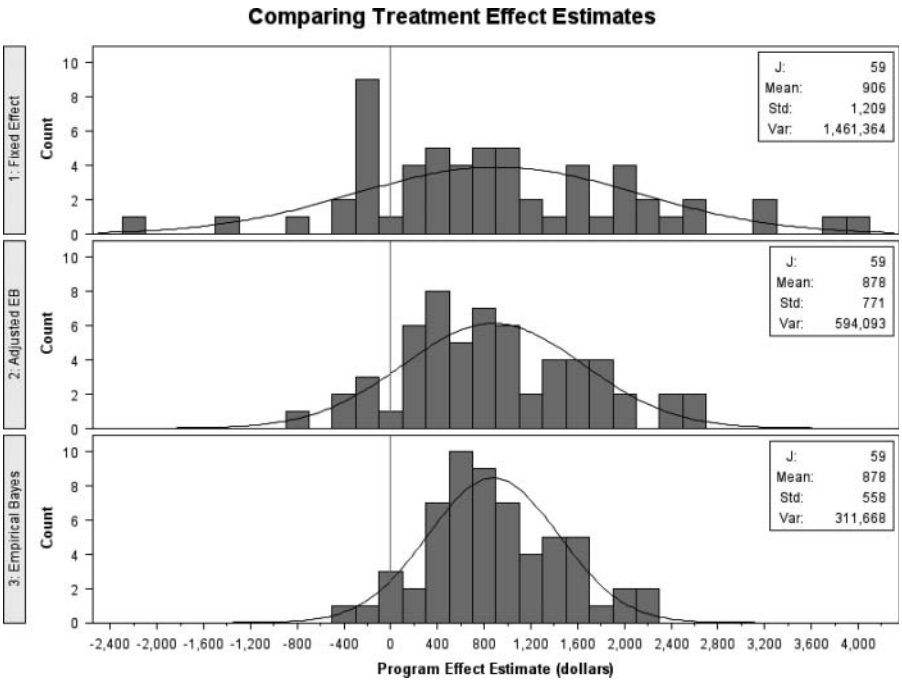

Figure 1. Cross-site distributions of estimated effects of 59 welfare-to-work programs.

these negative effects tend to be much larger than those for our constrained empirical Bayes estimates.

Diagnostics

Figures 2–4 provide useful diagnostics 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 Excel using SAS output, is a simple way to illustrate what is known and not known about site-specific impacts and thereby provides a visual representation of the "raw material" for our analysis of cross-site impact variation. 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% 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; the more precise these estimates are on average, the narrower the confidence band around these estimates and their slope will be.

A second tool for assessing site-specific impact estimates is the "profile likelihood" plot in Figure 3 (Murphy & Vander der Vaart, 2000). This plot, which was produced using a forthcoming version of 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 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 expand from a single point for $\hat{\tau}_B^2 = 0$ to a broad band of lines as $\hat{\tau}_B^2$ increases. This pattern

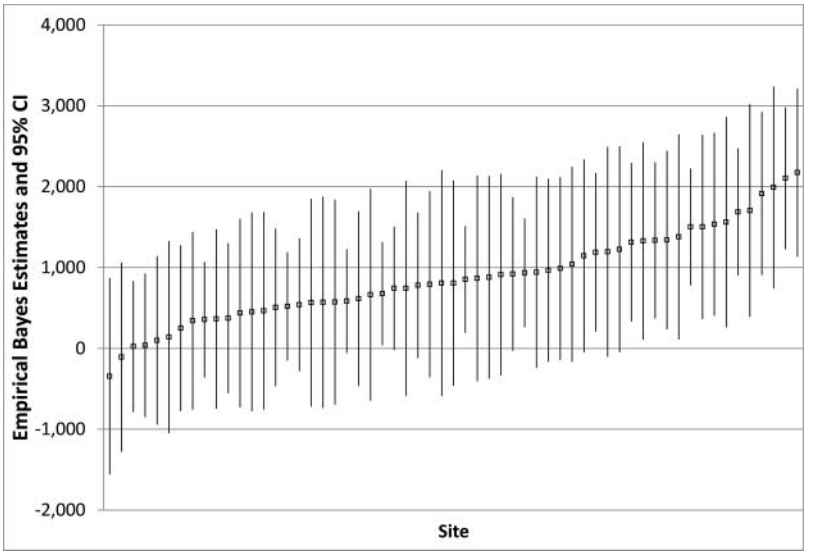

Figure 2. Caterpillar plot of empirical Bayes estimates of the effects of 59 welfare-to-work programs.

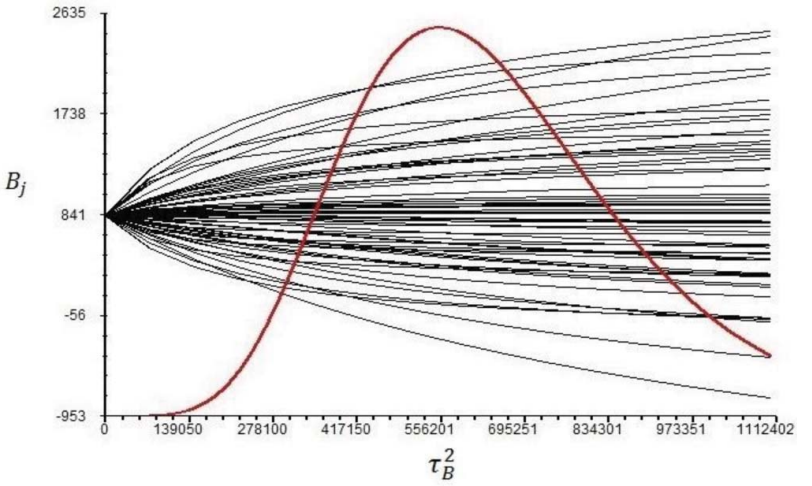

Figure 3. Profile likelihood graph of empirical Bayes estimates of the effects of 59 welfare-to-work programs.

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 for $\hat{B}_j^{OLS}$ increases relative to the weight for $\hat{\beta}$. 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 $\hat{B}_j^{OLS}$.[22]

What makes a profile likelihood plot particularly useful is the superimposed plot of the profile likelihood function evaluated for each possible value of $\hat{\tau}_B^2$. (See Rubin, 1981, which introduced this idea using a slightly different approach.) This is the inverted U-shaped curve in the figure. The vertical axis of this curve provides a measure of the relative plausibility of

[22] The value for $\hat{\beta}$ here is its fixed-effect estimate.

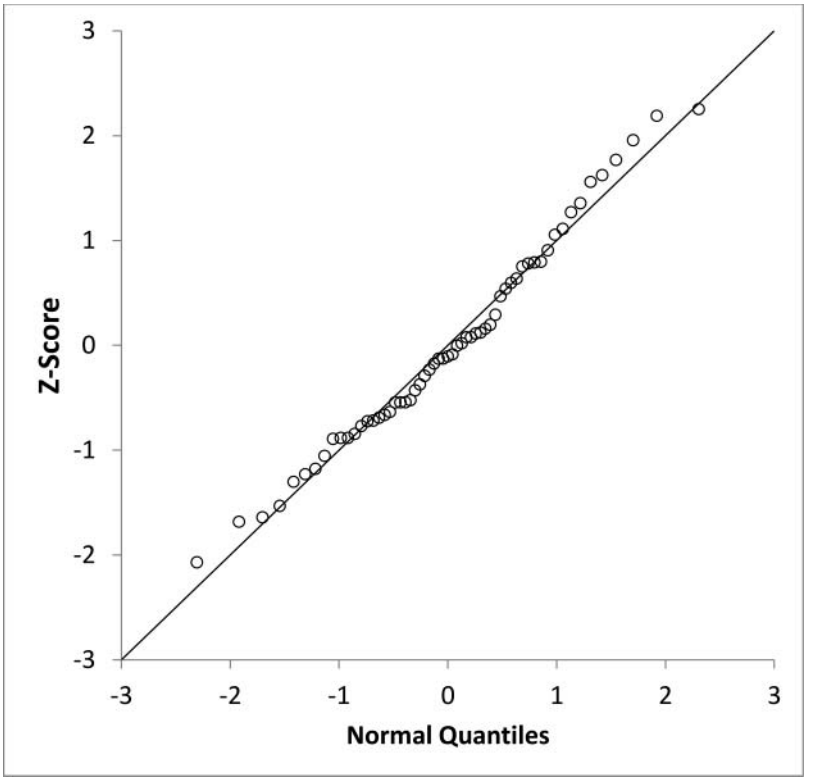

Figure 4. Q-Q plot of $\frac{\hat{B}_j^{OLS} - \hat{\beta}^{FIXED}}{\sqrt{\hat{V}_j}}$ for the welfare-to-work example.

each possible value of $\hat{\tau}_B^2$ given the existing data. The steeper and narrower this profile likelihood 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 peaked at a point that differs substantially from zero with a very small proportion of the weight of the profile plot being near zero. Thus we can be confident that program effects vary appreciably across sites. This conclusion is consistent with the fact that the 95% confidence interval produced by HLM for our estimate of τ_B^2 ranges from $(527)^2$ to $(1{,}049)^2$, which implies a 95% confidence interval for τ_B of \$527 to \$1,049.[23]

This relatively high degree of confidence about the presence of substantial cross-site impact variation reflects the especially large site samples 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 it.

A profile likelihood plot also can help to identify sites with especially large or small impacts. In Figure 3 there are several sites with estimated impacts near \$1,700, which is about twice the cross-site average. At the opposite extreme, there are several sites with estimated negative impacts that are small relative to the size of the positive

[23] 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, HLM uses a logarithmic transformation of the variance and the associated transformation of the information matrix. HLM then exponentiates the upper and lower limits of this interval to obtain an appropriate asymmetric confidence interval for $\hat{\tau}_B{}^2$.

estimated impacts for the majority of sites. Furthermore, over the plausible range of values for τ_B^2 (i.e., values for which the likelihood is high), the same sites are at these extremes. Hence, our belief about the relative program impacts for these sites is not dependent on errors of estimation for τ_B^2.[24]

Another useful diagnostic for analyses of cross-site impact variation based on the present approach is a Q-Q plot of the values for $\frac{\left(\hat{B}_j^{OLS} - \hat{\beta}^{FIXED}\right)}{\sqrt{\hat{V}_j}}$, which visually illustrates the extent to which these estimates are normally distributed. We focus on this issue because in order to use a Q statistic to test the statistical significance of $\hat{\tau}_B^2$, the values of $\frac{\left(\hat{B}_j^{OLS} - \hat{\beta}^{FIXED}\right)}{\sqrt{\hat{V}_j}}$ must be normally distributed. Figure 4 presents a Q-Q plot for our welfare-to-work example. It plots standardized z-scores for the values of $\frac{\left(\hat{B}_j^{OLS} - \hat{\beta}^{FIXED}\right)}{\sqrt{\hat{V}_j}}$ on the vertical axis based on the mean and standard deviation of these values against the quantile equivalent of these values for a normal distribution on the horizontal axis. The more closely these points approximate a 45-degree line, the more closely they reflect a normal distribution. As can be seen, this approximation is very good for the present example. Thus we can be confident in our test rejecting the null hypothesis that $\tau_B^2 = 0$.

Next, consider the implications for the welfare-to-work example of two potential biases discussed earlier for our impact estimation model. First, is a potential bias in our estimator of the cross-site mean program effect ($\hat{\beta}$) due to a correlation between the precision weights used to average impact estimates across sites and the mean program impact for each site (B_j). Raudenbush and Schwartz (2017) present an estimator of this bias. Online Appendix D discusses this estimator and extends it to produce an approximate confidence interval estimate of an upper bound on the bias. Applying Equation D.8 to the work–welfare data, we estimate a bias of $-$ \$29 for our mean program impact estimate ($\hat{\beta}$) of \$878.[25] Thus our single best estimate of this bias is very small.

However, it is difficult to estimate a confidence interval for this estimated bias. Thus to deal with this limitation, online Appendix D derives an upper-bound estimator of the bias for which it then derives an approximate confidence interval. Using this approach we estimate a negative bias with a maximum magnitude of \$74 and an approximate 95% confidence interval of $-$ \$232 to \$83. This is further evidence that the bias is small.

In addition, our estimator of β is more precise than its consistent equally weighted alternative, with an estimated standard error of \$139 for the present estimator versus \$157 for its equally weighted alternative.[26] Consequently, an equally weighted estimator would need a sample that is 28% larger than that for the present estimator to achieve the same precision.[27]

The second potential bias in our method discussed earlier is the possibility that if: (a) sample sizes differ substantially across sites, (b) individual-level residual variances differ

[24] Rather than examining the profile likelihood in this way, a fully Bayesian approach would study the posterior distribution of τ_B^2 and its implications for the *ensemble* of site-specific impact estimates (Seltzer, Wong, & Bryk, 1996).

[25] Online Appendix D refers to the estimated cross-site mean impact produced by our proposed model as β_{FIRC}, whereas the text of the present article simply refers to this estimator as $\hat{\beta}$.

[26] The estimated standard error ($\hat{se}\left(\hat{\beta}^{EW}\right)$) of an equally weighted cross-site mean impact estimate ($\hat{\beta}^{EW}$) was computed as $\hat{se}(\hat{\beta}^{EW}) = \frac{\sqrt{\sum_{j=1}^{J}(\hat{B}_j^{OLS} - \hat{\beta}^{EW})^2/(J-1)}}{J}$.

[27] To see this, note that the error variance of an estimator (its standard error squared) is inversely proportional to its sample size and $\frac{(\widehat{se}(\hat{\beta}^{EW}))^2}{(\widehat{se}(\hat{\beta}^{FIRC}))^2} = \frac{(157)^2}{(139)^2} = 1.28$, where $\hat{\beta}^{FIRC}$ represents the present estimator.

substantially across sites, *and* (c) these factors are substantially correlated, estimates of τ_B can be biased substantially. One sensitivity test for such bias is to stratify sites according to their sample sizes and specify a separate residual variance for each stratum when estimating our basic impact variation model (Equations 7–9). Doing so reduces the potential for bias because the variation in site sample size for each estimated residual variance (one for each treatment-group or control-group stratum) can be much smaller than the variation in sample size for corresponding full-sample estimates of residual variances. In addition, estimating residual variances within fairly large strata avoids the other potential bias in $\hat{\tau}_B$ from estimating separate residual variances for many small sites.

Online Appendix E describes how we grouped our 59 welfare-to-work sites into five treatment group strata of roughly equal size and five control group strata of roughly equal size. We then estimated our basic cross-site impact variation model with one residual variance for each stratum. This produced an estimate of τ_B equal to \$724 ($p \leq 0.0001$), which is very similar to our original estimate of \$771 ($p \leq 0.0001$). This finding is especially encouraging because the new estimate reflects a cross-site standard deviation of site sample sizes per estimated residual variance that is only *one third* of its full-sample counterpart for treatment group members and *one fourth* of its full-sample counterpart for control group members. Thus our original results are robust to potential violation of the assumption of cross-site homoskedasticity. Furthermore, our original estimate of β (\$878, $p \leq 0.0001$) is almost identical to that for the new model (\$871, $p \leq 0.0001$).

Of course, these findings represent only one empirical example, and further research is needed to assess their generalizability. Fortunately, however, if for a given application our sensitivity test suggests a serious bias from using full-sample estimates of residual variances, one can increase the number of strata until estimates of τ_B are no longer sensitive to further stratification. When using this strategy we recommend that each treatment-group stratum or control-group stratum contain at least 100 sample members to avoid bias in estimates of τ_B from small-sample estimates of residual variances (see online Appendix Table A.1).

Concluding Thoughts

The method we propose can be a valuable component of multisite evaluations of ongoing public programs such as Head Start (Puma et al., 2010) or Job Corps (Schochet, Burghardt, & McConnell, 2008); multisite studies of large-scale government interventions such as New York City's Small High Schools of Choice (Bloom & Unterman, 2014) or Massachusetts's charter schools (Angrist et al., 2013); and multisite studies of demonstration projects such as the Enhanced Reading Opportunities demonstration for high school students (Somers et al., 2010) or the enhanced coaching project for college students (Bettinger & Baker, 2014).

Specifically, this method is appropriate when there is interest in: (a) studying the effectiveness of interventions as they are operated by sites, (b) going beyond just studying the average effectiveness of interventions to quantify how this effectiveness varies across sites, and (c) going beyond just inferring findings to a study sample in order to generalize them to a population of sites.[28] We believe that the current growing interest in multisite "effectiveness" trials (to evaluate interventions under normal operating conditions in diverse settings) and

[28] As noted earlier, we believe that such generalizations are important whether or not it is possible to clearly and simply define the population represented by a study sample.

"scale-up" studies (to evaluate interventions in a large numbers of diverse settings) reflects a growing interest among policymakers, practitioners, researchers, and research funders in studying these issues. Thus we expect to see a growing number of situations for which our proposed method is appropriate. Nonetheless, other methods are appropriate for studies that focus on mean intervention effects for individuals instead of sites or for studies that focus on mean intervention effects for a specific sample of sites.[29]

One issue to consider when *designing* future multisite trials to study cross-site impact variation for a population of sites is the number and size of site samples needed to do so with adequate statistical power or precision. Although this issue is beyond the scope of the present article, analyses reported in Table 2 of Bloom and Spybrook (2017) indicate that a balanced trial with 20 sites and 100 sample members per site or with 50 sites and 50 sample members per site can have a minimum detectable cross-site effect-size standard deviation of roughly 0.16σ.[30]

To help interpret this finding, note that estimates reported in Weiss et al. (2017, Table 4) based on data from 16 multisite trials in education and training research indicate that the cross-site standard deviation of program effect sizes is often between 0.10σ and 0.25σ. This suggests that study samples with 2,000 to 2,500 persons from between 20 and 50 sites might be adequately powered to detect cross-site impact variation of a realistic magnitude.

One issue to consider when *analyzing* future multisite trials to study cross-site impact variation is whether the results of an "omnibus" statistical test of the existence of such variation should be used to determine whether to try to predict it.[31] 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 moderator that is anticipated to have substantial predictive power. This is because under that condition, an omnibus test can have less statistical power than a "focused" test about a difference between mean program effects for subgroups of sites (e.g., rural versus urban). Hedges and Pigott (2001) note this power difference in the context of meta-analysis and Rosenthal and Rosnow (1985) note it in the context of experimental design. In addition, online Appendix F explores the factors that determine this trade-off and provides a numerical example of their potential influence. On the other hand, if an omnibus test fails to detect cross-site impact variation, it should serve as a major caveat for ex post facto exploratory hypothesis tests about site-level impact predictors.

Last, we note the potential limitation of the present method that was identified earlier—that it can provide inconsistent estimates if there is an appreciable correlation between site precision weights and true program effects and thus alternative methods might be needed (Raudenbush & Bloom, 2015; Raudenbush & Schwartz, 2017). 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 trade-off between these methods. Although this trade-off was favorable for the present method in our welfare-to-work example, it remains

[29] For example, Schochet (2008) discusses, among other things, methods for estimating mean effects for a fixed sample of individuals or sites.

[30] This minimum detectable effect size standard deviation assumes: (a) statistical significance at the 0.05 level, (b) 80% statistical power, (c) a site-level intraclass correlation of 0.15, and (d) a within-site R-square for a baseline covariate (e.g., a pretest score) equal to 0.4.

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

to be seen what it will be for a broader range of programs, participant populations, and local settings.

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 (see Weiss et al., 2014). Likewise it will be essential for the next generation of multisite trials to collect high-quality data on those predictors. We hope that together, these new statistical methods, conceptual frameworks, and high-quality data can produce considerable knowledge for helping policymakers, practitioners, and researchers better understand when, how, why, and for whom programs do or do not work.

Acknowledgments

The authors thank Christina Weiland for simulation findings that informed the article; Jeffrey Smith, Winston Lin, and several anonymous referees for their feedback on the article; and Himani Gupta for her careful checking of the information in the article.

Funding

This paper was funded by grant #201500035 from the Spencer Foundation; grant #183631 from the William T. Grant Foundation; subcontract C604-3071-10-2, contract 90YR0049/01, and contract HHSP2332009564WC/HHSP23337043Y from the Office of Policy Research and Evaluation of the U.S. Department of Health and Human Services; grant R305D140012 from the Institute of Education Sciences of the U.S. Department of Education; and the Judith Gueron Fund of MDRC. All views expressed herein are those of the authors and do not necessarily reflect those of our funders or reviewers.

ARTICLE HISTORY

Received 3 March 2014
Revised 14 November 2016
Accepted 21 November 2016

References

Angrist, J. D., Pathak, P., & Walters, C. R. (2013). Explaining charter school effectiveness. *American Economic Journal: Applied Economics*, *5*(4), 1–27. doi:10.1257/app.5.4.1

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.

Blom, G. (1958). *Statistical estimates and transformed beta variables*. New York, NY: John Wiley and Sons.

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

Bloom, H. S., & Spybrook, J. (2017). Assessing the precision of multisite trials for estimating the parameters of a cross-site population distribution of program effects. *Journal of Research on Educational Effectiveness.* Advance online publication. doi:10.1080/19345747.2016.1271069

Bloom, H. S., & Unterman, R. (2014). Can small high schools of choice improve educational prospects for disadvantaged students? *Journal of Policy Analysis and Management, 33*(2), 290–319.

Bloom, H. S., & Weiland, C. (2015, March). *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, NY: MDRC.

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

Greene, W. H. (2008). *Econometric analysis* (6th ed.). Upper Saddle River, NJ: Prentice Hall.

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.

Hedges, L.V., & Olkin, I. (2014). *Statistical methods for meta analysis*. San Diego, CA: Academic Press.

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

Judkins, D. R., & Liu, J. (2000) Correcting the bias in the range of a statistic across small areas. *Journal of Official Statistics*, *16*(1), 1–13.

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

Konstantopoulos, S. (2011). 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.

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.,. . . Tagnotta N. (2013). *Evaluation of the i3 scale-up of Reading Recovery: Year one report, 2011–12* (CPRE Research Report #RR-76). Philadelphia, PA: Consortium for Policy Research in Education.

Morris, C. N. (1983). Parametric empirical Bayes inference: Theory and applications. *Journal of the American Statistical Association, 78*(381), 47–55.

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

Pearson, E. S. (1931). The analysis of variance in cases of non-normal variation. *Biometrika*, *23*(1–2), 114–133.

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

Rao, J. N. K., & Molina, I. (2015). *Small area estimation* (2nd ed.). Hoboken, NJ: John Wiley.

Raudenbush, S. W. (1994). Analyzing effect sizes: Random effects models. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 302–332). New York, NY: Russell Sage Foundation.

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

Raudenbush, S. W., & Bryk, A. S. (1985). Empirical Bayes meta-analysis. *Journal of Educational and Behavioral Statistics*, *10*(2), 75–98.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage Publications.

Raudenbush, S. W., Reardon, S., & Nomi, T. (2012). Statistical analysis for multisite trials using instrumental variables. *Journal of Research on Educational Effectiveness, 5*(3), 303–332.

Raudenbush, S. W., & Schwartz, D. (2017). *Estimation in multisite randomized trials with heterogeneous treatment effects* (Occasional Paper). Department of Sociology, University of Chicago, Chicago, IL.

Riccio, J., & Friedlander, D. (1992). *GAIN: Program strategies, participation patterns and first-year impacts in six counties*. New York, NY: MDRC.
Rosenthal, R., & Rosnow, R. L. (1985). *Contrast analysis: Focused comparisons in the analysis of variance*. Cambridge, England: Cambridge University Press.
Rubin, D. (1981). Estimation in parallel randomized experiments. *Journal of Educational Statistics, 6* (4), 377–401.
Scheffe, H. (1959). *The analysis of variance*. New York, NY: John Wiley & Sons.
Schochet, P. Z. (2008). Statistical power for random assignment evaluations of education programs. *Journal of Educational and Behavioral Statistics, 33*(1), 62–87.
Schochet, P. Z., Burghardt, J., & McConnell, S. (2006). *National Job Corps Study and longer-term follow-up study: Impact and cost-benefit findings using survey and summary earnings records, final report*. Washington, DC: U.S. Department of Labor.
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.
Seltzer, M. H., Wong, W. W., & Bryk, A. S. (1996). Bayesian analysis in applications of hierarchical models: Issues and methods. *Journal of Educational and Behavioral Statistics*, *21*(2), 131–167.
Somers, M-A., Corrin, W., Sepanik, S., Salinger, T., Levin, J., & Zmach, C. (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.
Spiegelhalter, D., Thomas A., Best N., & Lunn N. (2003). *WINBUGS user manual*. Retrieved from http://www.mrc-bsu.cam.ac.uk/software/bugs/
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., Bloom, H. S., & Brock, T. (2014). A conceptual framework for studying the sources of variation in program effects. *Journal of Policy Analysis and Management*, *33*(3), 778–808.
Weiss, M. J., Bloom, H. S., Verbitsky Savitz, N., Gupta, H., Vigil, A., & Cullinan, D. (2017). *How much do the effects of education and training programs vary across sites? Evidence from past multisite randomized control trials*. Manuscript submitted for publication.