How Much Do the Effects of Education and Training Programs Vary Across Sites? Evidence from Existing Multisite Randomized Control Trials

Michael J. Weiss[a]

Howard S. Bloom[a]

Natalya Verbitsky Savitz[b]

Himani Gupta[a]

Alma Vigil[b]

Dan Cullinan[a]

May 13, 2016
Manuscript Under Review

[a]MDRC
[b]Mathematica

Funding: This paper was funded by a grant to MDRC from the Institute of Educational Sciences and a grant from the Spencer Foundation. The views expressed are those of the authors.

Abstract

Multisite randomized control trials randomly assign individuals to treatment arms within multiple sites. In so doing, they provide a unique opportunity to estimate both the overall average effect of a treatment (relative to its counterfactual condition) and the amount by which treatment effects vary across sites. Cross-site variation in these effects is of considerable substantive interest; furthermore, the design of multisite trials requires *a priori* knowledge of the magnitude of this effect variation in order to conduct accurate statistical power or precision analyses. Currently there is little empirical information and almost no theoretical foundation for anticipating how much treatment effects will vary across sites. This article is a first-step toward filling this void. To do so, we estimate the cross-site standard deviation of site-average treatment effects using data from 15 extant multisite randomized controlled trials of educational and training programs. We then illustrate the implications of these findings for the statistical precision of multisite trials and present hypotheses about factors that may predict the amount of cross-site variation in impacts of a given intervention (relative to its counterfactual).

Background

The last 15 years have seen a dramatic increase in the number of randomized control trials (RCTs) of educational programs (Spybrook, 2014). This surge has led to an unprecedented increase in information about the *average* causal effects of these interventions. It has also led to increased appreciation of, and interest in, variation in program effectiveness (Schochet, Puma, & Deke, 2014; Weiss, Bloom, & Brock, 2014), the importance of which has been long acknowledged (e.g., Abadie, Angrist, & Imbens, 2002; Bitler, Gelbach, & Hoynes, 2006; Bryk & Raudenbush, 1988; Djebbari & Smith, 2008; Friedlander & Robins, 1997; Heckman, 2001; Heckman, Smith, & Clements, 1997; Raudenbush & Liu, 2000). With this in mind, one important step toward improved understanding of variation in program effectiveness is to quantify the amount of variation that exists, which is the goal of the present paper.

There are many reasons why program effects might vary. Weiss et al. (2014) offer a simple framework for considering the sources of this variation. At the broadest level, this framework posits three sources of variation in program effects (referred to by the authors as the three C's): (1) Client Characteristics (e.g., the program is more effective than average for at-risk students), (2) Context (e.g., the program is more effective than average for students attending schools in low-crime neighborhoods), and (3) Treatment Contrast – the difference between services and activities experienced with and without access to the program (e.g., the program is more effective than average for students with a larger than average *difference* between the amount of time they practice word problems with and without access to the program).

While this framework was used by Weiss et al. (2014) to describe sources of variation in program effects across individuals, it applies equally well to variation in program effects across sites. This is because different sites might serve different types of clients, operate in different contexts, and have different treatment contrasts.

Cross-site variation in program effects produced by the three C's has important implications for policy and practice. For example, a program with effects that vary substantially and unpredictably can be a risky option for local decision-makers, even if its average effects are positive. In addition, variation in program effects can have important consequences for equity or fairness. For example, a remedial reading curriculum intended for struggling readers might have positive effects on initially low-achieving students and little effect on initially high-achieving students, thereby reducing disparities in reading achievement. Furthermore, variation in program effects, if predictable, can allow for the targeting of resources towards those who are most likely to benefit. Moreover, natural variation in treatment contrasts, variation in treatment effects on intermediate outcomes, and variation in treatment effects on important longer-term outcomes can provide opportunities to learn about the mechanisms through which programs work (Bloom, Hill, & Riccio, 2003; Reardon & Raudenbush, 2013).

Finally, the extent to which program effects vary is related to the generalizability of an evaluation study's results (Cole & Stuart, 2010; Tipton, 2013a, 2013b), which typically reflect the observed experience of a convenience sample of sites (Allcott, 2015; Olsen, Orr, Bell, & Stuart, 2013; Stuart, Cole, Bradshaw, & Leaf, 2011). If program effects vary substantially and unpredictably across sites, a given convenience sample might not yield results that are broadly applicable to one's inference population. Conversely, if program effects are homogenous, a given convenience sample could yield results that are highly relevant to the inference population.

In addition to the above *substantive* consequences of program effect variation for policy and practice, this variation has important *methodological* consequences for research. As discussed below, the variance of a site-level program effect distribution influences the statistical power and precision of a study, and therefore must be considered when designing and planning a study (Bloom & Spybrook, under review; Dong & Maynard, 2013; Hedges & Pigott, 2001; Raudenbush & Liu, 2000). The more program effects vary across sites, the larger a study must be in order to attain a given level of statistical power, all else being equal.

However, there is little empirical information about the magnitude of cross-site impact variation for different types of interventions, student populations, or counterfactual conditions. This represents a major information void for the design of multisite evaluations. The goal of the present paper is to provide a first step toward filling this void. We do so by estimating the cross-site standard deviation of site mean intent-to-treat effects using client-level data from 15 past multisite RCTs of educational and training programs. These estimates provide an empirical starting point for anticipating how much program effects vary across sites when designing multi-site RCTs. In addition, we explore hypotheses about when we might expect treatment effects to vary across sites a little or a lot – hypotheses that may further guide the design and planning of multi-site RCTs.

Section 2 of the paper, which follows below, describes the data sources and methodology we used to estimate the cross-site standard deviation of site mean program effects for a range of interventions and study populations. Section 3 describes the results of these analyses. Section 4 considers the implications of the results for planning future multisite RCTs and offers preliminary hypotheses about when to expect treatment effects to vary across sites by a lot or a little.

Data and Method

Data Sources

The analyses that follow are based on data from a convenience sample of 15 multisite RCTs that were chosen based on several criteria. The core selection criterion was that each study be able to provide internally valid estimates of average program effects by site. Datasets were then restricted to multisite trials that randomized *individuals* within sites, for as many sites as possible, with as many individuals per site as possible. This was done to maximize the precision of estimates of our parameter of interest, the cross-site standard deviation of site mean program effects.[1] Additional criteria for selecting datasets included how readily available to and well understood by our project team they were, and how well our final collection of datasets represented a broad spectrum of educational levels from pre-school to primary school to secondary school to post-secondary education and job training programs. We obtained the individual-level data from all studies and re-analyzed the data as described in later sections.

Appendix A briefly summarizes the studies used for the present analysis. Each study summary describes the program and target population examined, the research design and primary outcome measures we analyzed, and the number of sites and sample members involved. Citations for each study are included to help readers learn more about them.

Table 1 summarizes important design features for each study. It includes: the total sample size (N), the number of sites (J), the number of random assignment blocks (K), and the proportion of sample members assigned to treatment ($\bar{T}$). Across studies the total number of sample members ranges from about 1,400 to 69,000 persons; the total number of sites ranges

[1] Consequently we excluded: (1) cluster-randomized trials (CRTs) where classrooms were randomized because of their typically limited site-level precision, (2) CRTs where whole schools were randomized because of their inability to estimate average effects for individual schools, (3) regression discontinuity designs (RDD) because of their typically limited site-level precision, and (4) cluster-level RDDs because of their doubly-limited site-level precision.

from 9 to almost 300. The treatment and control groups are roughly evenly split, with the percentage of sample members randomized to the treatment group ranging from about 40 to 62 percent.

[INSERT Table 1 AROUND HERE]

Estimands and Estimation Model

Our analyses focus on two parameters of a cross-site distribution of site mean intent-to-treat effects (B_j) – the mean (β) and standard deviation (τ). By definition:[2]

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

and

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

We use the following 2-level hierarchical linear model to estimate β and τ:

Level 1: Sample Members

$$Y_{ij} = \sum_{r=1}^{R} \alpha_r \, RA\_Block_{rij} + B_j T_{ij} + \sum_{l=1}^{L} \gamma_l X_{lij} + e_{ij} \qquad (3)$$

Level 2: Sites

$$B_j = \beta + b_j \qquad (4)$$

where:

$$e_{ij} \sim N\left(0, \sigma^2_{|\boldsymbol{X}\alpha_r}(T_{ij})\right)$$
$$b_j \sim N(0, \tau^2)$$
$$Cov(e_{ij}, b_j) = 0$$

In this model, Y_{ij} is the value of the outcome measure for individual i in site j, RA_Block_{rij} equals one if individual i in site j belongs to random assignment block r and zero

[2] See Raudenbush and Bloom (2015) and Raudenbush (2015) for discussions of related estimands.

otherwise, T_{ij} equals one if individual i in site j was assigned to treatment and zero otherwise. We also include baseline covariates, X_{lij} to improve the precision of parameter estimates.

The model has a set of fixed random assignment block intercepts (α_r),[3] which account for the fact that individuals were randomly assigned within blocks and that the proportion of sample members randomized to treatment can vary across blocks.[4] The model allows for site-specific program-effect coefficients (B_j) that can vary randomly across sites. The B_j's are modeled as representing a cross-site population distribution with a mean value of β and a standard deviation of τ. Hence, the site-level random error term, b_j has a mean of zero and a standard deviation of τ. Finally, the model allows for the variability of level-1 residuals to differ by treatment group. The individual-level random error term, e_{ij}, is assumed to have a mean of zero and a variance of $\sigma^2_{|\boldsymbol{X}\alpha_r}(T_{ij})$, which can be different for treatment group members and control group members.[5]

Analytic Decisions

The following analytic decisions were made when applying our estimation model.

Defining "Sites"

We define a site to be the physical location where the program operated. This makes it possible to define sites consistently across studies and provides an intuitively appealing site

[3] An alternative way to fit this model is to exclude fixed random assignment block indicators and instead to group-mean center the outcome, treatment indicator, and the other covariates around their random assignment block means (see Raudenbush (2009) for more information on this type of centering). We found that in some cases, such as Teach for America—pooled analysis, the alternative model was more stable.

[4] For some studies, random assignment blocks were defined as the sites of interest (e.g., schools) and thus $K = J$. For other studies, blocks were nested within sites (e.g., when random assignment was conducted separately for multiple grades and/or student cohorts from a school) and thus $K > J$.

[5] The "$|\boldsymbol{X}\alpha_r$" is used to distinguish the residual variance (after accounting for covariates and the random assignment block indicators), referred to here, from the total outcome variance, discussed later. Bloom, Raudenbush, Weiss, and Porter (revise and resubmit) provide further information about this model and Raudenbush and Bloom (2015) and Raudenbush (2015) explore its properties. Appendix B provides SAS code for implementing this estimation model.

definition. For example, the Career Academies study was conducted in nine high schools across the U.S, with each high school operating its own Career Academy. Thus each high school is a site in our analysis. Notably, random assignment to Career Academies was conducted separately for up to three student cohorts per high school. Thus each cohort for a given high school is a separate random assignment block. Consequently, our estimation model for Career Academies contained fixed intercepts for 20 random assignment blocks and randomly varying impact coefficients for 9 sites.

Choosing Covariates

Our estimation model includes individual-level baseline covariates to improve the precision of our parameter estimates by predicting some of the observed variation in outcomes across individual sample members (Bloom, Richburg-Hayes, & Black, 2007). For most studies, we ran our estimation model including all covariates used by the original researchers plus any additional covariates that we thought may improve the model's predictive ability. For pre-school, elementary, and secondary education programs, we usually were able to include measures of: a pretest score, race/ethnicity, gender, parental educational attainment, socio-economic status, English language status, special education status, and whether a student was overage for grade. For post-secondary education programs, we usually were able to include measures of: prior educational attainment, the presence of dependents, and whether a sample member was the first in his family to attend college. For job-training programs we usually were able to include measures of: earnings at baseline, prior educational attainment, the presence of dependents, and household structure.

Handling Missing Data

For each baseline covariate in Equation 3, a binary indicator was added to the model to indicate sample members with missing data for the covariate. The value of each covariate was then set to zero for all sample members with missing data for that covariate (Gerber & Green, 2012).[6] This approach prevents the loss of observations due to missing covariate data. In some cases the binary missing-data indicators can improve the predictive power of the model and thus the precision of its parameter estimates. Puma, Olsen, Bell, and Price (2009) demonstrate that this approach does not introduce bias for randomized trials.[7] Observations with missing *outcome* data were not included in the analysis – this is consistent with the procedures used by nearly all of the original researchers.

Outcome Metrics and Reporting

For seven of the programs we examine (Early College High Schools, Small Schools of Choice, Career Academies, the two community college programs, and the two workforce development programs), all outcome measures have meaningful "natural" units (e.g., degree completion or not, number of credits earned, or annual earnings). We denote such outcome measures as Y and present findings for them in their natural units. For example, estimates of cross-site mean program effects on earnings in their natural units, β_Y, and the corresponding cross-site standard deviation of earnings effects in natural units, τ_Y, are reported in dollars.

The other studies include outcomes on metrics without natural meaning (e.g., test scaled scores). To maximize interpretability and facilitate cross-study comparisons, we z-score these

[6] There are two exceptions: For the Welfare to Work study, baseline covariates used in the estimation model were imputed by the original study team and non-imputed data were not available. For the national Head Start Impact Study, two baseline covariates (mother's age and teen mom) used in the estimation model were imputed by the original study team and non-imputed data were not available.

[7] Although it can introduced bias for observational studies (Jones, 1996).

outcomes, denoting them as Z, and present their findings in *standardized effect size units*. An individual's z-score is calculated as:

$$Z = \frac{Y_i - \bar{Y}}{\sigma} \qquad (5)$$

where Y_i is the outcome for individual i, $\bar{Y}$ is the average outcome in its natural units across a reference group of individuals (described later), and σ is the standard deviation of the outcome in its natural units across a reference group of individuals (described later).

Findings for z-scores are presented in standardized effect size units, which are defined as a multiple of an individual-level standard deviation (σ) for the original outcome measure involved. For example, an estimate of a cross-site mean program effect size, β_Z, might be reported as 0.20σ. This implies that the average program site increased the mean outcome by a magnitude that equals 0.20 individual-level standard deviations of the outcome measure in its original units. A corresponding estimate of the cross-site standard deviation of site mean program effect sizes, τ_Z, might be reported as 0.15σ. This implies that the magnitude of the cross-site standard deviation of site mean program effect size equals 0.15 individual-level standard deviations of the outcome measure in its original units.

To construct z-scores using Equation 5 (and ultimately, to estimate corresponding standardized parameters), one must choose an individual-level mean ($\bar{Y}$) and standard deviation (σ) of the outcome measure in its natural units for a specific *reference group* – that is, across which individuals are $\bar{Y}$ and σ calculated? This decision – which arises whenever a study finding is reported as a standardized effect size – is especially important for cross-study comparisons like the present analysis. We consider two common reference groups: (1) the nation or state, and (2) the study's control group.[8]

To make the units of effect size measures from the different studies in our analyses as comparable as possible, our preferred reference group is the nation (for outcomes that are measured as scores on a nationally-normed test) or for a given state or states (for outcomes that are measured as scores on a statewide standardized test). We refer to this broad reference group as a *reference population*. An advantage of defining an effect size in this way is that the interpretation of findings on this metric are *not* tethered to the particular individuals in a study sample, as is the case when the reference group is the study's control group. Consequently, when this approach is applied consistently across studies, it facilitates cross-study comparisons. Furthermore, by basing the effect size on the amount of individual outcome variation that exists for a relevant reference population of interest, it grounds the meaning of resulting findings in a broad-based (national or state), meaningful outcome variation.

Another common approach defines standardized effect sizes with respect to the standard deviation of the outcome measure (in its natural units) for the study's *control group*. This control-group-based approach to standardizing outcome measures (and their related parameters) is very convenient because it can be used for any outcome measure, including those that have not been normed for a relevant population. However, because a control-group based z-score and its related parameters are defined relative to the heterogeneity of a particular study sample, the meaning of such measures and parameters can differ across studies, thereby complicating cross-

[8] In both cases, for multi-grade studies, we calculate effect sizes in reference to *within* grade variability only – that is, z-scoring is done separately within grade. In addition, when the study's control group is the reference group, z-scores are calculated separately within metric. For example, if different assessments were administered to different clients, z-scores are calculated separately for clients taking different assessments. This frequently occurs in multi-state studies when state-administered tests are the outcome.

study comparisons.[9] Nonetheless, because it is common practice (and often the only option) to use control-group-based standardization for scaling program effect sizes, we also report our findings in this metric. Table 2 provides a glossary of the notation described in this section and used in the results section.

[INSERT Table 2 AROUND HERE]

Results

Table 3 presents the key results of our analyses. These findings are reported either in their natural units (Y) for outcomes with meaningful natural units or in reference population-based standardized units (Z) for outcomes without meaningful natural units.[10] The first column in the table reports the unadjusted mean control group value of each outcome measure from each study. The second column reports the estimated cross-site mean treatment effect (β_Y) on outcomes measured in their natural units or their counterpart (β_Z) for outcomes measured in reference population-based standardized units. These results are generally consistent with their counterparts from the original studies. Columns three and four report estimated standard errors and p-values for our cross-site mean impact estimates. Column five reports estimates of cross-site standard deviations of program effects, in natural units (τ_Y) when these units are meaningful or in reference population-based effect-size units (τ_Z) when natural units are not meaningful.

[9] This same problem exists—we expect to a lesser extent—with estimates from studies that standardize on reference populations from different states; or comparisons of effects from a state-normed estimate and effects from a national-normed estimate. For example, variation in outcomes for students in New Hampshire differs from variation in outcomes in New Mexico. However, using the 2013 National Assessment of Educational Progress (NAEP) we find that the national standard deviation of student test scores across public schools is very similar to the average within-state standard deviation of test scores. Thus, we do not believe that the standard deviation of outcomes nationally differs dramatically from the standard deviation of outcomes within states (on average). However, standard deviations of test scores seem to vary considerably across states, thus there is ambiguity regarding how much of an issue this presents.

[10] Two exceptions are the Head Start Impact Study (HSIS) and Tennessee STAR. For HSIS we were able to obtain the mean and standard deviation from a national norming sample for the Woodcock-Johnson measures, but not for the PPVT or the Externalizing Behavior and Self-Regulation scales – thus they are not included in this table. For Tennessee STAR, we were not able to obtain information on a relevant reference population, thus we do not present estimates in this table. In both cases results are presented in control-group-based standardized units in Table 4.

Lastly, column six reports estimates of the p-values for our estimates of τ_Y or τ_Z.[11] Studies are listed in the table according to the developmental stage of the intervention they represent, from early childhood education to primary and secondary education to postsecondary education and adult workforce development.

[Insert Table 3 around here]

Findings in the table reflect a broad range of different cross-site impact distributions. Below we consider several illustrative examples:

Consistent zero average impact across sites. The After School Reading Program has an estimated cross-site mean reading-achievement effect size near zero ($\hat{\beta}_Z$ = –0.02σ) and a small estimated cross-site effect-size standard deviation ($\hat{\tau}_Z$ = 0.04σ). This suggests that most program sites had little effect on reading achievement relative to "business as usual" in the same after-school center. Figure 1 graphically illustrates this situation by summarizing the cross-site distribution of constrained empirical Bayes effect-size estimates.[12]

[INSERT Figure 1 AROUND HERE – AFTER SCHOOL]

Near zero average impact with a lot of cross-site variation. Charter middle schools also have an estimated cross-site mean reading achievement effect size ($\hat{\beta}_Z$) that is near zero (–0.02σ for the first follow-up year and –0.07σ for the second follow-up year), relative to their counterfactual traditional public schools. However, the estimated cross-site effect-size standard deviation for charter middle schools is substantial ($\hat{\tau}_Z$ = 0.15σ for the first follow-up year and $\hat{\tau}_Z$ = 0.16σ for the second follow-up year). Hence, the near zero mean effect of these charter schools masks substantial cross-site variation, which is illustrated by Figure 2.

[11] P-values were obtained from estimates of a Q-statistic described in Appendix C.

[12] These estimates are constrained to ensure a cross-site variance equal to $\hat{\tau}_Z$ (see Bloom et al. (revise and resubmit)). This constraint adjusts for the fact that conventional empirical Bayes estimates tend to *understate* true variability across estimates (Raudenbush & Bryk, 2002).

[INSERT Figure 2 AROUND HERE – CHARTER]

Consistent positive impacts across sites. A different scenario is reflected by the results for Career Academies. For example, consider the effects of this intervention on future earnings. During the first four follow-up years the cross-site mean earnings effect of assignment to a Career Academy ($\hat{\beta}_Y$) was an increase of $1,883 per year (11 percent of control group earnings) and during the second four follow-up years this mean effect was $2,313 per year (8 percent of control group earnings). Interestingly, there is not evidence of statistically discernable cross-site variation in these effects, which suggests that Career Academies produced consistently large positive effects on future earnings.

Large average impacts with a lot of cross-site variation. Yet another scenario was observed for New York City's Small High Schools of Choice (SSCs), the national Job Corps Program, and the welfare to work (WtW) programs. Each of these interventions appeared to produce a large positive cross-site mean effect *and* have substantial cross-site variation in effects. This suggests that a majority of sites for these interventions produced positive effects with widely varying magnitudes and a minority of sites produced negative effects with modestly varying magnitudes.

For example, random assignment to the average SSC (versus some other New York City public high school) increased the percentage of students who were on track toward graduation at the end of their ninth grade (i.e., first year of high school) by 10.3 percentage points and increased four-year (i.e., on-time) high school graduation rates by 6.7 percentage points. The cross-site standard deviations of these effects are 15.3 and 11.5 percentage points, respectively. Assuming approximate normality, this implies that roughly two thirds of New York City's SSCs produced "ninth grade on-track effects" that ranged from 26 percentage-points *higher* than their counterfactual schools to 5 percentage points *lower*. The corresponding range for high school graduation rates was positive 18 percentage points to negative 5 percentage points.[13]

Table 4 provides another point of reference for future researchers by reporting parameter estimates for each of the 15 multisite RCTs in control-group-based effect-size units. This information is directly applicable for planning studies that will use control-group-based z-scores as their outcome measures. In all cases where comparison is possible, $\hat{\beta}_Z$ in control-group-based z-scores and reference-population-based z-scores is of similar magnitude and statistical significance. The same is true with respect to $\hat{\tau}_Z$ for the Head Start Impact Study, the two After School programs, Enhanced Reading Opportunities, and Teach for America-pooled analysis. The magnitude of $\hat{\tau}_Z$ is slightly larger in control-group-based z-scores compared to reference-population-based z-scores in the Charter Middle School study, and Teach for America: Math, although the statistical significance is consistent.[14]

Discussion

In this final section of the paper we consider how the preceding findings can inform the design of future multisite trials. We first illustrate how to use the findings to assess the statistical precision of multisite sample designs. We next pose preliminary hypotheses about when to expect a lot or a little cross-site impact variation. These findings and hypotheses can help researchers make educated guesses about the magnitude of cross-site impact variation to expect for future trials. We conclude by considering some limitations of the present research.

[13] We report our findings for SSCs in terms of the effect of intent to treat (ITT) in order to be consistent with the findings we report for all other studies. Hence, the magnitudes of present SSC findings are smaller than those reported by the original authors (e.g. Bloom and Unterman (2014)) which are in terms of local average treatment effects.

[14] Notably, in HSIS, After School Reading and Math, and ERO, four of the studies where the reference-population-based $\hat{\tau}_Z$ is very similar to the control-group-based $\hat{\tau}_Z$, all students took the same study administered assessments. In contrast, in Charters, TFA-Math, and TFA-pooled, students located in sites in different states took different state-administered tests. The resulting z-scores may contribute to the observed differences in $\hat{\tau}_Z$ for Charters and TFA-Math for the two z-score reference groups.

Using the Present Findings to Assess the Statistical Precision of a Planned Multi-Site RCT

The design of a planned multi-site RCT typically involves an educated estimate of the statistical precision of the design under assumptions about factors like the number of sites, number of sample members per site, how much the treatment effect will vary across sites, etc. Power or minimum detectable effect (MDE) calculations are used to determine sample size requirements to ensure a study will be well-positioned to detect meaningful program effects. By definition, the MDE is the smallest true mean effect that a study design can detect at a specified level of statistical significance (typically 0.05 for a two-tailed test) with a specified level of statistical power (typically 80 percent). Dong and Maynard (2013) and Bloom and Spybrook (under review) provide formulas to calculate the minimum detectable *effect*, MDE_Y, for an outcome, Y, that is measured in its natural units or a minimum detectable *effect size*, $MDES_Z$, for a standardized outcome, Z, that is measured in standard deviation units.[15]

Some Illustrative Examples

This section illustrates how to use the present empirical findings to help assess the MDE_Y or $MDES_Z$ of alternative multisite sample designs when planning an evaluation. We recommend presenting the MDE and findings for outcomes in their natural units (e.g., credits earned, degree completion, or earnings) whenever these units are meaningful. Therefore, we begin the discussion by focusing on the corresponding formula:[16]

$$MDE_Y = M_{j-1}\sqrt{\left(\frac{1}{J}\right)\left(\tau_Y^2 + \frac{(1-\rho)(1-R_{within}^2)\sigma_Y^2}{n\bar{T}(1-\bar{T})}\right)} \qquad (6)$$

[15] Raudenbush and Liu (2000) demonstrate how to determine the statistical power of impact estimates from multisite trials.

[16] This expression, which is a good approximation for many situations (see Bloom and Spybrook, under review), assumes, for simplicity, equal number (n) of sample members at every site, and equal proportion $\bar{T}$ of the sample members at every site randomized to the treatment group, and that the individual-level residual outcome variance ($(1-\rho)(1-R_{within}^2)\sigma_Y^2$) is the same for all sites and the same for treatment and control group members.

where τ_Y is the cross-site standard deviation of site-average program effects in the natural units of the outcome measure (Y), σ_Y is the individual-level *control-group* standard deviation of the outcome in its natural units,[17] ρ is the proportion of the total outcome variance (σ_Y^2) that is *between* random assignment blocks (i.e., the control group intra-class correlation),[18] R_{within}^2 is the proportion of within-random assignment block outcome variation for control group members that is explained by baseline covariates (X), and M_{J-1} is a multiplier that approaches 2.8 (for a two-tail test at 0.05 significance and 80 percent power) as J increases.[19]

Equation 6 illustrates that, other things being equal, MDE_Y increases directly with the amount of cross-site impact variation (τ_Y) that exists and with the amount of individual-level residual outcome variation that exists $((1-\rho)(1-R_{within}^2)\sigma_Y^2)$. In words, studies evaluating interventions that are expected to have higher variation in impacts across sites or higher levels of individual-level residual outcome variation (for example, because they have fewer strong baseline predictors of the outcome) will have lower statistical power to detect an average impact of the intervention. Equation 6 also indicates how J, n and $\bar{T}$ influence MDE_Y.

Findings in Tables 3 and A.1 (augmented for some applications by parameter estimates from existing sources described later) can be used with Equation 6 to help assess the minimum detectable effect of a proposed multisite RCT. For example, consider a design for testing the effect of an intervention like a performance-based scholarship on total credits earned by community college students during their first year of the scholarship offer. Assume that you want

[17] This represents the *total* control-group variation in the outcome measure within and between sites.

[18] The term ρ represents the proportion of σ_Y^2 that is accounted for by the random assignment block indicators in our estimation model. As noted earlier, random assignment blocks are sometimes sites, in which case this is the intraclass correlation. However, in many cases random assignment blocks are nested within sites and are thus defined as unique site by grade combinations or site by cohort combinations.

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

to assess the minimum detectable effect of a multi-site RCT with 25 sites (J), 75 students per site (n) and 0.6 of the students at each site randomized to treatment ($\bar{T}$) in a single random assignment block per site. In order to "guesstimate" a minimum detectable effect (MDE_Y) for the intervention, outcome, and sample design, you must assume values for the parameters on the right-hand side of Equation 6. You should then test the sensitivity of your MDE_Y to your assumptions.

A good starting place is the findings in Tables 3 and A.1 for the performance-based scholarship evaluation. Findings in Table A.1 indicate that the individual-level control group standard deviation of the outcome measure (σ_Y) equals 10.7 credits, the intraclass correlation for the outcome (ρ) equals 0.09 and the within-block R-square of your covariates for control group members (R^2_{within}) equals 0.02. In addition, and most central to the present paper, findings in Table 3 indicate that the estimated cross-site standard deviation of the impacts of performance based scholarships on credits earned during students' first year (τ_Y) equals 0.8 credits. Together the preceding parameter values imply a minimum detectable effect (MDE_Y) of 1.5 credits. Of course, the intervention and the nature of the sample for your proposed study might differ in important ways from that for the performance-based scholarship evaluation. Thus, you should carefully consider these differences and modify your assumed parameter values accordingly. In many cases this will require a great deal of subjective judgment.

Equation 6 can also be used to assess the MDE_Y for a binary outcome with impacts expressed in percentage points. For example, findings in Tables 3 and A.1 for the effects of Early College High Schools (ECHSs) on the percentage of incoming ninth graders who are "on track to graduate" at the end of ninth grade (σ_Y = 31.7 percentage points, ρ = 0.11, R^2_{within} = 0.26 and τ_Y = 8.2 percentage points) suggest that MDE_Y would be 6.0 percentage points for our illustrative sample design (with J = 25, n = 75 and $\bar{T}$ = 0.6). Of course, one should only consider findings for ECHSs as a starting point for assessing precision for a related intervention, outcome and sample structure.

Equation 7 below indicates how to compute a minimum detectable effect *size*.

$$MDES_Z = M_{J-1}\sqrt{\left(\frac{1}{J}\right)\left(\tau_Z^2 + \frac{(1-\rho)(1-R_{within}^2)\sigma_Z^2}{n\bar{T}(1-\bar{T})}\right)} \qquad (7)$$

Equations 6 and 7 are very similar, the key difference is that in equation 6 parameters τ^2 and σ^2 are in natural units (Y), whereas in equation 7 they are in z-score units (Z). To be clear, σ_Z is the expected individual-level standard deviation of the outcome for the study's control group on the z-score metric and τ_Z is the anticipated cross-site standard deviation of site-average program effects on the z-score metric. In the simple case, $\sigma_Z^2 = \frac{\sigma_Y^2}{\sigma_{Y(RG)}^2}$ and $\tau_Z^2 = \frac{\tau_Y^2}{\sigma_{Y(RG)}^2}$, where σ_Y^2 and τ_Y^2 are defined as in equation 6 and $\sigma_{Y(RG)}^2$ is the individual-level variance of the outcome in natural units across the relevant *reference group* (e.g., a broad reference population or the study's control group).[20] As before, we must distinguish between effects sizes defined relative to two reasonable reference groups – a broad reference population (e.g., the state or nation) or the study's control group.

We begin with an example of an $MDES_Z$ for the reference-population-based effect size, using findings from Tables 3 and A.1 to assess the precision of our illustrative multi-site trial (with J = 25, n = 75 and $\bar{T} = 0.6$). Assume that we want to design a study of the effects of a high-school reading intervention on student scores on a nationally-standardized test of reading achievement. For this purpose, we might want to *combine* information from Table 3 on cross-site

[20] Exceptions to the simple case arise due to the way z-scores are constructed in practice, when multiple metrics are used to measure outcomes.

variation in reference-population-based effect sizes (τ_Z) with information about intraclass correlations (ρ) and within block R-squares (R^2_{within}) from the extant literature on design parameters for cluster randomized trials (e.g. see (Bloom, Bos, & Lee, 1999; Bloom et al., 2007; Bloom et al., 2008; Hedges & Hedberg, 2014; Jacob, Zhu, & Bloom, 2009; Westine, Spybrook, & Taylor, 2014; Xu & Nichols, 2010).

For high school standardized achievement test scores, the school-level intra-class correlation reported by this literature tends to range from about 0.10 to 0.20; so assume a starting value of 0.15 for ρ. The corresponding individual-level R-square for covariates, including a pretest, tend to range from about 0.3 to 0.6; so assume a starting value of 0.45 for R^2_{within}. From findings in Table 3 for the Enhanced Reading Opportunities (ERO) program note that the estimated cross-site reference-population-based effect size standard deviation (τ_Z) equals 0.08σ for the GRADE test of reading comprehension and zero for the GRADE test of reading vocabulary. Thus as a starting point, assume that τ_Z = 0.04. Lastly note, that ERO findings indicate that $\sigma_Z = \frac{\sigma^2_Y}{\sigma^2_{Y(RP)}} = 0.68$ for both versions of the GRADE reading test. Thus assume this value. Given Equation 7, the preceding assumed parameter values imply that $MDES_Z$ = 0.07σ.

Equation 7 can also be used to compute the minimum detectable effect size when, as is common practice, researchers use a standardized outcome measure that is based on the control-group standard deviation of the original outcome measure. As noted earlier, this metric depends on the heterogeneity of a study’s sample, which can vary markedly across studies, thus it can be difficult to compare such findings across studies. Nonetheless, this approach is often necessary because the outcome of interest does not have meaningful natural units and it is not possible to standardize the outcome measure based on data for a relevant reference population. Conveniently, when the $MDES_Z$ is in control-group-based effect size units, σ^2_Z (the individual-level control-group variance of the outcome on the z-score metric) has value of one by construction, and thus the formula simplifies.

Now consider how to use Equation 7 with findings from Tables 4 and A.1 to assess the statistical precision of our illustrative multi-site sample design (with J = 25, n = 75 and $\bar{T}$= 0.6) for a control-group-based standardized outcome measure (Z). For example, our findings for the class-size effects of Tennessee STAR on students' scores on the SAT-7 total math test ρ = 0.21, R^2_{within} = 0.03, and τ_Z = 0.26σ suggest that $MDES_Z$ = 0.19σ. We could also base this determination on values for ρ and R^2_{within} reported in the literature on cluster randomized trials noted earlier.[21]

The Influence of τ on MDES

To conclude this part of the discussion, Figure 3 illustrates how cross-site impact variation influences the precision – and thus the sample size needs – of a multi-site impact study. Based on Equation 7, the figure plots $MDES_Z$ as a function of J, which ranges from 5 to 70, and for two values of τ_Z, setting, ρ = 0.20, R^2_{within} = .30 , σ^2_Z = 1, n = 75 and $\bar{T}$ = 0.5. The two values for τ_Z are 0.04σ, which reflects cross-site variation in the reference-population-based effect sizes of the After School Reading Program and 0.16σ, which reflects cross-site variation in the year two reading reference-population-based effect sizes of Charter Middle Schools.

[INSERT Figure 3 AROUND HERE]

[21] A special issue arises when guesstimating ρ for calculating the $MDES_Z$ in control group-based effect size units in a study where the outcomes for clients in different sites are on different metrics. This can occur in multi-state RCTs, such as the Charter Middle School evaluation, that rely on administrative records (e.g., state tests) that are not the same metric across sites. In such cases, to create a single standardized outcome measure it is necessary to z-score the outcome data within clients whose outcome is on the same metric – for example, within state. Doing so artificially reduces the intraclass correlation (ρ) by eliminating any outcome variation between clients whose outcomes were measured on a different metric. In such instances, it may be prudent to assume ρ = 0 for MDES calculations.

Figure 3 demonstrates the marked effect that τ_Z has on the number of sites required to attain a given $MDES_Z$ in a multisite trial, and thus the importance of accurately guesstimating τ_Z when designing a study. Notice that the $MDES_Z$ is always larger for the larger value of τ_Z. This illustrates that the more effects vary across sites, the larger is the $MDES_Z$. The top and bottom red horizontal lines mark the number of sites required to design a study to detect an effect size of 0.20σ and 0.10σ, respectively. Given the assumptions above, if τ_Z = 0.04σ only eight sites (N = 600) are needed for an $MDES_Z$ of $.20\sigma$. In contrast, if τ_Z = 0.16σ a total of 13 sites (N = 975) are needed for an $MDES_Z$ of 0.20σ – a 63 percent increase in required number of sites. For an $MDES_Z$ of 0.10σ, the number of sites jumps to 27 and 46 (N = 2,025 and 3,450), for τ_Z of 0.04σ and 0.16σ, respectively – a 70 percent increase in sites needed to detect the same effect size for this fourfold increase in τ_Z.

When to Expect a Lot or a Little Cross-site Impact Variation

Estimates of τ in Tables 3 and 4 provide a starting point for a deeper understanding of how the magnitude of cross-site impact variation is related to the educational intervention, student population, and/or counterfactual educational options involved. Thus, as a first step toward building a theory of cross-site impact variation, we use information from the tables to *explore* some hypotheses about when to expect a lot or a little variation. Because our hypotheses, and the nascent theory they reflect, were developed both before and after examining the present findings, our discussion represents a mix of ex ante and post hoc speculations. Therefore, rather than trying to test hypotheses against the combined weight of the present evidence, we simply illustrate each hypothesis with a specific empirical example.

The central focus of our hypotheses about the magnitude of cross-site program effect variation is on cross-site variation in the *treatment contrast.* The treatment contrast is the *difference* between the services and activities experienced with and without access to the program that are expected to affect the outcome of interest. Focusing on this difference emphasizes the often-ignored fact that program effects are defined *relative* to some other counterfactual condition; or in the words of Paul Holland (1986), "the effect of a cause is *always* relative to another cause."

Thus when as is often the case, evaluators state their conclusions in terms like "program X works well" or "program Y works poorly," the inherently *relative* nature of program effects tends to get lost. The main point here is that cross-site variation in program effects can reflect cross-site variation *both* in services and activities experienced by treatment group members and in services and activities experienced by control group members. We illustrate this point – which cannot be over-emphasized – in numerous ways.

A treatment contrast can have many *components*. For example, in the performance-based scholarship evaluations noted earlier, some sites had two core components: (1) a financial incentive for meeting student performance benchmarks and (2) student advising. Moreover, each program component can be described by its *features*. For example, student advising can be thought of in terms of its content (*What* issues were addressed?), its quantity (*How much* advising took place?), and its conveyance (Through *what* medium and by *whom* was the advising delivered?).[22] Cross-site variation in components and/or features of components of the treatment contrast can produce corresponding variation in its effects relative to counterfactual conditions.

To explore the implications of this point, we introduce the following notation. Let $\overline{TC}_{cfj}$ be the average treatment contrast for component c (e.g., advising), along feature f (the content, quantity, or conveyance of component c), at site j. By definition, $\overline{TC}_{cfj}$ is site j's difference

[22] See Weiss et al. (2014) for a discussion of these factors.

between: (1) mean services and activities experienced by treatment group members, $\bar{S}_{cfj|T=1}$, and (2) mean counterfactual services and activities experienced by control group members, $\bar{S}_{cfj|T=0}$. That is,

$$\overline{TC}_{cfj} \equiv \bar{S}_{cfj|T=1} - \bar{S}_{cfj|T=0} \qquad (8)$$

We denote the cross-site variance of $\overline{TC}_{cfj}$ as $Var(\overline{TC}_{cf})$, which can be expressed as:

$$Var(\overline{TC}_{cf}) = Var(\bar{S}_{cf|T=1}) + Var(\bar{S}_{cf|T=0}) - 2Cov(\bar{S}_{cf|T=1}, \bar{S}_{cf|T=0}) \qquad (9)$$

We hypothesize that other things being equal, when $Var(\overline{TC}_{cf})$ is large for key components or features that drive educational production (or some other outcome of interest), cross-site impact variation (τ) will tend to be large. For this discussion we have expanded what is contained within the definition of “services and activities” (and therefore the treatment contrast) to include anything that individual’s experience that might reasonably influence the outcomes of interest, rather than just the specific services that comprise the intervention itself.

To help unpack this hypothesis, Equation 9 suggests that we focus on three factors that influence $Var(\overline{TC}_{cf})$: (1) cross-site variation in services and activities experienced by treatment group members that influence their outcomes, $Var(\bar{S}_{cf|T=1})$, (2) cross-site variation in counterfactual services and activities experienced by control group members that influence their outcomes, $Var(\bar{S}_{cf|T=0})$ and (3) the covariance between these two potential service and activity experiences, $Cov(\bar{S}_{cf|T=1}, \bar{S}_{cf|T=0})$. Therefore woven throughout the following discussion are specific hypotheses about when to expect $Var(\bar{S}_{cf|T=1})$ to be large, when to expect $Var(\bar{S}_{cf|T=0})$ to be large, and when to expect $Cov(\bar{S}_{cf|T=1}, \bar{S}_{cf|T=0})$ to be close to zero (or negative). Other things being equal, each of these conditions tends to increase $Var(\overline{TC}_{cf})$.

(1) Cross-site Variation in Service Experiences when Assigned to the Program

All else equal, greater cross-site variation in services and activities experienced by treatment group members $Var(\bar{S}_{cf|T=1})$ produces greater cross-site variation in a program's treatment contrast, $Var(\overline{TC}_{cf})$. So, under what circumstances might we expect $Var(\bar{S}_{cf|T=1})$ to be large?

Low Specificity of the Intervention: One situation that can result in a lot of cross-site variation in the services and activities experienced by clients (a large value for $Var(\bar{S}_{cf|T=1})$) is when the program's planned services and activities are not clearly defined (i.e., they have low specificity). With little clear guidance about planned services and activities, it is likely that the services and activities experienced by clients will vary substantially across sites. Consequently, we might anticipate a lot of cross-site variation in program impacts.

Recall, for example, that estimated cross-site impact variation ($\hat{\tau}$) was substantial for charter middle schools. Charter schools are public schools that are operated under a contract with a school board (or another authorizer) that often releases the school from many of the state and district regulations on staffing, budgeting, and curriculum decisions, but holds it accountable for the quality of student outcomes (Gleason, Clark, Tuttle, & Dwoyer, 2010, p.1). The charter schools participating in the study were created and operated by different organizations with different educational philosophies and priorities. Furthermore, the schools did not share a common curriculum, pedagogy, teacher selection strategy, or course schedule. Without these commonalities it seems almost inevitable that $Var(\bar{S}_{cf|T=1})$ will be substantial, which in turn, can produce substantial $Var(\overline{TC}_{cf})$, and thus substantial cross-site impact variation.

In contrast, the After School Reading Program was a highly structured and carefully prescribed intervention that used an adaptation of the *Success for All* reading curriculum, which is known for its specificity.[23] Of course, like any program operated by service delivery agents with different backgrounds, priorities, and preferences (e.g. after-school teachers), implementation of the After School Reading program must have varied somewhat across sites. Nonetheless, the fact that program services were centrally planned and highly specific, that technical assistance was provided to increase implementation fidelity with the program model, and that implementation was closely monitored by an external organization, suggests that the services and activities experienced by treatment group members were probably fairly uniform across sites. This might be partly responsible for the small cross-site variation in the program's effects on student reading achievement that were observed.

(2) Cross-site Variation in Counterfactual Service Experiences

All else equal, greater cross-site variation in counterfactual service experiences, $Var(\bar{S}_{cf|T=0})$, should produce greater cross-site variation in a program's treatment contrast, $Var(\overline{TC}_{cf})$, which in turn, should produce greater cross-impact impact variation (τ). We expect $Var(\bar{S}_{cf|T=0})$ to be large in most situations because it represents unplanned variation in the educational or workforce development environment.

Our reasoning for this expectation is similar to that for expecting $Var(\bar{S}_{cf|T=1})$ to be large for low-specificity program models. In particular, we would expect that without organizational coordination across sites, *planned* counterfactual services – and thus *actual* counterfactual services – will reflect existing differences in the philosophies, priorities, preferences, past experiences, and comparative advantages of the organizations and individuals

[23] According to the program's evaluation, it was "a structured reading model with daily lessons that involve switching quickly from one teacher-led activity to the next. It… builds cooperative learning into its daily classroom routines, which also include reading a variety of selected books and frequent assessments built into lessons to monitor progress" (Black, Somers, Doolittle, Unterman, & Grossman, 2009, p.xxvi).

providing these services. Because such coordination is rare and difficult to manage effectively, we expect many interventions to have counterfactual services and activities that vary widely across sites. Thus, unless $Var(\bar{S}_{cf|T=0})$ covaries with $Var(\bar{S}_{cf|T=1})$, we expect this factor to be a major source of cross-site impact variation for many interventions. Unfortunately, this potentially major source of cross-site impact variation is often overlooked by evaluation studies, especially those with implementation analyses that focus almost exclusively on the program being tested (such as treatment fidelity analyses).

(3) The Critical Covariance

Equation 9 indicates that other things being equal, the cross-site variance of the treatment contrast, $Var(\overline{TC}_{cf})$, decreases as the cross-site *covariance* between its two elements, $(\bar{S}_{cf|T=1})$ and $(\bar{S}_{cf|T=0})$, increases. To understand this fundamental fact, consider a simplified example where $(\bar{S}_{cf|T=1})$ and $(\bar{S}_{cf|T=0})$ are *perfectly positively correlated* across sites and their cross-site variances are both equal to one. In this case, the cross-site covariance is equal to one and there is zero cross-site variation in the treatment contrast ($Var(\overline{TC}_{cf})$ = 0). This is the case because for every unit increase (or decrease) in $(\bar{S}_{cf|T=1})$ across sites there is a corresponding unit increase (or decrease) in $(\bar{S}_{cf|T=0})$. Consequently the *difference* between these two potential service experiences (the treatment contrast) is the same for all sites. So, what factors might predict $Cov(\bar{S}_{cf|T=1}, \bar{S}_{cf|T=0})$?

A High Proportion of Services and Activities are Altered by the Intervention: We hypothesize that the proportion of the services and activities that the intervention alters is related to the cross-site covariance, $Cov(\bar{S}_{cf|T=1}, \bar{S}_{cf|T=0})$.

Consider first an example where a low proportion of formal education (an important driver of education production) is altered by the intervention and thus the cross-site covariance, $Cov\left(\bar{S}_{cf|T=1}, \bar{S}_{cf|T=0}\right)$, is likely to be positive and large along many components and features that drive educational production. By design, the After School Math Program influenced a small portion of students' formal schooling. This intervention replaced the "business as usual" informal after-school experience of control group members with a structured math curriculum for treatment group members that was delivered through formal instruction. However, the after school program, as intended, had no influence on the roughly 162 hours per year of math instruction received by students during their regular school day (Black et al., 2009). Consequently, although the after school program increased average formal after-school math instruction from 11.4 hours per year for control group members to 59.8 hours per year for treatment group members, the resulting 48-hour annual increment was only 28 percent of the underlying 173.4 hour control group base (162 hours during the school day plus 11.4 hours during after school). As a result, cross-site variation in the increment was a small fraction of the control group base, which might help to explain why impacts varied little across sites for this intervention. Stated differently, during the entire regular school day program and control group students *within each site* had very similar types of experiences. Thus, despite the fact that program group members' regular school day experiences vary across sites and control group members' regular school day experiences vary across sites, $Var\left(\overline{TC}_{cf}\right)$ is small for most components and features of educational production because those experiences nearly perfectly covary.

In contrast, consider an intervention like New York City's small schools of choice (SSCs), which determines almost all of the high school experience of their students because those students attend an SSC throughout each school day, week, and year. SSCs were created and operated by organizations and individuals with different philosophies, priorities, and past experiences. The substantial time spent by students in an SSC and the lack of specificity of the intervention means a high proportion of experiences that drive education production may be distinctive relative to counterfactual experiences and among SSCs.

SSCs influence many components and features of students' educational experiences, such as the curriculum, instructional strategies, academic and personal guidance, and the peers to which they are exposed. To the extent that the quality of these components and features differ across SSCs, the extensive and extended student exposure to SSCs should result in large $Var\left(\bar{S}_{cf|T=1}\right)$. Moreover, while nearly all of the high school experiences of program group students are determined by their SSC, their control group counterpart's entire high school experience is determined outside the SSC. While it is theoretically possible that those experiences could covary with their counterfactuals, given the multi-dimensional nature of services and activities that students experience in any given school, it is almost certain that $Var\left(\overline{TC}_{cf}\right)$ must be large in many important ways. This might help to explain why SSC impacts on student academic attainment vary substantially.

We expect that the proportion of services altered by the intervention depends on whether treatment and control group members in a given site are served in the same or a different physical location. For example, programs like small schools of choice, charter middle schools, Head Start, early college high schools, welfare-to-work, and Job Corps serve treatment group members in settings (schools or centers) that usually are physically and organizationally separate from those which serve control group members. Hence, the forces that shape the program group and control group clients' experiences—such as policies and practices; teachers, counselors, or other service providers; and curricula, textbooks, or other materials—are likely to differ in ways that are not highly correlated across treatment and control locations within the site. Other things being equal, this would tend to increase cross-site variation in a program's treatment contrast and therefore increase cross-site variation in the program's impacts.

In contrast, programs like After School Reading, After School Math, Enhanced Reading Opportunities, and Performance-based Scholarships serve treatment and control group members in the same location. Hence they are more likely to have a strong positive covariance between the total educational services and activities experienced by treatment and control group members. In fact, with the exception of the hour or two per day that students experience the intervention (or its counterfactual services), the services and activities experienced during the remainder of the day may covary perfectly across sites. This suggests less cross-site variation in the treatment contrast (and impact) of educational programs that serve treatment and control group members in the same locations.

Between these two extremes are programs like Teach for America, Career Academies, Tennessee STAR, and Learning Communities, where treatment and control group members are served *in different classrooms within the same schools* (for the most part) throughout the school day. In theory, one might expect their covariances between treatment and control group services and activities to be somewhere between those of the preceding two cases. Other things being equal, this suggests that their cross-site variation in treatment contrasts and program effects might be between those of the preceding cases.

Variation in the Effectiveness of Individual Service Providers

One issue that cuts across all three factors which contribute to $Var\left(\overline{TC}_{cf}\right)$ is variation in the effectiveness of individual service providers, such as school teachers or program case managers. This variation can influence both the amount of cross-site variation in program effects that is observed and the meaning of this variation.

For example, it is well-documented that teachers vary in their effectiveness, with some estimates indicating that such variation accounts for 10 percent of the total variation in student achievement (Nye, Konstantopoulos, & Hedges, 2004). Therefore, underlying cross-site variation in the site-mean differences of teacher effectiveness for treatment and control group members could contribute substantially to the amount of cross-site impact variation that is observed.

This implies that cross-site variation in actual program effectiveness is *confounded* with cross-site variation in the implicit teacher effectiveness contrast. In an individually randomized experiment it is rarely possible to break this confound because teachers are almost never randomly assigned to implement a treatment or control condition.[24] Furthermore, for sites with a small number of treatment or control group teachers, little of the chance variation in teacher effectiveness will be averaged out when making cross-site comparisons. Hence, the influence of variation in the teacher effectiveness contrast on observed variation in program impacts will be magnified.

The national Head Start Impact Study (HSIS) is a good example of this situation. In that study, 71 percent of Head Start centers had eight or fewer children in the treatment group.[25] Thus many of these centers probably had only one or two instructors serving the treatment group. This could produce substantial cross-site variation in the teacher effectiveness contrast, which in turn could produce substantial cross-site variation in observed program effects. In general then, we

[24] There are examples of cluster randomized trials that randomize teachers/classrooms or schools – this can successfully break the confound. However, it is extremely rare to randomize individual students and, independently, to randomize instructors.

[25] In 72 percent of sites the control group had 5 or fewer children.

would expect studies with small numbers of individual service providers per site to have substantial observed cross-site impact variation, even if the effectiveness of the program itself (for a teacher with given effectiveness) were the same for all sites.

To examine this issue empirically, we turn to the Tennessee STAR class-size experiment, which is perhaps unique among large-scale educational evaluations because it randomized students and teachers to the treatment condition (reduced-size classes) or the control condition (regular-size classes) independently.[26] Moreover, within any given school (program site) there was often more than one teacher in the treatment or control conditions.[27] These two conditions make it possible to attempt to separate observed cross-variation in impacts on student outcomes due to variation in the teacher effectiveness contrast from that due to cross-site variation in the effectiveness of smaller classes. In other words, this property of the Tennessee STAR experimental design makes it possible to break the confound that typically exists between these two sources of cross-site program effect variation.

The following three-level hierarchical model was estimated for this purpose.

Level 1: Students

$$Y_{ijk} = A_{jk} + \sum_{l=1}^{L} \gamma_l X_{lijk} + e_{ijk} \qquad (10)$$

Level 2: Teachers

$$A_{jk} = \sum_{r=1}^{R} \alpha_r \, RA\_Block_{rjk} + B_k T_{jk} + a_{jk} \qquad (11)$$

Level 3: Schools (Sites)

$$B_k = \beta + b_k \qquad (12)$$

[26] We combine the regular-size classes and the regular-size classes with teaching aides to create the control condition, consistent with many earlier analyses.

[27] In half of the random assignment blocks there were at least two reduced-size classes and at least two regular-size classes.

Where:

$$e_{ijk} \sim N\left(0, \sigma^2_{|\boldsymbol{X}\alpha_r}(T_{ijk})\right)$$
$$a_{jk} \sim N(0, \eta^2)$$
$$b_k \sim N(0, \tau^2)$$
$$Cov(e_{ijk}, a_{jk}, b_k) = 0$$

The key difference between this model and that represented by Equations 3 and 4 is the addition of a middle level for teachers.[28] This middle level reflects the variability in student outcomes produced by the variability in teacher effectiveness and classroom composition beyond that produced by variation in the effects of class-size reduction.

Using the original two-level model represented by Equations 3 and 4, we estimate τ_Z to be 0.26σ in math and 0.23σ in reading. This represents the total amount of cross-site impact variation that is due both to cross-site differences in the effects of class size reduction and cross-site differences in the teacher effectiveness contrast. Using the three-level model represented by Equations 10 – 12 to remove the influence of variation in the teacher effectiveness contrast, we find that τ_Z decreases to 0.16σ (or by almost 40 percent) for math and to 0.09σ (or by almost 60 percent) for reading.[29] This suggests that between 40 and 60 percent of the total observed cross-site impact variation for Tennessee STAR is due to variation in the teacher effectiveness contrast. Essentially, there is a lot of variation in effects across schools in the study—with the treatment group doing much better than the control group in some schools and a much smaller difference in others. However, a lot of this variation arose simply because the treatment classroom was assigned a more/less effective teacher than the control classroom.

[28] In this model, Y_{ijk} is the value of the outcome measure for individual i in classroom j in site k. RA_Block_{rijk} equals one if individual i in classroom j in site k belongs to random assignment block r and zero otherwise. Here, the random assignment blocks refer to each unique school by grade combination. T_{ijk} equals one if individual i in classroom j in site k was assigned to treatment and zero otherwise.

[29] Nye, Hedges, and Konstantopoulos (2000) conduct related analyses and similarly find much more cross-site variation in treatment effects using a two-level model compared to a three level model.

It is difficult to know whether the Tennessee STAR findings are idiosyncratic or indicative of the general importance of service deliverers in individually randomized experiments – both with respect to interpreting the overall average treatment effect, as detailed by Weiss (2010), as well as when considering (and interpreting) how much treatment effects vary across sites, as described here. We suspect that it is quite important in both cases.

What about Client Characteristics?

The preceding discussion of the sources of cross-site variation in treatment effects focused on cross-site variation in the *treatment contrast*. Of course, cross-site differences in client characteristics can also produce cross-site impact variation. When designing a study, it may prove challenging to anticipate substantial cross-site impact variation resulting from client characteristics-based moderators because it requires knowledge of two factors: (1) strong evidence or theory regarding the type of clients who benefit least/most from the intervention, and (2) a priori knowledge that there will be substantial cross-site variation in the client characteristics that moderate the treatment effect. Our experience is that such knowledge and empirical evidence is uncommon. Additionally, attempts to replicate experimental findings regarding differential effects based on client characteristics are often unsuccessful – even less successful than efforts to replicate findings for treatment effect in the main sample (Rothwell, 2005; Tipton et al., forthcoming; Yusuf, Wittes, Probstfield, & Tyroler, 1991). Therefore, we believe that using client characteristics as a predictor of how much cross-site impact variation to expect may be difficult and unreliable.

Moreover, in general, we *suspect* that the majority of client characteristic-based impact heterogeneity occurs within sites rather than between sites. Consider perhaps the most common client characteristic hypothesized to moderate treatment effects: individual sample member’s risk of having poor outcomes. In education, prior achievement is the strongest predictor of future achievement and thus usually serves as the proxy for “at-risk.” Evidence shows that, on average, between 15 and 25 percent of variation in individual student achievement is between schools, with the vast majority of variation in achievement lying within schools (Hedges & Hedberg, 2014). Consequently, if prior achievement is a source of treatment effect heterogeneity, then the best place to look for this source of treatment effect heterogeneity is within schools, not between schools. This again indicates that, even when client characteristics do moderate treatment effects, they may not result in substantial cross-site impact variation if the characteristic mostly varies within sites.

It is worth acknowledging that other client characteristics may exhibit different distributions and some evaluations may be designed intentionally to select sites that serve dramatically different client types. In such cases, if relevant characteristics that vary across sites have strong theory suggesting those characteristics moderate treatment effects, then we may anticipate a larger than average τ.

What about Context?

Cross-site differences in context may also produce cross-site impact variation. We expect that this will most often occur through cross-site variation in the treatment contrast. For example, the same program’s effectiveness may vary dramatically when operating in a service poor environment compared to a service rich environment. In the former situation, the treatment contrast may be large, whereas in the latter situation it may be small. If we consider the treatment contrast a mediator of treatment assignment on the outcome, then we expect context to be an important moderator of treatment assignment on the treatment contrast, and thus the outcome. In other words, contextual moderation will generally operate through the treatment contrast (or other mediators). As a result, attention to the treatment contrast remains paramount.

Limitations of and Future Directions for the Present Research

This paper provides estimates of the standard deviation of site-average treatment effects – a key design parameter for planning multi-site RCTs. Importantly, what is presented are *estimates* of τ, not the true τ. Despite being estimated from large multi-site RCTs –we suspect that these estimates are still fairly imprecise in many cases.[30] Unfortunately, most (if not all) commercially available software packages, including SAS and HLM, do not provide confidence intervals for estimates of τ. While the estimates presented here may provide a good starting point for power calculations, we recommend using a range of values for τ to examine the sensitivity of your study design to different assumptions about variability of treatment effects across sites and, ultimately, to ensure adequate statistical precision of the study design. In future work we aim to include confidence intervals for $\hat{\tau}$.

As noted at the outset of this paper – the evaluations included in this article were purposively sampled and intended as a starting point for improving knowledge about how much program effects vary across sites. Consequently, the results presented are limited with respect to the situations and conditions they represent and are not necessarily broadly generalizable. This empirical contribution is analogous to the early contributions made by past researchers with respect to estimating key parameters to help guide the design of future cluster-randomized trials for educational research. That earlier research used existing databases to estimate key parameters (intra-class correlations for outcome measures and R-square values for baseline covariates) that

[30] See Bloom and Spybrook (under review) for a formula for the minimum detectable τ. Inserting the study design characteristics and relevant parameter estimates from the studies examined in this paper into this formula provides hints regarding the precision of the estimates of τ presented here.

are needed to design studies that randomize classrooms or whole schools to estimate program effects on student outcomes (Bloom et al., 1999; Bloom et al., 2007; Bloom et al., 2008; Hedges & Hedberg, 2014; Jacob et al., 2009; Westine et al., 2014; Xu & Nichols, 2010). Those findings have been used to design many (perhaps most) of the cluster-randomized studies of educational and child development interventions funded to date by IES, HHS, and private foundations. In addition, those early findings plus later updates to them have been integrated into the Optimal Design software program (Spybrook et al., 2011) and been made available via websites (Hedges & Hedberg, 2014) to make this empirical information easily available when using the software to design studies that randomize schools, pre-schools, or Head Start centers.

To place the present work in historical perspective, note that estimates of key design parameters for cluster-randomized studies were not obtained at one time from a single study. Rather they accumulated over time from a series of studies. Thus, while it was not possible to broadly generalize findings from the first publically-available estimates (Bloom et al., 1999), it is now possible to do so from the findings which have accumulated over the more than 15 years since. In addition, it is now possible to tailor the use of this accumulated information to the level of schooling (pre-school, primary school, secondary school or post-secondary education) and the target outcome of interest (achievement tests or social-emotional outcomes). The original findings set a precedent, provided a template, and encouraged subsequent work, and illustrated its usefulness to education researchers. We hope that the present article can do likewise.

Appendix A – Synopsis of Studies

Early Childhood—Elementary School

Head Start Impact Study (HSIS)
Program: The Head Start (HS) program seeks to better-prepare children for school. The most common HS programs operate as center-based programs, engaging with children primarily in a classroom setting and providing at least two home visits per year. Other models include the home-based program, the family child care program, and the combination program. The program can last for two years.
Target Population: Low-income children (3-4 year olds) in a nationally representative sample of HS programs, excluding programs intended to serve certain target populations.
Study Design: Individual random assignment within HS centers. The study compares students who were selected to enroll in Head Start with students who were not allowed to enroll in Head Start.
Outcomes: Cognitive measures from an abbreviated version of the Peabody Picture Vocabulary Test-III, and the letter-word identification, oral comprehension, and applied problems subscales of the Woodcock-Johnson III. Socio-emotional measures created based on parent-reported items from the Child Behavior Checklist and the Leiter-R Assessor Report. All assessments were done at the end of the year the student enrolled in Head Start.
Sample Size: Around 300 HS centers and around 3,500 children.
Report: Bloom and Weiland (2015)

After School - Reading
Program: Academically rigorous, structured after-school program using the Success for All reading curriculum. The program lasts for up to two years.
Target Population: Students in grades two to five attending an after-school program.
Study Design: Individual random assignment within each unique after-school center, grade, and cohort block. The study compares students who were selected to attend an academically-oriented after-school program with students who attended less formal academic support offered in a regular after-school program (both within the same center).
Outcomes: SAT-10 Total Reading at the end of their first year in the study.
Sample Size: 25 after-school centers and around 2,300 students. (Black et al., 2009, p.xvii)
Report: Black et al. (2009)

After School - Math
Program: Academically rigorous, structured after-school program using the Harcourt math curriculum. The program lasts for up to two years.
Target Population: Students in grades two to five attending an after-school program.
Study Design: Individual random assignment within each unique after-school center, grade, and cohort block. The study compares students who were selected to attend an academically-oriented after-school program with students who attended less formal academic support offered in a regular after-school program (both within the same center)
Outcomes: SAT-10 Total Math at the end of their first year in the study.
Sample Size: 25 after-school centers and around 2,500 students. (Black et al., 2009, p.xvii)
Report: Black et al. (2009)

Teach for America–Pooled Analysis of National TFA and i3 TFA Scape-Up Evaluations (TFA - Pooled)
Program: The TFA program provides an alternative pathway to teacher certification with the goal of providing high-quality teachers of hard-to-staff subjects in high-poverty schools. TFA is a highly selective program that recruits career professionals and recent graduates from some of the top colleges and universities across the United States. TFA teachers participate in a five-week summer training before they start teaching and receive ongoing support, such as one-on-one coaching, group meetings, and access to additional classroom resources and assessments, during their teaching commitment period. TFA requires its teachers to commit to two years of teaching.
Target Population: Elementary school students (grades K–5 and 1-5 for i3 TFA Scale Up and National TFA evaluations, respectively) attending high poverty schools.
Study Design: This analysis combines samples from two separate evaluations—National TFA Evaluation conducted in the 2001-2002 and 2002-2003 school years and i3 TFA Scale-up Evaluation conducted in the 2012-2013 school year. Both studies used individual random assignment within school and grade. This analysis compares students taught by TFA teachers to those taught by non-TFA teachers.
Outcomes: The two studies used a combination of study-administered and end-of-year state assessments in mathematics and reading. In the National TFA evaluation, the authors administered the Iowa Test of Basic Skills for mathematics and reading to students in grades 1–5. In the i3 TFA Scale Up, the authors administered Woodcock-Johnson III Normative Update Test of Achievement in mathematics and reading to students in grades K–2 and used end-of-year state assessments in mathematics and reading for students in grades 3–5. All outcomes were measured one year after random assignment.
Sample Size: 46 public schools and about 3,330 students across two studies
Reports: Clark, Isenberg, Liu, Makowsky, and Zukiewicz (2015); Decker, Mayer, and Glazerman (2004).

Tennessee STAR
Program: Small classes (13-17 students) and regular-size classes, either with or without a full-time teaching aide. The duration of the programs was up to four years. The original study team was intentional in working with schools to incorporate the study and interventions into the schools' existing frameworks and not encouraging or promoting any changes in curriculum or school schedule.
Target Population: Students in K-3 from schools with at least 57 students per grade, in 42 school systems in Tennessee.
Study Design: Individual and teacher random assignment into three experimental arms, within each unique school and grade. The original study compares students in small, regular-size, and regular-size with full-time teaching aide classrooms. As is common in reanalysis of this evaluation, we pool regular-size and regular-size with full-time teaching aide into one grouping.
Outcomes: SAT-7 Total Reading and SAT-7 Total Math at the end of a student's first year in the study.
Sample Size: 80 schools and around 10,000 students.
Report: Word et al. (1990)

Middle School-High School

Charter Middle Schools (Charters)
Program: Public schools that are released from many state and district regulations, including those involving staffing, curriculum, and budget decisions but are accountable for the quality of student outcomes and may be closed by their authorizer if they fail to meet expectations.
Target Population: Middle school students (grades 4-7) who applied to oversubscribed charter schools and participated in the schools' admissions lotteries.
Study Design: Individual random assignment within schools. This study compares students who applied and were offered admission to a charter middle school ("lottery winners") to those who applied but were not offered admission to a charter middle schools ("lottery losers").
Outcomes: End-of-year state assessments in mathematics and reading, one and two years after random assignment.
Sample Size: 29 charter middle schools across 15 states and about 2,100 students. (Gleason et al., 2010, p.12)
Report: Gleason et al. (2010)

Teach for America – Math (TFA – Math)
Program: TFA provides an alternative pathway to teacher certification with the goal of providing high-quality teachers of hard-to-staff subjects in high-poverty schools. TFA is a highly selective program that recruits career professionals and recent graduates from some of the top colleges and universities across the United States. TFA teachers receive training and supports before and after they start teaching. TFA requires its teachers to commit to two years of teaching.
Target Population: Middle and high school students (grades 6–12) attending high poverty schools.
Study Design: Individual random assignment within school, grade, and subject. In each participating school, the authors identified "classroom matches"—two or more classes covering the same math course and level, where at least one was assigned to a TFA teacher and at least one assigned to a non-TFA teacher. In each classroom match, students were randomly assigned to a class taught by a TFA teacher or a non-TFA teacher. The study compares students assigned to classes taught by TFA teachers to those taught by non-TFA teachers.
Outcomes: For students in grades 6–8, the authors used end-of-year state assessment in mathematics one year after random assignment. For students in grades 9–12, the authors used end-of-course math assessment developed by Northwest Evaluation Association (NWEA).
Sample Size: 45 public middle and high schools across 11 school districts in 8 states, and about 4,570 students. (Clark et al., 2013, p.13)
Report: Clark et al. (2013)

Enhanced Reading Opportunity (ERO)
Program: Ninth grade students take a supplemental reading course in place of an elective class, using either the Reading Apprenticeship Academic Literacy (RAAL) program or the Xtreme Reading program. The program lasted for one year.
Target Population: Ninth grade students whose reading ability was at least two years below grade level.
Study Design: Individual random assignment within each school by cohort block. Schools were randomly assigned to one of the two reading curricula. The study compares students who were selected to enroll in the supplemental reading class (using either the RAAL or the Xtreme Reading program) with students who took another, elective class regularly offered by the school.
Outcomes: Reading comprehension and reading vocabulary GRADE assessment, and credits earned as a percentage of credits required for graduation (program year and follow-up year).
Sample Size: About 34 public schools and about 5,500 students. (Somers et al., 2010, p. ES-4)
Report: Somers et al. (2010)

Small Schools of Choice (SSC)
Program: Small (typically under 400 students) and academically nonselective schools operating under the design principles of academic rigor, personalization, and community partnerships. The program can last for at least four years.
Target Population: New York City High School students who indicated interest in enrolling in an SSC.
Study Design: Individual random assignment based on a complicated lottery-based admission system. The study compares students who indicated interest in enrolling in an SSC and were admitted to the SSC to those who indicated interest but were not admitted.
Outcomes: 9th grade "on-track" indicator and 4-year graduation rate.
Sample Size: 84 to 87 schools and 17,000 to 26,000 students (depending on the outcome)
Report: Bloom and Unterman (2014)

Career Academies
Program: A "school-within-school" structure to foster a supportive learning community, integrated academic and career curricula, and partnerships with local employers to provide work-based learning opportunities. The program can last for three or four years.
Target Population: High school students who were academy applicants in the 8th or 9th grade.
Study Design: Individual random assignment within each school by cohort block. The study compares students who were selected to enroll in the career academy with students who were not allowed to enroll in the career academy, but instead received regular high school services (both within the same school).
Outcomes: Five-year graduation rates, enrollment in post-secondary within 14 months of expected high school graduation, and earnings and employment in years 1-4 after expected high school graduation and years 5-8.
Sample Size: 9 High Schools and around 1,500 students.
Report: Kemple (2001); Kemple (2008)

Early College High School (ECHS)
Program: Provides students with concurrent high and college experiences. Students attend high school on a college campus, enroll in college courses, and are expected to complete two-years of transferable college credits or an associate's degree by the time they earn their high school diploma.
Target Population: High school students that are underrepresented in college – first in their family to go to college; low-income students; members of racial and ethnic groups that are underrepresented.
Study Design: Schools were selected based on whether they were overenrolled and agreed to use a lottery system to place students. Students were randomly assigned by cohort within each school. In some cases, the lottery for a given school/cohort combination assigned students probabilities for selection into the program group.
Outcomes: 9th grade "on-track" indicator and 5-year graduation rate
Sample Size: 19 schools and nearly 4,000 students (depending on the outcome)
Report: Edmunds et al. (forthcoming)

Post-Secondary Education

Learning Communities (LC)
Program: Small cohorts (around 20-25 students) co-enroll in two or more classes. Instructors are encouraged to collaborate to integrate curricula, assessments, and use similar approaches to support struggling students. The program lasts for 1 semester.
Target Population: Community college students in need of developmental/remedial education in math and/or English.
Study Design: Individual random assignment within each campus by cohort block. The study compares students who were allowed to enroll in the learning community with students who could enroll in their college's usual courses and services, just not learning communities.
Outcomes: Credit accumulation at the end of the program semester and after three semesters, including: targeted credits earned (in developmental education classes) and total credits.
Sample Size: 11 community college campuses and nearly 7,000 students.
Report: Weiss, Visher, Weissman, and Wathington (2015)

Performance-Based Scholarships (PBS)
Program: A form of conditional-cash transfer. Scholarships are not contingent upon past performance, but are awarded based on meeting pre-specified benchmarks. Scholarship amounts and performance criteria varied across colleges. Maximum scholarship amounts ranged from $600 to $1,500 per term. Scholarship durations ranged from 2-4 semesters. Some colleges also had a service component.
Target Population: Low-income community college students.
Study Design: Individual random assignment within each campus by cohort block. The study compares students offered a PBS to those who were not offered a PBS.
Outcomes: Total credits earned at the end of one and three years, and three-year graduation rates.
Sample Size: 15 community college campuses and nearly 7,000 students.
Report: Mayer, Patel, Rudd, and Ratledge (2015)

Labor/Workforce Development

Job Corps
Program: The largest federal education and training program. It is an intensive, comprehensive program whose major service components include academic education, vocational training, residential living, health care and health education, counseling, and job placement assistance. Services are provided in over 100 centers across the country. Students stay for an average of 8 months but the length of the program varies.
Target Population: Disadvantaged youth between ages of 16 and 24.
Study Design: Individual random assignment. Center was identified based on the counselors' predictions prior to random assignment. The study compared the outcomes of individuals who were offered admission to a Job Corps program to those who were not.
Outcomes: Annual employment and earnings outcomes for the first four years after random assignment; attainment of high school diploma or GED (for youth who did not have them prior to random assignment); percent arrested at least once during the first four years.
Sample Size: 100 centers and over 10,000 individuals.
Report: Schochet, Burghardt, and Glazerman (2001); Schochet, Burghardt, and McConnell (2008)

Welfare-to-Work Programs (WtW)
Program: Conditional cash assistance program. The types of services offered by welfare-to-work can vary on a number of dimensions, including: whether job searching or training are prioritized, the relationship between frontline staff and clients and the size of staff caseload, and how closely the programs monitor client activities. The program had no standard length – as long as clients were receiving welfare they were subject to the work requirement.
Target Population: Low-income welfare recipients.
Study Design: Individual random assignment within welfare centers. This study compares welfare recipients and applicants who were assigned to a new employment program comprised of services, regulations, and potential sanctions for non-compliance to those who were not assigned to the new program and thereby were only exposed to "business as usual."
Outcomes: Average annual earnings over two years.
Sample Size: 59 welfare centers including over 69,000 clients. The analysis sample was restricted to include only female sample members. Additionally, following the original analyses, some offices were excluded due to their small size, unusual client mix, or incomplete data.
Report: Bloom et al. (2003)

Appendix B – SAS Code

The following SAS code was used to implement the two-level model described in equations (3) and (4), using restricted maximum likelihood:

```
PROC MIXED DATA = <DATASET> METHOD = REML COVTEST;
 CLASS <RA_BLOCK> <SITE>;
 MODEL <OUTCOME> = <RA_BLOCK> <TREATMENT> <COVARIATES>
                 / DDFM = BW NOINT SOLUTION;
 RANDOM <TREATMENT> / SUBJECT = <SITE> SOLUTION TYPE = UN G GCORR;
 REPEATED / GROUP = <TREATMENT>;
RUN;
```

The input dataset (<*DATASET*>) is an individual-level file, with one record per study participant. The model, specified in the MODEL statement, models the outcome variable (<*OUTCOME*>) on a binary treatment indicator (<*TREATMENT*>), the vector of random assignment blocks (<*RA_BLOCK*>), and individual-level covariates (<*COVARIATES*>). The RANDOM statement allows the treatment effect coefficient to vary randomly across the specified SUBJECT, in this case, <*SITE*>. In other words, τ^2 is estimated as the cross-site variation in impacts. In order to model the residual variances ($\sigma^2_{|\mathbf{X}\alpha_r}(T_{ij} = 1)$ and $\sigma^2_{|\mathbf{X}\alpha_r}(T_{ij} = 0)$) separately for each experimental group, the REPEATED statement is included, where the GROUP variable is the treatment indicator.[31]

[31] We used an alternative approach, which excluded random assignment block indicators and instead included group-mean centering the outcome, treatment indicator, and other covariates around their random-assignment block means, for Teach for America—pooled analysis. This model was estimated using HLM software.

Appendix C – Calculation of the Q-Statistic for Significance Testing of $\hat{\boldsymbol{\tau}}$

To determine the statistical significance of the key parameter of interest, $\hat{\tau}$, we borrow the Q-statistic and its associated p-value from meta-analysis, as presented by Hedges and Pigott (2001). We refer to the population average treatment effect at site j as B_j and its sample-based estimate as $\hat{B}_j$, with corresponding known variance v_j.[32]

Formally, the hypothesis test of interest is:

$$H_0: B_1 = B_2 = \cdots = B_J,$$

which implies a null hypothesis of $\tau^2 = 0$. The alternative is that at least one B_j is not equal to the others, or $\tau^2 > 0$. The exact small-sample test of H_0 is based on the Q-statistic, where

$$Q = \sum_{j=1}^{J} \frac{\left(\hat{B}_j - \bar{B}.\right)^2}{v_j} \tag{C.1}$$

If all sites have the same B_j, that is, if there is no variation in effects across sites and $\tau^2 = 0$, then the test statistic Q has a χ^2 distribution with $J - 1$ degrees of freedom. A comparison of Q to the χ^2 distribution with $J - 1$ degrees of freedom yields a p-value indicating the probability of observing the amount of variation in site-average treatment effects seen in the data (or greater), under the null hypothesis. Small p-values provide evidence to reject the null hypothesis and accept the alternative hypothesis.

Estimates of B_j and v_j are obtained using the following fixed-effects model:

$$Y_i = \sum_{r=1}^{R} \alpha_r RA\_Block_{ri} + \sum_{j=1}^{J} B_j Site_{ji} T_i + \sum_{l=1}^{L} \gamma_l X_{li} + e_i \tag{C.2}$$

Here, Y_i is the value of the outcome measure for sample member i. RA_Block_{ri} is a vector of R *random assignment block* indicators, where each block indicator is set equal to one if individual i

[32] A noted by Hedges and Pigott (2001), the calculations described are only approximate when the v_j's are not known exactly, as is the case.

was randomly assigned within that block and set equal to zero otherwise. $Site_{ji}$ is a vector of J *site* indicators, where each site indicator is set equal to one if individual i was randomly assigned within that site and set equal to zero otherwise. T_i is a binary indicator, set equal to one if individual i was randomly assigned to the treatment group, and zero otherwise. X_{li} is a vector of L covariates. We allow the variance of the residual error term to differ for the treatment and control groups, i.e., $\varepsilon_i \sim N\left(0, \sigma^2_{|\boldsymbol{X}\alpha_r}(T_{ij})\right)$. In those studies where probability of assignment to treatment was unequal within blocks, the regression model adjusts for unequal probability of assignment to treatment within blocks using individual-level weights created by the original study teams.

We use the standard errors of $\hat{B}_j$ from the regression model above as the estimate for the v_j's. $\bar{B}.$ is calculated as the weighted average of the $\hat{B}_j$, where the weights are equal to $1/\hat{v}_j$. Plugging these values into equation C.1 yields the Q-statistic.

TABLES

Table 1. Study Design Characteristics

Outcome	**Indiv.** **(N)**	**Sites** **(J)**	**RA Blocks** **(R)**	**P[t=1]** **($\bar{T}$)**
Early Childhood - Elementary School				
Head Start Impact Study (HSIS)[a]				
WJ-III AP Early numeracy, yr 1	3,566	25	316	0.63
WJ-III LW Early reading, yr 1	3,593	25	317	0.62
WJ-III OC Oral comprehension, yr 1	3,531	25	317	0.63
PPVT-III Receptive vocabulary, yr 1	3,586	25	317	0.62
Externalizing behavior problems, yr 1	3,601	25	318	0.61
Self-regulation skills, yr 1	3,616	25	318	0.62
After School Reading[b]				
SAT-10 total reading	2,340	25	140	0.57
After School Math[b]				
SAT-10 total math	2,500	25	150	0.56
Teach for America: Pooled (TFA-Pooled)				
Math	3,197	43	76	0.42
Reading	3,300	45	78	0.43
Tennessee STAR				
SAT-7 total math	10,045	80	280	0.26
SAT-7 total reading	9,883	80	280	0.26
Middle School-High School				
Charter Middle Schools (Charters)[b]				
Math, yr 1	2,110	29	30	0.62
Math, yr 2	2,000	29	30	0.62
Reading, yr 1	2,130	29	30	0.62
Reading, yr 2	2,000	29	30	0.62
Teach for America: Math (TFA-Math)[b]				
Math	4,570	45	110	0.50
Enhanced Reading Opportunities (ERO)[b]				
GRADE Reading comprehension	4,580	34	70	0.58
GRADE Reading vocabulary	4,580	34	70	0.58
Proportion of required credits earned, yr 1	5,230	34	70	0.57
Proportion of required credits earned, yr 2	4,550	34	70	0.57
Small Schools of Choice (SSC)[c]				
On track in ninth grade	25,925	87	356	0.49
Earned a high school diploma, yr 4	14,137	84	227	0.46
Career Academies[d]				
Earned HS diploma or equivalent, yr 5	1,533	9	20	0.55
Enrolled in post-secondary	1,482	9	20	0.55
Avg. annual earnings, yrs 1-4	1,458	9	20	0.55
Avg. annual earnings, yrs 5-8	1,405	9	20	0.55
Avg. months worked annually, yrs 1-4	1,458	9	20	0.55
Avg. months worked annually, yrs 5-8	1,405	9	20	0.55

(continued)

Table 1. (continued)

Outcome	Indiv. (N)	Sites (J)	RA Blocks (R)	P[t=1] ($\bar{T}$)
Early College High School (ECHS)				
On track in ninth grade	3,887	19	119	0.58
Earned a high school diploma, yr 5	2,847	19	83	0.58
Post-Secondary Education				
Learning Communities (LC)				
Targeted credits earned, sem 1	6,974	11	36	0.57
Cumulative targeted credits earned, sem 3	6,974	11	36	0.57
Total credits earned, sem 1	6,974	11	36	0.57
Cumulative total credits earned, sem 3	6,974	11	36	0.57
Performance-Based Scholarships (PBS)				
Cumulative total credits earned, yr 1	6,938	15	39	0.57
Cumulative total credits earned, yr 3	6,938	15	39	0.57
Earned a degree, yr 3	6,971	15	39	0.56
Labor/Workforce				
Job Corps Program				
Total annual earnings, yr 1	10,027	100	100	0.60
Total annual earnings, yr 2	10,218	100	100	0.61
Total annual earnings, yr 3	10,238	100	100	0.61
Total annual earnings, yr 4	10,001	100	100	0.61
Months worked, yr 1	10,027	100	100	0.60
Months worked, yr 2	10,218	100	100	0.61
Months worked, yr 3	10,238	100	100	0.61
Months worked, yr 4	10,234	100	100	0.61
Earned HS diploma or equivalent	7,819	100	100	0.60
Ever arrested	10,372	100	100	0.61
Welfare-to-Work Program (WtW)[e]				
Avg. annual earnings, qtrs 1-8	69,399	59	59	0.61

NOTES: [a] The HSIS team administered a shorter version of the Peabody Picture Vocabulary Test III (PPVT-III). WJ-III is Woodcock-Johnson III; LW stands for the Letter-Word Identification subscale; OC stands for the Oral Comprehension subscale; AP stands for the Applied Problems subscale. Externalizing behavior problems is based on parent-reported items from the Child Behavior Checklist. Self-regulation skills is based on parent-reported items from the Leiter-R Assessor Report.

[b] Data are from IES restricted-use files. Sample sizes and number of random assignment blocks have been rounded to the nearest tens.

[c] Due to availability of follow-up data for the SSC evaluation, the on track in ninth grade indicator is based on seven cohorts of students whereas the four-year graduation is based on four cohorts.

[d] In Career Academies, earned HS diploma or equivalent and enrollment in post-secondary were measured 14 months after expected HS graduation. Years 1 to 4 earnings and employment capture the first four years after expected HS graduation; Years 5 to 8 earnings and employment capture the next four years.

[e] Mirroring the approach of the original study team, the analysis sample for the WtW analysis was restricted to include only female sample members. Additionally, some offices were excluded due to their small size, unusual client mix, or incomplete data.

Table 2. Glossary of Notation

Notation	**Definition**
	Outcome Metrics
$\boldsymbol{Y}$:	The outcome in "natural" units (e.g., dollars) or original units (e.g., test scaled score)
$\boldsymbol{Z}$:	A "z-score" outcome metric constructed such that the reference group's mean is zero and standard deviation is one. $\boldsymbol{Z = \frac{Y_i - \bar{Y}}{\sigma_Y}}$
	Standard Deviation of the Outcome ($\boldsymbol{\sigma}$*)*
$\boldsymbol{\sigma_Y}$:	The standard deviation of the outcome in the $\boldsymbol{Y}$ metric across individuals in the study's control group
$\boldsymbol{\sigma_{Y(RG)}}$:	The standard deviation of the outcome in the $\boldsymbol{Y}$ metric across individuals in a particular reference group
	Mean of the Cross-site Distribution of Site Mean Effects ($\boldsymbol{\beta}$*)*
$\boldsymbol{\beta_Y}$:	The mean of the cross-site distribution of site mean effects in the $\boldsymbol{Y}$ metric
$\boldsymbol{\beta_Z}$:	The mean of the cross-site distribution of site mean effect sizes in the $\boldsymbol{Z}$ metric. Commonly referred to as an "effect size" this parameter indicates the magnitude of the treatment effect relative to the total outcome variability of the reference group.
	Standard Deviation of the Cross-site Distribution of Site Mean Effects ($\boldsymbol{\tau}$*)*
$\boldsymbol{\tau_Y}$:	The standard deviation of the cross-site distribution of site mean effects in the $\boldsymbol{Y}$ metric.
$\boldsymbol{\tau_Z}$:	The standard deviation of the cross-site distribution of site mean effect sizes in the $\boldsymbol{Z}$ metric. This parameter indicates the magnitude of variation in site-average treatment effects relative to the outcome variability of the reference group.
	Minimum Detectable Effect (MDE) and Effect Size (MDES)
$\boldsymbol{MDE_Y}$:	The smallest true mean effect, in $\boldsymbol{Y}$ units, that a study design is likely to detect.
$\boldsymbol{MDES_Z}$:	The smallest true mean effect, in $\boldsymbol{Z}$ units, that a study design is likely to detect. Commonly referred to as the Minimum Detectable Effect Size (MDES)

Table 3. Estimates of the Mean and Standard Deviation of the Distribution of Site-Average Treatment Effects in Natural Units (*Y*) or Reference-population-based Effect Size Units (*Z*) on a Scale where the Reference Population Standard deviation=1

	Control	Distribution of Site-Average Treatment Effects				
	Mean	Mean (β)			Standard Deviation (τ)	
Outcome	Level	Estimate	(SE)	P	Estimate	P
Early Childhood - Elementary School						
Head Start Impact Study (HSIS)[a]						
WJ-III AP Early numeracy, yr 1 (Z)	-0.13	0.13	(0.03)	<0.0001	0.00	0.2225
WJ-III LW Early reading, yr 1 (Z)	-0.34	0.20	(0.04)	<0.0001	0.30	0.0004
WJ-III OC Oral comprehension, yr 1 (Z)	-0.47	0.02	(0.03)	0.5039	0.15	0.0254
PPVT-III Receptive vocabulary, yr 1	-	-	-	-	-	-
Externalizing behavior problems, yr 1	-	-	-	-	-	-
Self-regulation skills, yr 1	-	-	-	-	-	-
After School Reading						
SAT-10 total reading (Z)	-0.77	-0.02	(0.02)	0.3929	0.04	0.3703
After School Math						
SAT-10 total math (Z)	-0.23	0.07	(0.03)	0.0086	0.00	0.8782
Teach for America: Pooled (TFA-Pooled)						
Math (Z)	-0.76	0.10	(0.03)	0.0010	0.05	0.0770
Reading (Z)	-0.68	0.05	(0.03)	0.0480	0.02	0.1060
Tennessee STAR						
SAT-7 total math	-	-	-	-	-	-
SAT-7 total reading	-	-	-	-	-	-
Middle School-High School						
Charter Middle Schools (Charters)						
Math, yr 1 (Z)	0.37	-0.02	(0.04)	0.6252	0.14	0.0028
Math, yr 2 (Z)	0.48	-0.06	(0.07)	0.4066	0.30	<0.0001
Reading, yr 1 (Z)	0.44	-0.02	(0.04)	0.5721	0.15	0.0004
Reading, yr 2 (Z)	0.46	-0.07	(0.04)	0.1280	0.16	0.0022
Teach for America: Math (TFA-Math)						
Math (Z)	-0.60	0.08	(0.03)	0.0016	0.10	0.0016
Enhanced Reading Opportunities (ERO)						
GRADE Reading comprehension (Z)	-0.70	0.07	(0.02)	0.0032	0.08	0.0236
GRADE Reading vocabulary (Z)	-0.43	0.02	(0.02)	0.2380	0.00	0.4112
Proportion of required credits earned, yr 1 (perc pt)	20.5	0.5	(0.3)	0.0381	0.5	0.2685
Proportion of required credits earned, yr 2 (perc pt)	43.1	0.2	(0.5)	0.6569	1.0	0.2295
Small Schools of Choice (SSC)[b]						
On track in ninth grade (perc pt)	51.6	10.3	(1.9)	<0.0001	15.3	<0.0001
Earned a high school diploma, yr 4 (perc pt)	60.6	6.7	(1.6)	<0.0001	11.5	<0.0001
Career Academies[c]						
Earned HS diploma or equivalent, yr 5 (perc pt)	86.5	0.1	(1.7)	0.9326	0.0	0.8673
Enrolled in post-secondary (perc pt)	55.3	0.2	(2.8)	0.9353	3.5	0.4076
Avg. annual earnings, yrs 1-4 (2015 $)	17,518	1,883	(660)	0.0044	0	0.5664
Avg. annual earnings, yrs 5-8 (2015 $)	27,723	2,313	(1,249)	0.0644	0	0.9278
Avg. months worked annually, yrs 1-4	8.8	0.3	(0.2)	0.0503	0.0	0.5374
Avg. months worked annually, yrs 5-8	9.4	0.2	(0.2)	0.2468	0.1	0.3698

(continued)

Table 3. (continued)

	Control	Distribution of Site-Average Treatment Effects				
	Mean	Mean (β)			Standard Deviation (τ)	
Outcome	Level	Estimate	(SE)	P	Estimate	P
Early College High School (ECHS)						
On track in ninth grade (perc pt)	88.7	3.7	(2.1)	0.0767	8.2	<0.0001
Earned a high school diploma, yr 5 (perc pt)	82.7	2.5	(1.3)	0.0585	0.0	0.4141
Post-Secondary Education						
Learning Communities (LC)						
Targeted credits earned, sem 1	2.2	0.4	(0.2)	0.0374	0.5	<0.0001
Cumulative targeted credits earned, sem 3	4.3	0.3	(0.2)	0.1449	0.5	0.0040
Total credits earned, sem 1	6.5	0.5	(0.2)	0.0024	0.2	0.3027
Cumulative total credits earned, sem 3	15.4	0.4	(0.3)	0.1811	0.0	0.8270
Performance-Based Scholarships (PBS)						
Cumulative total credits earned, yr 1	16.5	1.3	(0.3)	0.0003	0.8	0.0231
Cumulative total credits earned, yr 3	35.6	1.8	(0.7)	0.0143	1.3	0.0688
Earned a degree, yr 3 (perc pt)	17.4	1.9	(1.0)	0.0646	1.5	0.5544
Labor/Workforce						
Job Corps Program						
Total annual earnings, yr 1 (2015 $)	7,172	-1,740	(179)	<0.0001	592	0.1613
Total annual earnings, yr 2 (2015 $)	10,619	-15	(261)	0.9538	1,072	0.0442
Total annual earnings, yr 3 (2015 $)	13,223	1,172	(317)	0.0002	1,716	0.0009
Total annual earnings, yr 4 (2015 $)	15,595	1,415	(358)	<0.0001	1,687	0.0135
Months worked, yr 1	4.5	-1.2	(0.1)	<0.0001	0.4	0.0410
Months worked, yr 2	5.6	-0.2	(0.1)	0.0250	0.6	0.0058
Months worked, yr 3	6.4	0.2	(0.1)	0.0540	0.6	0.0008
Months worked, yr 4	6.8	0.4	(0.1)	0.0002	0.2	0.5136
Earned HS diploma or equivalent (perc pt)	34.2	12.7	(1.3)	<0.0001	6.9	0.0040
Ever arrested (perc pt)	31.7	-3.2	(0.9)	0.0003	2.1	0.1279
Welfare-to-Work Program (WtW)[d]						
Avg. annual earnings, qtrs 1-8 (2015 $)	3,700	670	(104)	<0.0001	601	<0.0001

NOTES: Outcomes listed with (Z) appended are in effect size units where the mean and standard deviation are of the broadest population taking the assessment within the grade, based on data availability (e.g., the national or state).

All earnings outcomes have been converted to 2015 dollars.

Items marked "-" are outcomes for which we were unable to create reference-population-based z-score.

[a] The HSIS team administered a shorter version of the Peabody Picture Vocabulary Test III (PPVT-III). WJ-III is Woodcock-Johnson III; LW stands for the Letter-Word Identification subscale; OC stands for the Oral Comprehension subscale; AP stands for the Applied Problems subscale. Externalizing behavior problems is based on parent-reported items from the Child Behavior Checklist. Self-regulation skills is based on parent-reported items from the Leiter-R Assessor Report. Some baseline covariates used in the estimation model were imputed by the original study team and non-imputed data are not available for these covariates.

[b] Due to availability of follow-up data for the SSC evaluation, the on track in ninth grade indicator is based on seven cohorts of students whereas the four-year graduation is based on four cohorts.

[c] In Career Academies, earned HS diploma or equivalent and enrollment in post-secondary were measured 14 months after expected HS graduation. Years 1 to 4 earnings and employment capture the first four years after expected HS graduation; Years 5 to 8 earnings and employment capture the next four years.

[d] Mirroring the approach of the original study team, the analysis sample for the WtW analysis was restricted to include only female sample members. Additionally, some offices were excluded due to their small size, unusual client mix, or incomplete data. Baseline covariates used in the estimation model were imputed by the original study team and non-imputed data are not available.

Table 4. Estimates of the Mean and Standard Deviation of the Distribution of Site-Average Treatment Effects in Control-group-based Effect Size Units (*Z*) on a Scale where the Study Control Group's Standard Deviation=1.

	Distribution of Site-Average Treatment Effects				
	Mean (β)			Standard Deviation (τ)	
Outcome	**Estimate**	**(SE)**	**P**	**Estimate**	**P**
Early Childhood - Elementary School					
Head Start Impact Study (HSIS)[a]					
WJ-III AP Early numeracy, yr 1	0.12	(0.03)	<0.0001	0.00	0.2259
WJ-III LW Early reading, yr 1	0.18	(0.03)	<0.0001	0.27	0.0009
WJ-III OC Oral comprehension, yr 1	0.02	(0.03)	0.4669	0.16	0.0122
PPVT-III Receptive vocabulary, yr 1	0.17	(0.02)	<0.0001	0.09	0.2176
Externalizing behavior problems, yr 1	-0.05	(0.03)	0.1255	0.16	0.0142
Self-regulation skills, yr 1	-0.02	(0.03)	0.5758	0.18	0.0028
After School Reading					
SAT-10 total reading	-0.03	(0.03)	0.4248	0.06	0.3416
After School Math					
SAT-10 total math	0.08	(0.03)	0.0086	0.00	0.8833
Teach for America: Pooled (TFA-Pooled)					
Math	0.12	(0.04)	0.0010	0.07	0.1460
Reading	0.08	(0.03)	0.0230	0.04	0.0550
Tennessee STAR					
SAT-7 total math	0.17	(0.04)	<0.0001	0.26	<0.0001
SAT-7 total reading	0.15	(0.03)	<0.0001	0.23	<0.0001
Middle School-High School					
Charter Middle Schools (Charters)					
Math, yr 1	-0.04	(0.05)	0.4812	0.22	<0.0001
Math, yr 2	-0.04	(0.07)	0.5600	0.35	<0.0001
Reading, yr 1	-0.02	(0.04)	0.7313	0.15	0.0100
Reading, yr 2	-0.06	(0.06)	0.3022	0.25	<0.0001
Teach for America: Math (TFA-Math)					
Math	0.10	(0.04)	0.0054	0.16	<0.0001
Enhanced Reading Opportunities (ERO)					
GRADE Reading comprehension	0.10	(0.03)	0.0032	0.12	0.0236
GRADE Reading vocabulary	0.03	(0.03)	0.2380	0.00	0.4112
Proportion of required credits earned, yr 1	0.05	(0.03)	0.0381	0.05	0.2685
Proportion of required credits earned, yr 2	0.01	(0.03)	0.6569	0.05	0.2295
Small Schools of Choice (SSC)[b]					
On track in ninth grade	0.21	(0.04)	<0.0001	0.31	<0.0001
Earned a high school diploma, yr 4	0.14	(0.03)	<0.0001	0.23	<0.0001
Career Academies[c]					
Earned HS diploma or equivalent, yr 5	0.00	(0.05)	0.9326	0.00	0.8673
Enrolled in post-secondary	0.00	(0.06)	0.9353	0.07	0.4076
Avg. annual earnings, yrs 1-4	0.16	(0.06)	0.0044	0.00	0.5664
Avg. annual earnings, yrs 5-8	0.09	(0.05)	0.0644	0.00	0.9278
Avg. months worked annually, yrs 1-4	0.10	(0.05)	0.0503	0.00	0.5374
Avg. months worked annually, yrs 5-8	0.06	(0.05)	0.2468	0.02	0.3698

(continued)

Table 4. (continued)

	Distribution of Site-Average Treatment Effects				
	Mean (β)			Standard Deviation	
Outcome	Estimate	(SE)	P	Estimate	P
Early College High School (ECHS)					
On track in ninth grade	0.12	(0.07)	0.0767	0.26	<0.0001
Earned a high school diploma, yr 5	0.07	(0.03)	0.0585	0.00	0.4141
Post-Secondary Education					
Learning Communities (LC)					
Targeted credits earned, sem 1	0.12	(0.06)	0.0374	0.16	<0.0001
Cumulative targeted credits earned, sem 3	0.06	(0.04)	0.1449	0.10	0.0040
Total credits earned, sem 1	0.08	(0.03)	0.0024	0.04	0.3027
Cumulative total credits earned, sem 3	0.03	(0.02)	0.1811	0.00	0.8270
Performance-Based Scholarships (PBS)					
Cumulative total credits earned, yr 1	0.12	(0.03)	0.0003	0.08	0.0231
Cumulative total credits earned, yr 3	0.06	(0.03)	0.0143	0.04	0.0688
Earned a degree, yr 3	0.05	(0.03)	0.0646	0.04	0.5544
Labor/Workforce					
Job Corps Program					
Total annual earnings, yr 1	-0.19	(0.02)	<0.0001	0.07	0.1613
Total annual earnings, yr 2	0.00	(0.02)	0.9538	0.09	0.0442
Total annual earnings, yr 3	0.09	(0.02)	0.0002	0.13	0.0009
Total annual earnings, yr 4	0.09	(0.02)	<0.0001	0.11	0.0135
Months worked, yr 1	-0.29	(0.02)	<0.0001	0.10	0.0410
Months worked, yr 2	-0.05	(0.02)	0.0250	0.12	0.0058
Months worked, yr 3	0.05	(0.02)	0.0540	0.14	0.0008
Months worked, yr 4	0.07	(0.02)	0.0002	0.04	0.5136
Earned HS diploma or equivalent	0.27	(0.03)	<0.0001	0.14	0.0040
Ever arrested	-0.07	(0.02)	0.0003	0.04	0.1279
Welfare-to-Work Program (WtW)[d]					
Avg. annual earnings, qtrs 1-8	0.10	(0.02)	<0.0001	0.09	<0.0001

NOTES: [a]The HSIS team administered a shorter version of the Peabody Picture Vocabulary Test III (PPVT-III). WJ-III is Woodcock-Johnson III; LW stands for the Letter-Word Identification subscale; OC stands for the Oral Comprehension subscale; AP stands for the Applied Problems subscale. Externalizing behavior problems is based on parent-reported items from the Child Behavior Checklist. Self-regulation skills is based on parent-reported items from the Leiter-R Assessor Report. Some baseline covariates used in the estimation model were imputed by the original study team and non-imputed data are not available for these covariates.

[b]Due to availability of follow-up data for the SSC evaluation, the on track in ninth grade indicator is based on seven cohorts of students whereas the four-year graduation is based on four cohorts.

[c]In Career Academies, earned HS diploma or equivalent and enrollment in post-secondary were measured 14 months after expected HS graduation. Years 1 to 4 earnings and employment capture the first four years after expected HS graduation; Years 5 to 8 earnings and employment capture the next four years.

[d]Mirroring the approach of the original study team, the analysis sample for the WtW analysis was restricted to include only female sample members. Additionally, some offices were excluded due to their small size, unusual client mix, or incomplete data. Baseline covariates used in the estimation model were imputed by the original study team and non-imputed data are not available.

Appendix Table A.1. Estimates of σ, ρ, and R^2_{within}

	Natural Units (Y)			Standardized Effect Size Units (Z)					
				Reference			Control Group		
Outcome	σ	ρ	R^2_{within}	σ	ρ	R^2_{within}	σ	ρ	R^2_{within}
Early Childhood - Elementary School									
Head Start Impact Study (HSIS)[a]									
WJ-III AP Early numeracy, yr 1	-	-	-	1.10	0.14	0.16	1.00	0.14	0.16
WJ-III LW Early reading, yr 1	-	-	-	1.08	0.10	0.22	1.00	0.10	0.22
WJ-III OC Oral comprehension, yr 1	-	-	-	0.98	0.34	0.16	1.00	0.34	0.16
PPVT-III Receptive vocabulary, yr 1	-	-	-	-	-	-	1.00	0.25	0.23
Externalizing behavior problems, yr 1	-	-	-	-	-	-	1.00	0.10	0.21
Self-regulation skills, yr 1	-	-	-	-	-	-	1.00	0.05	0.07
After School Reading									
SAT-10 total reading	-	-	-	0.74	0.17	0.21	1.00	0.17	0.21
After School Math									
SAT-10 total math	-	-	-	0.89	0.20	0.18	1.00	0.18	0.19
Teach for America: Pooled (TFA-Pooled)									
Math	-	-	-	0.94	0.20	0.15	1.00	0.00	0.14
Reading	-	-	-	1.08	0.27	0.15	1.00	0.00	0.16
Tennessee STAR									
SAT-7 total math	-	-	-	-	-	-	1.00	0.21	0.03
SAT-7 total reading	-	-	-	-	-	-	1.00	0.21	0.05
Middle School-High School									
Charter Middle Schools (Charters)									
Math, yr 1	-	-	-	1.06	0.20	0.57	1.00	0.01	0.55
Math, yr 2	-	-	-	1.23	0.21	0.48	1.00	0.00	0.47
Reading, yr 1	-	-	-	0.94	0.22	0.44	1.00	0.00	0.43
Reading, yr 2	-	-	-	1.02	0.19	0.42	1.00	0.00	0.41
Teach for America: Math (TFA-Math)									
Math	-	-	-	0.91	0.31	0.24	1.00	0.08	0.24
Enhanced Reading Opportunities (ERO)									
GRADE Reading comprehension	-	-	-	0.68	0.04	0.13	1.00	0.04	0.13
GRADE Reading vocabulary	-	-	-	0.68	0.05	0.12	1.00	0.05	0.12
Proportion of required credits earned, yr 1	10.1	0.11	0.17	-	-	-	1.00	0.11	0.17
Proportion of required credits earned, yr 2	17.4	0.17	0.14	-	-	-	1.00	0.17	0.14
Small Schools of Choice (SSC)[b]									
On track in ninth grade	50.0	-	-	-	-	-	1.00	-	-
Earned a high school diploma, yr 4	48.9	-	-	-	-	-	1.00	-	-
Career Academies[c]									
Earned HS diploma or equivalent, yr 5	34.2	0.03	0.13	-	-	-	1.00	0.03	0.13
Enrolled in post-secondary	49.7	0.05	0.13	-	-	-	1.00	0.05	0.13
Avg. annual earnings, yrs 1-4	11,884	0.05	0.08	-	-	-	1.00	0.05	0.08
Avg. annual earnings, yrs 5-8	25,913	0.01	0.05	-	-	-	1.00	0.01	0.05
Avg. months worked annually, yrs 1-4	3.3	0.03	0.06	-	-	-	1.00	0.03	0.06
Avg. months worked annually, yrs 5-8	3.4	0.01	0.05	-	-	-	1.00	0.01	0.05

(continued)

Appendix Table A.1. (continued)

	Natural Units (Y)			Standardized Effect Size Units (Z)					
				Reference			Control Group		
Outcome	**σ**	**ρ**	**R^2_{within}**	**σ**	**ρ**	**R^2_{within}**	**σ**	**ρ**	**R^2_{within}**
Early College High School (ECHS)									
On track in ninth grade	31.7	0.11	0.26	-	-	-	1.00	0.11	0.26
Earned a high school diploma, yr 5	37.8	0.02	0.22	-	-	-	1.00	0.02	0.22
Post-Secondary Education									
Learning Communities (LC)									
Targeted credits earned, sem 1	3.0	0.10	0.04	-	-	-	1.00	0.10	0.04
Cumulative targeted credits earned, sem 3	5.1	0.10	0.05	-	-	-	1.00	0.10	0.05
Total credits earned, sem 1	5.8	0.07	0.06	-	-	-	1.00	0.07	0.06
Cumulative total credits earned, sem 3	14.9	0.08	0.06	-	-	-	1.00	0.08	0.06
Performance-Based Scholarships (PBS)									
Cumulative total credits earned, yr 1	10.7	0.09	0.02	-	-	-	1.00	0.09	0.02
Cumulative total credits earned, yr 3	29.0	0.12	0.02	-	-	-	1.00	0.12	0.02
Earned a degree, yr 3	37.9	0.03	0.01	-	-	-	1.00	0.03	0.01
Labor/Workforce									
Job Corps Program									
Total annual earnings, yr 1	9,080	0.01	0.17	-	-	-	1.00	0.01	0.17
Total annual earnings, yr 2	11,982	0.01	0.13	-	-	-	1.00	0.01	0.13
Total annual earnings, yr 3	13,061	0.02	0.11	-	-	-	1.00	0.02	0.11
Total annual earnings, yr 4	15,041	0.01	0.09	-	-	-	1.00	0.01	0.09
Months worked, yr 1	4.2	0.02	0.18	-	-	-	1.00	0.02	0.18
Months worked, yr 2	4.7	0.02	0.11	-	-	-	1.00	0.02	0.11
Months worked, yr 3	4.6	0.02	0.09	-	-	-	1.00	0.02	0.09
Months worked, yr 4	4.7	0.01	0.07	-	-	-	1.00	0.01	0.07
Earned HS diploma or equivalent	47.4	0.02	0.02	-	-	-	1.00	0.02	0.02
Ever arrested	46.5	0.05	0.10	-	-	-	1.00	0.05	0.10
Welfare-to-Work Program (WtW)[d]									
Avg. annual earnings, qtrs 1-8	6,832	0.02	0.14	-	-	-	1.00	0.02	0.14

NOTES: Items marked "-" are unavailable or not relevant.

[a] The HSIS team administered a shorter version of the Peabody Picture Vocabulary Test III (PPVT-III). WJ-III is Woodcock-Johnson III; LW stands for the Letter-Word Identification subscale; OC stands for the Oral Comprehension subscale; AP stands for the Applied Problems subscale. Externalizing behavior problems is based on parent-reported items from the Child Behavior Checklist. Self-regulation skills is based on parent-reported items from the Leiter-R Assessor Report. Some baseline covariates used in the estimation model were imputed by the original study team and non-imputed data are not available for these covariates.

[b] Due to availability of follow-up data for the SSC evaluation, the on track in ninth grade indicator is based on seven cohorts of students whereas the four-year graduation is based on four cohorts.

[c] In Career Academies, earned HS diploma or equivalent and enrollment in post-secondary were measured 14 months after expected HS graduation. Years 1 to 4 earnings and employment capture the first four years after expected HS graduation; Years 5 to 8 earnings and employment capture the next four years.

[d] Mirroring the approach of the original study team, the analysis sample for the WtW analysis was restricted to include only female sample members. Additionally, some offices were excluded due to their small size, unusual client mix, or incomplete data. Baseline covariates used in the estimation model were imputed by the original study team and non-imputed data are not available.

FIGURES

Figure 1. Histogram of Site-level Constrained Empirical-Bayes Impact Estimates - After School Reading Program

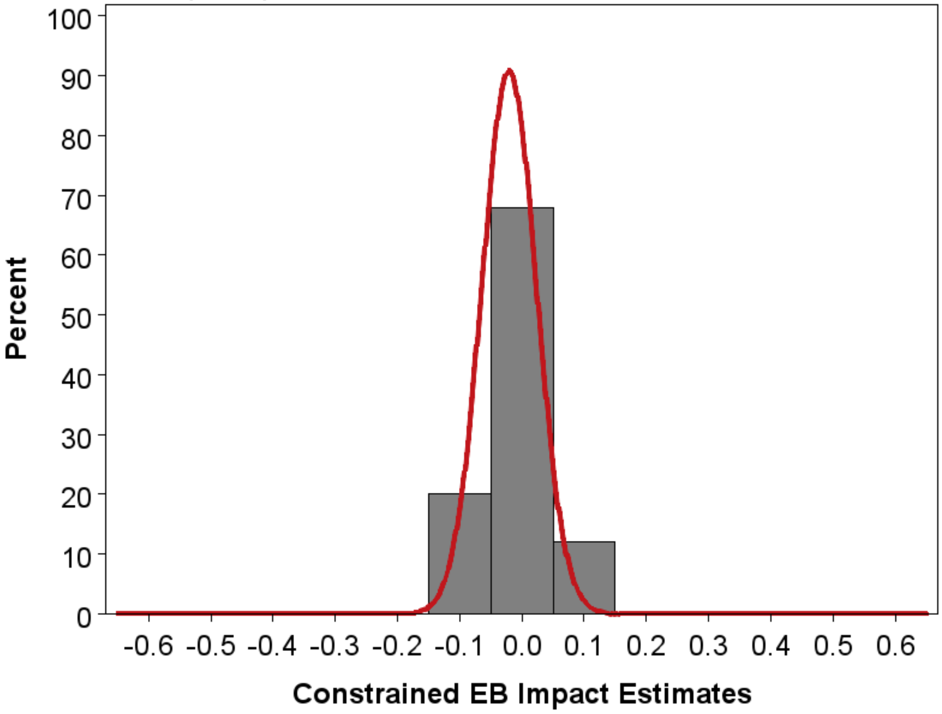


Figure 2. Histogram of Site-level Constrained Empirical-Bayes Impact Estimates – Charter Middle School – Year 2 Reading

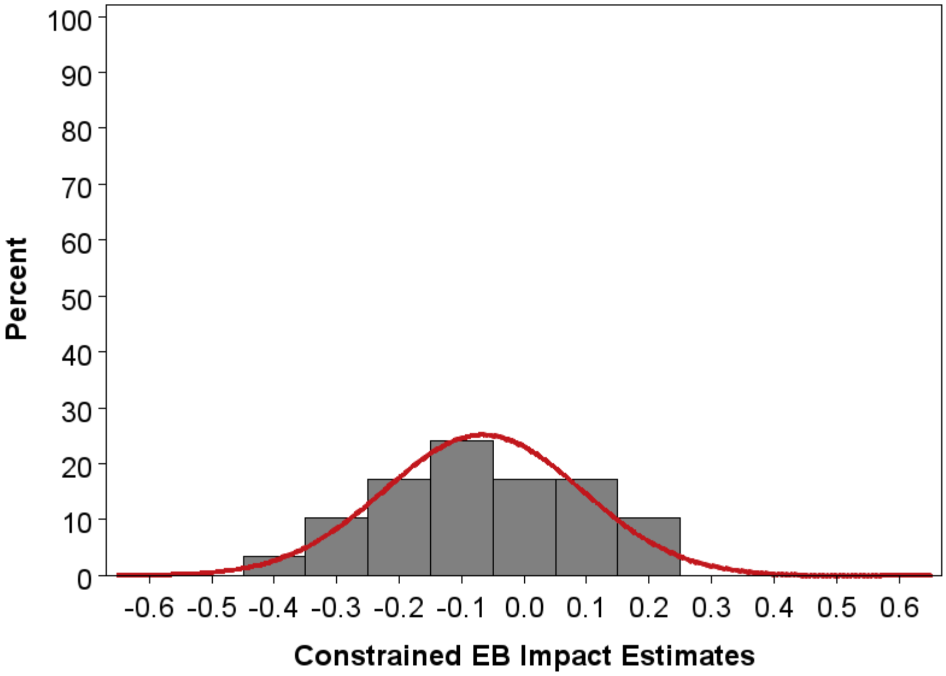
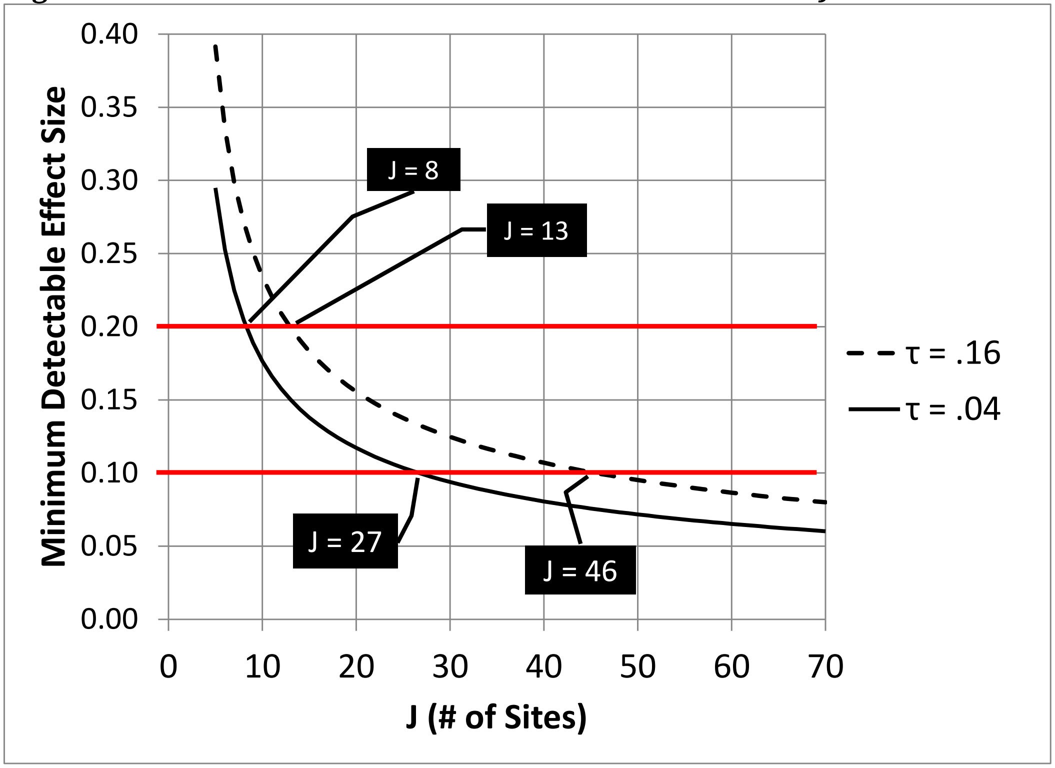


Figure 3. How Minimum Detectable Effect Size Varies with *J* and *τ*



Notes: Assumes α=0.05, Power=0.80, ρ=0.20, R^2_{within}=0.30, σ_Z=1, n=75, and $\bar{T}$=0.50.

References

Abadie, A., Angrist, J., & Imbens, G. (2002). Instrumental Variables Estimates of the Effect of Subsidized Training on the Quantiles of Trainee Earnings. *Econometrica, 70*(1), 91-117. doi: 10.1111/1468-0262.00270

Allcott, H. (2015). Site Selection Bias in Program Evaluation. *The Quarterly Journal of Economics*. doi: 10.1093/qje/qjv015

Bitler, M. P., Gelbach, J. B., & Hoynes, H. W. (2006). What Mean Impacts Miss: Distributional Effects of Welfare Reform Experiments. *American Economic Review, 96*(4), 988-1012. doi: doi: 10.1257/aer.96.4.988

Black, A. R., Somers, M.-A., Doolittle, F., Unterman, R., & Grossman, J. B. (2009). The Evaluation of Enhanced Academic Instruction in After-School Programs Final Report (NCEE 2009-4077). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Bloom, H. S. (2005). *Learning More from Social Experiments: Evolving Analytic Approaches*. New York: Russell Sage Foundation.

Bloom, H. S., Bos, J. M., & Lee, S.-W. (1999). Using Cluster Random Assignment to Measure Program Impacts: Statistical Implications for the Evaluation of Education Programs. *Evaluation Review, 23*(4), 445-469. doi: 10.1177/0193841x9902300405

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., Raudenbush, S. W., Weiss, M. J., & Porter, K. (revise and resubmit). Using Multisite Experiments to Study Cross-site Variation in Effects of Program Assignment.

Bloom, H. S., Richburg-Hayes, L., & Black, A. R. (2007). Using Covariates to Improve Precision for Studies That Randomize Schools to Evaluate Educational Interventions. *Educational Evaluation and Policy Analysis, 29*(1), 30-59.

Bloom, H. S., & Spybrook, J. (under review). Determining Minimum Detectable Cross-site Mean Effect Sizes, Minimum Detectable Cross-Site Variation in Effect Sizes and Minimum Detectable Effect Size Differences for Categories of Sites for Multi-site Trials.

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. doi: 10.1002/pam.21748

Bloom, H. S., & Weiland, C. (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. MDRC Working Papers on Research Methodology: MDRC. 16 East 34th Street 19th Floor, New York, NY 10016-4326. Tel: 212-532-3200; Fax: 212-684-0832; e-mail: publications@mdrc.org; Web site: http://www.mdrc.org.

Bloom, H. S., Zhu, P., Jacob, R., Raudenbush, S., Martinez, A., & Lin, F. (2008). Empirical Issues in the Design of Group-Randomized Studies to Measure the Effects of Interventions for Children. MDRC Working Papers on Research Methodology *MDRC Working Papers on Research Methodology*: MDRC. 16 East 34th Street 19th Floor, New York, NY 10016-4326. Tel: 212-532-3200; Fax: 212-684-0832; e-mail: publications@mdrc.org; Web site: http://www.mdrc.org.

Bryk, A. S., & Raudenbush, S. W. (1988). Heterogeneity of Variance in Experimental Studies: A Challenge to Conventional Interpretations. *Psychological Bulletin, 104*(3), 396-404.

Clark, M. A., Chiang, H. S., Silva, T., McConnell, S., Sonnenfeld, K., & Erbe, A. (2013). The Effectiveness of Secondary Math Teachers from Teach For America and the Teaching Fellows Programs (NCEE 2013-4015). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

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

Cole, S. R., & Stuart, E. A. (2010). Generalizing Evidence From Randomized Clinical Trials to Target Populations: The ACTG 320 Trial. *American Journal of Epidemiology, 172*(1), 107-115. doi: 10.1093/aje/kwq084

Decker, P. T., Mayer, D. P., & Glazerman, S. (2004). The Effects of Teach For America on Students: Findings from a National Evaluation. Discussion Paper no. 1285-04 (pp. 1-82). Princeton, NJ: Mathematica Policy Research.

Djebbari, H., & Smith, J. (2008). Heterogeneous impacts in PROGRESA. *Journal of Econometrics, 145*(1-2), 64-80. doi: 10.1016/j.jeconom.2008.05.012

Dong, N., & Maynard, R. (2013). PowerUp!: A Tool for Calculating Minimum Detectable Effect Sizes and Minimum Required Sample Sizes for Experimental and Quasi-Experimental Design Studies. *Journal of Research on Educational Effectiveness, 6*(1), 24-67. doi: 10.1080/19345747.2012.673143

Edmunds, J., Unlu, F., Glennie, E., Bernstein, L., Fesler, L., Furey, J., & Arshavsky, N. (forthcoming). Smoothing the Transition to Postsecondary Education: The Impact of the Early College Model. *Journal of Research on Educational Effectiveness*.

Friedlander, D., & Robins, P. K. (1997). The Distributional Impacts of Social Programs. *Evaluation Review, 21*(5), 531-553. doi: 10.1177/0193841x9702100501

Gerber, A. S., & Green, D. P. (2012). *Field Experiments: Design, Analysis, and Interpretation*. New York: W.W. Norton.

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

Heckman, J. J. (2001). Micro Data, Heterogeneity, and the Evaluation of Public Policy: Nobel Lecture. *Journal of Political Economy, 109*(4), 673-748. doi: 10.1086/322086

Heckman, J. J., Smith, J., & Clements, N. (1997). Making The Most Out Of Programme Evaluations and Social Experiments: Accounting For Heterogeneity in Programme Impacts. *Review of Economics and Statistics, 64*(4), 487-535. doi: 10.2307/2971729

Hedges, L. V., & Hedberg, E. C. (2014). Intraclass Correlations and Covariate Outcome Correlations for Planning Two- and Three-Level Cluster-Randomized Experiments in Education. *Evaluation Review, 37*(6), 445-489. doi: 10.1177/0193841X14529126

Hedges, L. V., & Pigott, T. D. (2001). The Power of Statistical Tests in Meta-Analysis. *Psychological Methods, 6*(3), 203-2017. doi: 10.1037//1082-989X.6.3.203

Holland, P. W. (1986). Statistics and Causal Inference. *Journal of the American Statistical Association, 81*(396), 945-960.

Jacob, R., Zhu, P., & Bloom, H. S. (2009). New Empirical Evidence for the Design of Group Randomized Trials in Education *MDRC Working Papers on Research Methodology* (pp. 1-48). New York: MDRC. 16 East 34th Street 19th Floor, New York, NY 10016-4326.

Tel: 212-532-3200; Fax: 212-684-0832; e-mail: publications@mdrc.org; Web site: http://www.mdrc.org.

Jones, M. P. (1996). Indicator and Stratification Methods for Missing Explanatory Variables in Multiple Linear Regression. *Journal of the American Statistical Association, 91*(433), 222-230. doi: 10.2307/2291399

Kemple, J. J. (2001). Career Academies: Impacts on Students' Initial Transitions to Post-Secondary Education and Employment. MDRC, 16 East 34th Street, New York, NY 10016. Tel: 212-532-3200. For fulltext: http://www.mdrc.org/sites/default/files/full_47.pdf.

Kemple, J. J. (2008). Career Academies: Long-Term Impacts on Labor Market Outcomes, Educational Attainment, and Transitions to Adulthood *MDRC Report*. New York: MDRC.

Mayer, A. K., Patel, R., Rudd, T., & Ratledge, A. (2015). Designing Scholarships to Improve College Success *MDRC Report* (pp. 41). New York: MDRC.

Nye, B., Hedges, L. V., & Konstantopoulos, S. (2000). The Effects of Small Classes on Academic Achievement: The Results of the Tennessee Class Size Experiment. *American Educational Research Journal, 37*(1), 123-151. doi: 10.3102/00028312037001123

Nye, B., Konstantopoulos, S., & Hedges, L. V. (2004). How Large Are Teacher Effects? *Educational Evaluation and Policy Analysis, 26*(3), 237-257.

Olsen, R. B., Orr, L. L., Bell, S. H., & Stuart, E. A. (2013). External Validity in Policy Evaluations That Choose Sites Purposively. *Journal of Policy Analysis and Management, 32*(1), 107-121. doi: 10.1002/pam.21660

Puma, M. J., Olsen, R. B., Bell, S. H., & Price, C. (2009). What to Do when Data Are Missing in Group Randomized Controlled Trials. (Vol. NCEE 2009-0049.). Washington DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Raudenbush, S. W. (2009). Adaptive Centering with Random Effects: An Alternative to the Fixed Effects Model for Studying Time-varying Treatments in School Settings. *American Education Finance Association*, 468-491.

Raudenbush, S. W. (2015). *Estimation of Means and Covariance Components in MultisiteTrials*. University of Chicago. Unpublished Manuscript.

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

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods* (2^{nd} ed.). Newbury Park, CA: Sage Publications.

Raudenbush, S. W., & Liu, X. (2000). Statistical Power and Optimal Design for Multisite Randomized Trials. *Psychological Methods, 5*(2), 199-213.

Reardon, S. F., & Raudenbush, S. W. (2013). Under What Assumptions Do Site-by-Treatment Instruments Identify Average Causal Effects? *Sociological Methods & Research, 42*(2), 143-163. doi: 10.1177/0049124113494575

Rothwell, P. M. (2005). Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. *The Lancet, 365*, 176-186.

Schochet, P. Z., Burghardt, J., & Glazerman, S. (2001). National Job Corps Study: The Impacts of Job Corps on Participants' Employment and Related Outcomes. Princeton, NJ: Mathematica Policy Research.

Schochet, P. Z., Burghardt, J., & McConnell, S. (2008). Does Job Corps Work? Impact Findings From the National Job Corps Study. *The American Economic Review, 98*(5), 1864-1886.

Schochet, P. Z., Puma, M., & Deke, J. (2014). Understanding Variation in Treatment Effects in Education Impact Evaluations: An Overview of Quantitative Methods *NCEE Report*. Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Analytic Technical Assistance and Development.

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. *National Center for Education Evaluation and Regional Assistance*.

Spybrook, J. (2014). An Introduction to a Special Issue on Design Parameters for Cluster Randomized Trials in Education. *Evaluation Review*. doi: 10.1177/0193841x14527758

Spybrook, J., Bloom, H. S., Congdon, R., Hill, C. J., Martinez, A., & Raudenbush, S. W. (2011, October 16, 2011). *Optimal Design Plus Empirical Evidence: Documentation for the "Optimal Design" Software Version 3.0.*

Stuart, E. A., Cole, S. R., Bradshaw, C. P., & Leaf, P. J. (2011). The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 174*(2), 369-386. doi: 10.1111/j.1467-985X.2010.00673.x

Tipton, E. (2013a). Improving Generalizations From Experiments Using Propensity Score Subclassification: Assumptions, Properties, and Contexts. *Journal of Educational and Behavioral Statistics, 38*(3), 239-266.

Tipton, E. (2013b). Stratified Sampling Using Cluster Analysis: A Sample Selection Strategy for Improved Generalizations From Experiments. *Evaluation Review, 37*(2), 109-139. doi: 10.1177/0193841x13516324

Tipton, E., Yeager, D., Schneider, B., & Iachan, R. (forthcoming). *Designing Probability Samples to Identify Sources of Treatment Effect Heterogeneity*.

Weiss, M. J. (2010). The Implications of Teacher Selection and the Teacher Effect in Individually Randomized Group Treatment Trials. *Journal of Research on Educational Effectiveness, 3*(4), 381-405.

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).

Weiss, M. J., Visher, M. G., Weissman, E., & Wathington, H. (2015). The Impact of Learning Communities for Students in Developmental Education: A Synthesis of Findings From Randomized Trials at Six Community Colleges. *Educational Evaluation and Policy Analysis*. doi: 10.3102/0162373714563307

Westine, C. D., Spybrook, J., & Taylor, J. A. (2014). An Empirical Investigation of Variance Design Parameters for Planning Cluster-Randomized Trials of Science Achievement. *Evaluation Review, 37*(6), 490-519. doi: 10.1177/0193841X14531584

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

Xu, Z., & Nichols, A. (2010). New Estimates of Design Parameters for Clustered Randomization Studies: Findings from North Carolina and Florida *CALDER working papers* (pp. 1-64):

CALDER, The Urban Institute 2100 M Street N.W., Washington, D.C. 20037 202-261-5739 • www.caldercenter.org.

Yusuf, S., Wittes, J., Probstfield, J., & Tyroler, H. A. (1991). Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials. *JAMA, 266*(1), 93-98. doi: 10.1001/jama.1991.03470010097038