Article

American Journal of Evaluation
1-25

DOI: 10.1177/1098214015600515
aje.sagepub.com
SAGE

Learning About and From a Distribution of Program Impacts Using Multisite Trials

Stephen W. Raudenbush[1] and Howard S. Bloom[2]

Abstract

The present article provides a synthesis of the conceptual and statistical issues involved in using multisite randomized trials to learn about and from a distribution of heterogeneous program impacts across individuals and/or program sites. Learning *about* such a distribution involves estimating its mean value, detecting and quantifying its variation, and estimating site-specific impacts. Learning *from* such a distribution involves studying the factors that predict or explain impact variation. Part I of the article introduces the concepts and issues involved. Part II focuses on estimating the mean and variation of impacts of program *assignment*. Part III extends the discussion to variation in the impacts of program *participation*. Part IV considers how to use multisite trials to study *moderators* of program impacts (individual-level or site-level factors that influence these impacts) and *mediators* of program impacts (individual-level or site-level "mechanisms" that produce these impacts).

Keywords

causal inference, impact evaluation, multilevel evaluation, PreK–12 education, public policy, quantitative methods

Using Multisite Randomized Trials to Learn *About* and *From* a Distribution of Program Impacts

To make a valid causal statement about the impacts of a new reading program, drop-out prevention program or job training initiative, measuring the gains made by program participants is not enough. Estimating how participants would have fared without the program is also necessary. This requires a valid comparison group of nonprogram participants who were similar in all predictors of the outcome (measured or unmeasured) to participants at the outset of the study. The aim is to compare future outcomes for program participants and comparison group members under the assumption that

[1] University of Chicago, Chicago, IL, USA
[2] MDRC, New York, NY, USA

Corresponding Author:
Howard S. Bloom, MDRC, 16 East 34th Street, 19th floor, New York, NY 10016, USA.
Email: howard.bloom@mdrc.org

comparison group outcomes reflect what ''would have happened'' to program participants without the program.

To create a valid comparison group for testing new medical treatments, scientists embraced the randomized controlled trial (RCT) soon after World War II. This strategy, which had its origins in agricultural research during the early 20th century (Fisher, 1925), was to randomly assign persons to a treatment group or a nontreatment ''control group'' in order to create two statistically equivalent groups. Mean future control group outcomes then provide unbiased estimates of what mean future treatment group outcomes would have been without the treatment. Exploiting this singular strength, government agencies and other funders have recently sponsored many RCTs to evaluate social and educational programs, policies, and practices (Greenberg & Shroder, 2004; Spybrook, 2013). As a result, we are now learning much about the effectiveness of preschool education, charter schools, remedial math and reading interventions, after-school services, teacher professional development, career academies, job training programs, social service programs, criminal justice programs, and more.

Researchers and research funders met in two recent conferences to review the design and analysis of RCTs for social and educational program evaluation (*Learning from Variation in Program Effects* sponsored by the William T. Grant Foundation and the conference that inspired this forum). Participants at these conferences noted that past RCTs focused mainly on estimating the *average* impacts of new programs. Although an average impact is an essential inferential target of any RCT, participants reasoned that the average is not sufficient by itself for developing public policy, professional practice, or program theory when program impacts are heterogeneous. Participants thus agreed that better understanding of how and why program impacts vary is needed; in other words, we need to learn more about and from the *distribution* of program impacts.

Learning About *a Distribution of Program Impacts*

Learning about a distribution of program impacts involves estimating the mean value of this impact, quantifying its variation around the mean, assessing the equity of this variation, and studying site-specific impacts.

Estimating mean impacts. Although it is common practice to estimate mean program impacts, the conceptual and statistical issues involved in doing so for multisite trials with heterogeneous impacts are more subtle than is typically acknowledged. Specifically, if impacts can vary across individuals and sites, multiple possible definitions of the overall mean exist. For example, a researcher might want to know the mean impact for the population of program-eligible *persons* represented by the study sample or the researcher might want to know the mean impact for the population of program *sites* represented by the sample.

Quantifying impact variation. If persons or sites vary widely in their responses to a program, the overall average program impact is not useful for policy makers who might contemplate adopting the program or for practitioners who want to know how to improve it. Hence, knowing about the extent to which program impacts vary is essential for informing appropriate use of the RCT results.

Assessing impact equity. In multisite trials, the cross-site correlation between program impacts and control group mean outcomes can be studied. If sites that serve individuals who would do especially poorly without the new program produce above-average impacts, this suggests that the program will tend to reduce inequality. If program sites that serve individuals who would do especially well without the program produce above-average impacts, this suggests that the program will tend to increase inequality.

Studying site-specific impacts. We can also use multisite RCTs to produce site-specific estimates of program impacts. We can thus quantify the effectiveness of the most and least effective sites. Knowing how effective a program *can be* is as important as knowing how effective it is on average, especially if one learns from best practice at effective sites how to improve performance at ineffective sites.

Learning From *a Distribution of Program Impacts*

Having learned about a distribution of program impacts, much can be learned *from* this distribution. Our point is that impact heterogeneity creates opportunities for testing theories about impact moderation and mediation.

Moderation of impacts. Program impacts vary because some types of persons are more likely than others to participate, because staff at some sites are more skilled than staff at other sites, or because existing services from outside of a program are more available and/or effective at some sites than at others. These factors are potential *moderators* of program impacts. More specifically, we define impact moderators as characteristics of clients or sites that (1) cannot be influenced by the program being tested and (2) facilitate or inhibit a program's effectiveness.

To explore potential moderators, evaluators often conduct impact analyses for sample subgroups defined in terms of factors such as gender, ethnicity, social background, and risk of failure. It is less common to find an evaluation that is founded on an explicit moderation *theory* about who is likely to benefit the most or the least from the program being studied and what organizational conditions are most important for its success. Testing such theories may significantly increase the utility of evaluations for future program design and practice.

Mediation of impacts: *mediators* (or mechanisms) of program impacts are those aspects of program implementation, staff practice, and short-term changes in participants' knowledge, skills, attitudes, or behavior that are (1) outcomes of random assignment and (2) predictors of long-term success.

In theory, sites with larger-than-average effects on program mediators will produce greater-than-average impacts on participant outcomes. Thus, heterogeneity of a program's effects on its mediators can explain heterogeneity of impacts on participants' outcomes. Nonetheless, most programs are founded on some theory about how program operations influence key mediators and produce long-term outcomes, but few rigorous evaluations explicitly test these theories, and impact heterogeneity is largely unexplained.

We allow for the possibility that treatment assignment can moderate the effect of treatment mediators. For example, assignment to a new job training program might increase participants' motivation to work, thereby mediating the program's impact on employment. In addition, program assignment might change the effect of motivation on employment.

The Importance of Multisite Trials for Studying a Distribution of Program Impacts

We focus here on multisite trials in which sample members are randomly assigned to a program or a control group *within* each of a number of sites. Sometimes sites are comparatively few in number, like the Moving to Opportunity (MTO) experiment conducted in five major U.S. cities (Katz, Kling, & Liebman, 2000). Other times, RCTs have many sites, like the national Head Start Impact Study, which was conducted in 350 Head Start centers from across the U.S. (Bloom & Weiland, 2015).

Although the research questions addressed and the statistical methods used depend on the number of sites and participants per site, all multisite trials represent "a fleet of randomized experiments." Hence, they are well suited for studying mean program impact and impact heterogeneity.

Moreover, multisite trials are prevalent, if not ubiquitous. For example, Spybrook (2013) found that more than two thirds of the 175 RCTs conducted by The Institute of Education Sciences since 1994 are multisite trials.

The Present Article

This article summarizes issues that arise and available options to consider when using multisite trials to study a distribution of program impacts. We recommend analytic approaches for addressing the issues, and we also identify new methodological frontiers as targets for future research. We now turn to focus on using multisite trials to study a distribution of impacts of program assignment (impacts of "intent to treat" [ITT]). Then we extend this discussion to impacts of program participation (complier average causal effects [CACE]). Finally, we consider moderators and mediators of program effects.

Learning *About* a Distribution of ITT Impacts

To lay a conceptual and methodological foundation, we begin with an individual-level distribution of ITT impacts *in a single site*. We then discuss how to use multisite RCTs to study a distribution of ITT impacts *across multiple sites*.

Studying the Distribution of ITT Impacts Across Individuals in a Single-Site RCT

The present discussion adopts the "potential outcomes" framework for causal inference, which is used widely in applied statistics.[1] We set $T = 1$ if a sample member is randomized to a new program (or treatment) and $T = 0$ if a sample member is randomized to a control group. Each individual has two potential outcomes: $Y(1)$ if the participant is assigned to the program and $Y(0)$ if the participant is assigned to the control group.[2] The causal effect of program assignment for an individual is the *difference* between his or her two potential outcomes:

$$B \equiv Y(1) - Y(0). \tag{1}$$

It is not possible to calculate an ITT impact for an individual because we can observe only one of his two potential outcomes. We can observe $Y(1)$ if the participant is assigned to the program group or $Y(0)$ if the participant is assigned to the control group. Although we cannot estimate person-specific impacts, we can estimate the *average* ITT impact for the site population of individuals under a key assumption that a person's potential outcomes do not predict treatment group assignment. Random assignment enables us to meet this assumption, so that, in an RCT, we can readily estimate a population average causal effect of program assignment or ITT (β_{ITT}):

$$\beta_{ITT} \equiv E[Y(1) - Y(0)] = E[Y(1)] - E[Y(0)], \tag{2}$$

where E denotes an "expectation" or population average. In other words, β_{ITT} equals the *difference* between the average outcome if the entire population were assigned to the program ($E[Y(1)]$) and the average outcome if the entire population were assigned to the control group ($E[Y(0)]$). We can use data from persons assigned to the treatment group ($T = 1$) to estimate how the entire population would fare, on average, if it were assigned to the program, that is, $E[Y(1)$, because, in an RCT, persons assigned to the treatment group are statistically representative of the entire population of interest. Similarly, the RCT enables us to use the data from persons assigned to the control group ($T = 0$) to estimate how the entire population would fare, on average, if assigned to the control condition. To do so we require that assigning the entire population to one of the two groups would not change the potential outcomes of individuals.[3]

Although we can estimate a population average impact from an RCT under mild assumptions, we cannot readily estimate the variance of ITT effects across individuals. To see this, note that, based on Equation 1:

$$Y(1) = Y(0) + B. \tag{3}$$

Hence, the variance of $Y(1)$ is:

$$\text{Var}[Y(1)] = \text{Var}[Y(0)] + \text{Var}[B] + 2 \cdot \text{Cov}[Y(0), B], \tag{4}$$

which implies that:

$$\text{Var}[Y(1)] - \text{Var}[Y(0)] = \text{Var}[B] + 2 \cdot \text{Cov}[Y(0), B], \tag{5}$$

where Cov[$Y(0)$, B] is the individual-level *covariance* between control group outcomes and program impacts. Although we can estimate the two variances Var[$Y(1)$] and Var[$Y(0)$] from sample data, we cannot estimate Cov[$Y(0)$, B] or Var(B) because we cannot observe both potential outcomes for individuals.

Further investigation (Bloom, Raudenbush, Weiss, & Porter, 2014; Bryk & Raudenbush, 1988) reveals that:

- If a program group and a control group have different individual-level outcome variances, we can conclude that ITT impacts vary across individuals.[4]
- If a program group and a control group do not have different individual-level outcome variances, we cannot conclude that ITT impacts do not vary across individuals.[5]
- If the program group variance is smaller than the control group variance, we can conclude that the program produces larger-than-average ITT impacts for persons who would fare worse than average without the program (i.e., program effects are compensatory).[6]
- If the program group variance is larger than the control group variance, however, we cannot conclude that the program produces larger-than-average impacts for persons who would fare better than average without the program.[7]

In summary, a single-site RCT provides full information about mean impact of program assignment at a single site and limited information about heterogeneity of this impact across individuals at that site.

Studying a Distribution of ITT Impacts Across Sites in a Multisite RCT

We now consider multisite analyses of the distribution of program impacts. We are interested in the mean of this distribution, the cross-site variation around the mean, and the cross-site correlation between program impacts and control group mean outcomes. In addition, we want to estimate site-specific impacts.

Consider first the case of a population mean ITT impact. Estimating mean impact is simple—*if* we assume impacts to be constant across sites. However, given the heterogeneity of organizational conditions and populations served across sites in many RCTs, the assumption of a constant impact seems implausible. In this case, defining and estimating a mean program impact can be tricky.

Defining a population mean impact. When program impacts vary across persons and/or sites, different ways to define a population mean impact exist. On one hand, we might want to generalize findings to a population of *sites* (e.g., we might want to know the mean of the mean Head Start impacts for all Head Start centers in the United States). Or, we might want to generalize findings to a population of *persons* (e.g., the mean Head Start impact for the national population of program-eligible children).

Statisticians often define a parameter of interest as a "target of inference" or "estimand." Ideally, researchers should be explicit about their estimands before designing a study. For example, suppose that prior to designing a study, we have information about the number of sites (J^* in our population of interest), and we also have information about the number of eligible persons (N_j in each site) j, there being $\sum_{j=1}^{J^*} N_j$ persons in the entire population. Each person i in each site j possesses a potential outcome $Y_{ij}(1)$ if assigned to the program and a potential outcome $Y_{ij}(0)$ if assigned to the control group. The ITT impact (B_{ij}) for each person i in site j is thus:

$$B_{ij} = Y_{ij}(1) - Y_{ij}(0). \tag{6}$$

The subpopulation mean ITT impact (B_j) for persons in site j is:

$$B_j = \sum_{i=1}^{N_j} B_{ij}/N_j. \tag{7}$$

If we wish to generalize to a population of sites, we define our estimand as the simple *mean of the site mean impacts* (β_{sites}), that is:

$$\beta_{\text{sites}} = \sum_{j=1}^{J^*} B_j/J^*. \tag{8}$$

If we wish to generalize to a population of persons, we define our estimand as the following *person-weighted mean of the site mean program impacts* (β_{persons}):

$$\beta_{\text{persons}} = \frac{\sum_{j=1}^{J^*}\sum_{i=1}^{N_j} B_{ij}}{\sum_{j=1}^{J^*} N_j} = \frac{\sum_{j=1}^{J^*} N_j B_j}{\sum_{j=1}^{J^*} N_j}. \tag{9}$$

If site-specific impacts are homogeneous, the site-average mean impact in Equation 8 will equal the person-average mean impact in Equation 9. Similarly, if all the sites have the same population size and the same fraction of persons assigned to treatment, the two estimates will also be equal. Otherwise, the estimands may differ from each other. For example, if programs in sites with large client populations are more effective than programs in sites with small client populations, the two population mean impacts will differ.

Designing a multisite trial. The choice of an estimand can influence the optimal design of a study. To see how, assume for simplicity that the cost of sampling children within sites and collecting data on program members and control group members is constant and that the individual outcome variance in the treatment and control groups is the same.

If the estimand of interest is β_{sites}, it is optimal first to (1) draw a simple random sample of sites from the population of sites, (2) draw a simple random sample of n persons from each site, and (3) assign persons from each site with equal probability to the program group or control group. These conditions produce a perfectly balanced design with $n/2$ persons from each experimental group per site. However, if the estimand of interest is β_{persons}, a good option is to draw a simple random sample of sites from the population of sites and set each site's sample size proportional to its number of program-eligible persons.[8]

Unfortunately, evaluators can rarely implement a probability sample of sites or persons and usually must select a sample of convenience. However, they can conceive of their study sites as representing a larger population of similar sites that might use the program, and they typically want their findings to apply to persons who might benefit if the program is found to be effective. Even in this setting, one must take care when choosing an estimand. For example, if one wanted to generalize findings to a population of *sites* β_{sites} should be used. If instead, one wanted to generalize findings to a population of persons β_{persons} should be used.[9]

Estimating a mean ITT impact. Having carefully defined an estimand of interest and designed a study accordingly, we determine how to estimate the desired mean impact. In so doing, we must confront the fact that site sample size (n_j), and the fraction of sample members randomized to the treatment ($\bar{T}_j$) will typically vary across sites. The fraction assigned to treatment is known as a "propensity score" (Rosenbaum & Rubin, 1983). In a multisite trial, propensity scores can vary across sites by design or, more often, because of unobserved site differences. For example, in a lottery-based study of charter schools, a highly popular charter school might have many applicants per available seat. For this school, the propensity score—that is, the chance of winning its lottery—is low. A less popular charter school might have fewer applicants per seat and thus have a higher propensity score. If these propensity scores are correlated with charter school impacts—which seems possible—one must take special care to account for the correlation.

To see how these challenges play out in practice, we need additional notation. Paralleling our discussion of potential outcomes for an individual, we define U_{1j} as the average outcome that would occur if the entire population of eligible persons in site j were assigned to the new program, and we define U_{0j} as the average outcome that would occur if the entire population at site j were assigned to the program's control group. The average impact of the new program at site j is thus $B_j = U_{1j} - U_{0j}$. If persons are randomly assigned to the program, we can estimate U_{1j} for site j without bias from the sample mean outcome (call it $\bar{Y}_{1j}$) for its program group. Similarly, we can estimate U_{0j} for site j from the sample mean outcome (call it $\bar{Y}_{0j}$) for its control group. The resulting estimate of the average program impact for site j is a simple difference of means $\hat{B}_j = \bar{Y}_{1j} - \bar{Y}_{0j}$, and its sampling variance (call it V_j) depends on the site's sample size and propensity score. These simple facts enable us to evaluate the bias associated with common estimators of alternative estimands.

To keep this discussion simple, we confine our attention to the case where the unweighted "mean of site means" defined by β_{sites} in Equation 8 is our estimand of interest. However, the logic of our inquiry would remain the same if we had focused on the estimand β_{persons} in Equation 9.

The "site fixed-effects" estimator. Perhaps the most common analytic strategy for estimating an average ITT effect for a multisite trial is the site fixed effects estimator. This estimator is obtained from the following regression model, where Y_{ij} is the outcome, T_{ij} is treatment assignment, α_j is a site fixed effect, and e_{ij} is a random error with zero mean, and, for simplicity, a constant variance (σ^2):[10]

$$Y_{ij} = \beta T_{ij} + \alpha_j + e_{ij}. \tag{10}$$

The resulting estimator is (Raudenbush, 2014) equivalent to the following weighted average of site-specific impact estimates $\hat{B}_j = \bar{Y}_{1j} - \bar{Y}_{0j}$:

$$\hat{\beta}_{\text{FE}} = \sum_{j=1}^{J} w_j \hat{B}_j \Big/ \sum_{j=1}^{J} w_j, \tag{11}$$

where $w_j = n_j \bar{T}_j (1 - \bar{T}_j)$. Here n_j is the sample size for site j and $\bar{T}_j$ is the propensity score (the proportion of sample members assigned to the program) for site j. Interestingly, the weight w_j for site j is

proportional to the reciprocal of the sampling variance of the site-specific impact estimates $\hat{B}_j$.[11] This estimator is optimal (it is unbiased and has minimum variance) when mean program impacts do not vary across sites.

When site impacts vary, things change. Now $\hat{\beta}_{FE}$ can be a biased estimate of β_{sites} if "true" site-specific impacts B_j are correlated with site weights w_j (see Raudenbush, 2014). This implies that if the sample size n_j or the propensity score $\bar{T}_j$ is statistically associated with B_j, we risk bias.

A simple average. Naturally, one may think that we can greatly simplify the preceding problem using a straightforward average of site-specific impact estimates. For this case, in which we are generalizing to a population of sites, consider the simple unweighted average estimator:

$$\hat{\beta}_{\text{unweighted}} = \sum_{j=1}^{J} \hat{B}_j / J. \tag{12}$$

This simple average is unbiased when we want to count all sites equally. However, it becomes imprecise when we give sites with very small samples equal importance to sites with very large samples.

A fixed-intercept random-coefficient estimator. Selecting between a site fixed effects estimator $(\hat{\beta}_{FE})$ and a simple unweighted site average estimator $(\hat{\beta}_{\text{unweighted}})$ creates a forced choice that many analysts find objectionable. Does a flexible alternative that is on a continuum between these two extremes and is sensible across a range of cross-site variation in impacts and sample sizes exist? The answer is a qualified "yes."

Consider a hierarchical linear model (HLM; Dempster, Rubin, & Tsutakawa, 1981; Lindley & Smith, 1972; Raudenbush & Bryk, 2002), which specifies site impacts that vary randomly around a population grand mean (β) with a variance τ^2 and removes cross-site differences in mean untreated (control group) outcomes by including a series of site-specific intercepts or fixed effects (α_j) as in Equation 10.[12] If τ^2 were known for the population of sites of interest and V_j were known for each site, we would have an estimator with site weights equal to the reciprocal of the *total* variance $\tau^2 + V_j$ of their site impact estimate.[1] This estimator has the same form as the fixed-effects estimator but with site-specific weights:

$$w_j = (\tau^2 + V_j)^{-1} = \left[\tau^2 + \frac{\sigma^2}{n_j \bar{T}_j (1 - \bar{T}_j)}\right]^{-1}. \tag{13}$$

When site-specific impacts are homogeneous ($\tau^2 = 0$), these weights are the same as those for the site fixed-effects estimator $\left(w_j = n_j \bar{T}_j (1 - \bar{T}_j)\right)$ in Equation 11, which is optimal for homogeneous impacts and heterogeneous site sample sizes. If site impacts are highly heterogeneous (relative to their sampling variances), $w_j \approx 1$, corresponding to the unweighted average estimator in Equation 12, which is optimal in this case. Thus, incorporating the "heterogeneity parameter" τ^2 into our weights creates a continuum of estimators that lie between the two preceding extremes (the site fixed-effects estimator of Equation 11 and the unweighted average of Equation 12).

Recall our answer to the question about a solution to the preceding dilemma was a "qualified" yes, as an important qualification exists. The previous paragraph's reasoning was based on the assumption that τ^2 for the population of sites and V_j for each site are *known*. The unknown part of V_j is the within-site variance σ^2, which can be estimated with considerable precision based on pooled data for even a moderately large RCT. However, precise estimation of τ^2 depends on the number of sites in the RCT. If τ^2 is estimated imprecisely, we will not likely land on the optimal place on the continuum between the fixed-effects estimator and the unweighted estimator. However, we will not land outside this continuum and estimator. Equation 13 can be computed using now-standard software for HLMs.

The Cross-Site Variance of ITT Impacts and the Cross-Site Covariance or Correlation Between ITT Impacts and Control Group Mean Outcomes

Defining a cross-site ITT impact variance. We've made the argument that, for multisite trials, the cross-site variance (or standard deviation) of mean program impacts as well as the cross-site mean should be estimated. But how should we do this?

First, we need to be careful in defining our estimand. An intuitively appealing definition of this cross-site variance is the mean squared discrepancy between site-specific impacts B_j and the unweighted cross-site mean impact β. This variance may be written as:

$$\tau^2_{\text{sites}} = \sum_{j=1}^{J^*} (B_j - \beta_{\text{sites}})^2 / J. \tag{14}$$

However, we may also be interested in a person-weighted average like:

$$\tau^2_{\text{persons}} = \sum_{j=1}^{J^*} N_j \left(B_j - \beta_{\text{persons}} \right)^2 \Big/ \sum_{j=1}^{J^*} N_j. \tag{15}$$

It may seem counterintuitive to define a variance as a person-weighted average. However, Equation 15 can be useful in characterizing the extent to which site differences explain person-specific variation in response to an intervention.

Estimating a cross-site ITT impact variance. Few studies have attempted to estimate a cross-site variance of ITT impacts, and we have not found literature providing guidance for doing so. Clearly, the optimal method depends on the estimand of interest and the study design. However, a broad class of weighted estimators (τ^2_w) will have the following form:

$$\hat{\tau}^2_w = \sum_{j=1}^{J} w_j \left[\left(\hat{B}_j - \hat{\beta} \right)^2 - \hat{V}_j \right] \Big/ \sum_{j=1}^{J} w_j, \tag{16}$$

where w_j is a weight for each site's contribution to the variance estimate. The idea here is that $(\hat{B}_j - \hat{\beta})^2 = [(\hat{B}_j - B_j) + (B_j - \hat{\beta})]^2$ which gives us, for each site, an unbiased estimate of the estimator's sampling variance $V_j = \text{Var}(\hat{B}_j - B_j)$ *plus* the true cross-site impact variance $\tau^2 = \text{Var}(B_j - \beta)$. Thus, we subtract the estimated sampling variance $\hat{V}_j$ from $(\hat{B}_j - \hat{\beta})^2$ to obtain an estimate of τ^2 for each site. We compute a weighted average of these site-specific estimates using a weight w_j. If this result is negative, we set the value $\hat{\tau}^2_w$ equal to zero.

Suppose we have a convenience sample, but regard our sites as representing an interesting if undefined universe of similar sites and our estimand is $\hat{\tau}^2_w$. Then, we'd be inclined to set $w_j = 1$ for Equation 16.[13]

This estimate will be "consistent," that is, it will converge to the correct value as the number of sites in the sample becomes ever larger. However, it might be very imprecise, particularly if small sites produce outlying estimates $(\hat{B}_j - \hat{\beta})^2 - \hat{V}_j$ because these outliers are given weights that are equal to the weights for far more precise estimates from much larger sites.

An alternative to the preceding simple average estimator is an HLM analysis based on maximum likelihood. Such an approach uses iteratively reestimated least squares to obtain, at iteration $m + 1$:

$$\hat{\tau}^{2(m+1)} = \sum_{j=1}^{J} w_j^{(m)} \left[\left(\hat{B}_j - B^{(m)} \right)^2 - \hat{V}_j^{(m)} \right] \Big/ \sum_{j=1}^{J} w_j^{(m)}, \tag{17}$$

where $w_j^{(m)} = (\tau^{2(m)} + V_j^{(m)})^{-2}$. Here, the weight is inversely proportional to the reciprocal of the *square* of the error variance of the site-specific estimate $\hat{B}_j$. This iterative estimator is optimal when

we assume no correlation between weight and site-specific impact estimates because it appropriately weights down outliers from small sample sites. However, if the true cross-site impact variance is large relative to site-specific estimation error, the HLM estimator will tend to converge to the unweighted estimator. Although Equation (17) can be estimated using now-standard packages for HLM, we need to learn more about the bias–precision trade-off that can arise in practice from this approach.[14]

Estimating a cross-site covariance or correlation between ITT impacts and control group mean outcomes. What is the cross-site covariance or correlation between program impacts and control group mean outcomes? This question is rarely addressed empirically, but the answer could be potentially informative. If the cross-site correlation between program impacts and control group mean outcomes is positive (i.e., sites with higher-than-average program impacts tend to have higher-than-average control group mean outcomes), this suggests that the program being tested will increase cross-site outcome inequality. However, if this correlation is negative, the program will tend to reduce cross-site outcome inequality. The influence of this correlation on the overall distribution of outcomes across all population members can be estimated without imposing strong theory or assumptions. Yet conventional methods do not provide a consistent estimate of this correlation.

Again, selecting an estimand is important. To estimate the covariance, suppose that we knew the true mean impact (β_{sites}) for our population of sites and each site-specific impact B_j. Suppose we also knew the true mean untreated counterfactual outcome (μ_0) for the population of sites and for each site U_{0j}. Then, for each site, we could compute the product $(B_j - \beta_{\text{sites}})(U_{0j} - \mu_0)$, which we could then average across sites. In practice, we might substitute corresponding sample estimates to compute $(\hat{B}_j - \hat{\beta})(\hat{U}_{0j} - \hat{\mu}_0)$ and then subtract the sampling covariance $(\hat{C}_j)$ to obtain a weighted average:

$$\hat{\tau}_{B0} = \sum_{j=1}^{J} w_j \left[\left(\hat{B}_j - \hat{\beta} \right) \left(\hat{U}_{0j} - \hat{\mu}_0 \right) - \hat{C}_j \right] \Big/ \sum_{j=1}^{J} w_j \tag{18}$$

where τ_{B0} is the cross-site *covariance* between mean program impacts and mean counterfactual untreated outcomes.[15] Again our choices depend upon our estimand of interest. For example, we could set $w_j = 1$ or $w_j = n_j$ or we could use maximum likelihood estimation of an HLM (although the latter is more complicated and beyond the scope of this discussion). More needs to be learned about how these methods work in practice.

Studying site-specific ITT impacts. If we could observe the impact B_j for each site, we could display the cross-site impact distribution and determine, for example, the 10th, 25th, 75th, or 90th percentile values of this distribution. The problem is that we cannot observe the true values of B_j. To address this problem, we might use our estimate $\hat{B}_j$. Unfortunately, doing so can grossly exaggerate the cross-site impact variation. This issue arises because the cross-site distribution of conventional ordinary least square impact *estimates* $\hat{B}_j$ reflects *two* sources of variation: (1) cross-site variation in true impacts (τ^2) and (2) cross-site variation in impact estimation error (V_j). This problem can also cause us to exaggerate how effective or ineffective the program is at its most and least effective sites. Furthermore, the problem can cause us to misrepresent the rank order of impacts for different sites because sites with the smallest samples will have the largest sampling error and thus tend to have the most extreme positive and negative impact estimates.

Perhaps the most popular way to address this problem is to compute, for each site, an "empirical Bayes," estimate (B_j^*) of the form:

$$B_j^* = \lambda_j \hat{B}_j + (1 - \lambda_j)\hat{\beta}. \tag{19}$$

This estimate is a weighted average of the site-specific impact estimate $\hat{B}_j$ and the overall mean impact estimate $\hat{\beta}$. The weight accorded to the site-specific estimate is its reliability (λ_j):

$$\lambda_j = \frac{\tau^2}{\tau^2 + V_j}. \tag{20}$$

Sites with large samples will tend to produce $\hat{B}_j$ values with a small sampling variance V_j and, thus, have high reliability. For those sites, $\hat{B}_j$ will have a large weight and $\hat{\beta}$ will have a small weight. For small sample sites with a large V_j and thus low reliabilities, the estimate of the true site impact will "shrink" toward the grand mean $\hat{\beta}$. We have considerable reason to believe that such empirical Bayes "shrinkage estimators" are the best possible predictors of true impacts under cross-validation (see Morris, 1983, for a review). The quantities V_j and τ^2 must be estimated.

Site-specific empirical Bayes estimators B_j^* are, in a sense, optimal for each site, given a reasonably large number of sites and sample members per site.[16] However, Louis (1984) noted that a histogram of empirical Bayes estimators will *understate* the variability of true impacts B_j. For this purpose, we recommend using "constrained" empirical Bayes estimators (Bloom et al., 2014).

Finally, shrinkage toward the overall mean is problematic when specific groups of sites vary markedly in their program effectiveness. In these cases, it might be more appropriate to shrink site-specific impact estimates toward a predicted value based on a theory of which kinds of sites have the largest effects (see Raudenbush & Bryk, 2002, Chapters 3 and 5). We consider such predictors (site-level moderators) in the next section.

Learning About the Distribution of Impacts of Program Participation

We have discussed the impact of random assignment of individuals to a program, known as an ITT effect. If everyone participates as assigned, whether as program or control members, we have an ideal situation called "perfect compliance" with random assignment. Unfortunately, perfect compliance rarely occurs. Instead, partial compliance results from two forms of behavior. First, some individuals assigned to a program will fail to participate. For example, in the MTO experiment (Kling, Liebman, & Katz, 2007), families living in public housing were randomly assigned to receive a voucher to pay rent in a low-poverty neighborhood. However, only 47% of the families assigned to receive the voucher actually used it. Second, individuals assigned to a control group can end up in the program being tested. For example, in lottery-based studies of charter schools, winners of the charter school lottery are invited to attend it and lottery losers are not. However, lottery losers may end up attending another charter school or even attend the charter school whose lottery they lost.

In studies where some persons assigned to the new program do not participate but no controls participate, the ITT impacts can be policy relevant. In these studies, the ITT effect represents the impact of a program on the persons for whom it was intended—that is, those who were assigned to it. However, in studies where some controls access the program, the ITT impact is of questionable relevance. In all cases of noncompliance, knowing the impact of actually participating in a program is important. For this purpose, a problem of selection bias arises, even in an RCT. This is because study participants shape the decision about whether to comply with random assignment. To cope with selection bias when estimating the impact of program participation, methodologists have widely adopted the method of instrumental variables (IVs) (Angrist, Imbens, & Rubin, 1996; Heckman & Vytlacil, 1998). For this approach, random assignment is conceived as an IV, which induces a subset of sample members to participate, and we can estimate the average impact of participation on those so induced ("compliers") under comparatively weak assumptions.

We first examine how the IV method works for a single site with homogeneous impacts. We then illustrate how the analysis becomes more complex—and more interesting—with program impacts

that vary across participants.[17] Finally, we consider how to use a multisite trial to estimate the cross-site mean and variance (or standard deviation) of the impacts of program participation.

The IV Method for a Single-Site Trial With Homogeneous Impacts

To understand the conventional IV method, consider the simple causal model in Figure 1. It begins with randomization of sample members to a program ($T = 1$) or control group ($T = 0$). This influences program participation, defined as $M = 1$ for participation and $M = 0$ for nonparticipation.[18] The impact of random assignment on program participation is denoted as γ, which is the difference between the probability of participating in the program if assigned to it and the probability of participating in the program if assigned to the control group. The impact of participating in the program on the outcome is denoted as δ.

Note that Figure 1 has no arrow between T and Y *and* thereby excludes a direct causal relationship between T and Y. This "exclusion restriction" is a key IV assumption and implies that the impact of program assignment on the outcome is produced entirely through the effect of program assignment on participation. In the language of path analysis, participation M "fully mediates" the ITT effect of T on Y, which we call β. This implies that the ITT effect is produced solely by the "indirect" effect of T on Y which operates through M, or:

$$\beta = \gamma\delta \tag{21}$$

The beauty of Equation 21 (when it holds for a situation) is that we can estimate δ without using M to predict Y. This eliminates potential selection bias noted earlier that occurs when trying to model Y as a function of M.

Instead, IV uses a two-stage approach. We estimate γ (the impact of T on M) and β (the impact of T on Y) without bias because T is randomly assigned. We then divide our estimate of β by our estimate of γ to obtain an approximately unbiased or, consistent estimate of δ:

$$\text{Impact of program participation} = \delta = \frac{\beta}{\gamma} = \frac{\text{ITT effect of } T \text{ on } Y}{\text{ITT effect of } T \text{ on } M}, \gamma > 0. \tag{22}$$

Another key assumption of Equation 22 is that assignment to the program increases the probability of participation, that is, $\gamma > 0$. This is easily checked, and it would be rare to find an experiment where this condition does not hold.

The IV Method for a Single-Site Trial With Heterogeneous Impacts

If we think that persons respond heterogeneously to a given program, constructing a person-specific path model of its impacts makes sense, as in Figure 2 (Raudenbush, Reardon, & Nomi, 2012). Here, Γ is the unique effect of assignment T on individual participation M. Γ is "compliance" with treatment assignment. The population-average compliance is $E(\Gamma) = \gamma$. Similarly, Δ is the person-specific causal effect of M on Y and with a population mean $E(\Delta) = \delta$. The "total effect" of T on Y for an individual is the *product* of the two causal effects Γ and Δ in the path from T to Y. The population mean effect of T on Y is:

$$\begin{aligned} E(B) = \beta = E(\Gamma\Delta) &= E(\Gamma)E(\Delta) + \text{Cov}(\Gamma, \Delta) \\ &= \gamma\delta + \text{Cov}(\Gamma, \Delta), \end{aligned} \tag{23}$$

as described by Raudenbush, Reardon, and Nomi (2012), based on Angrist, Imbens, and Rubin (1996).

As Equation 23 shows, the average effect of ITT (β) depends on the product $\gamma\delta$ of the two average causal effects and on the covariance across individuals $\text{Cov}(\Gamma, \Delta)$. This covariance is positive when

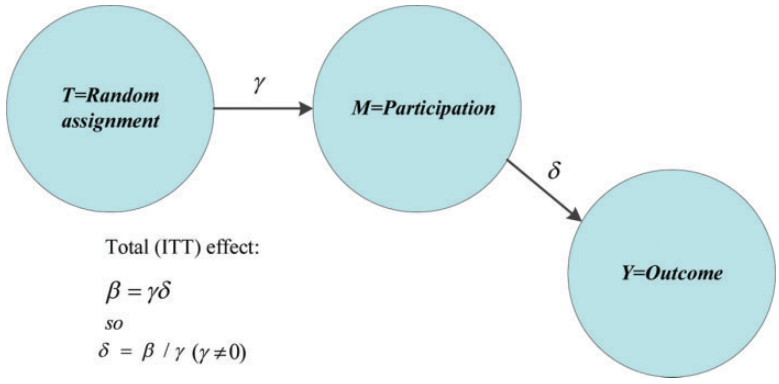

Figure 1. Single-site homogeneous impacts.

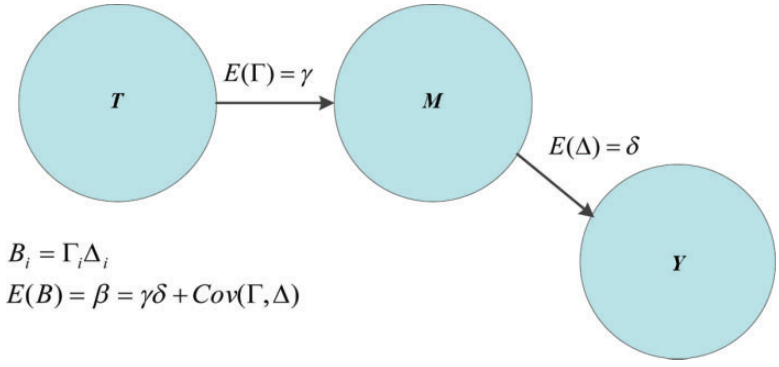

Figure 2. Single-site heterogeneous impacts: Person-specific causal model.

persons who comply with their assignment benefit more than others from participation. The covariance is negative if persons who comply with their assignment benefit less on average than others from participation.

How can we then estimate the average impact of program participation δ when the treatment effect is heterogeneous? We might assume as an approximation that $\mathrm{Cov}(\Gamma, \Delta) = 0$. Then, Equation 23 is equivalent to the conventional IV model in Equation 22, and we can identify the average treatment effect of M on Y as $\delta = \beta/\gamma$, $\gamma > 0$. However, sometimes the ''no compliance-effect covariance'' assumption may be implausible. For example, individuals may have knowledge about how much they will benefit from program participation, and this expectation may influence their willingness to participate (Roy, 1951). In other cases, highly disadvantaged persons who might benefit greatly from an intervention may be less able than others to comply with treatment assignment.

Rather than assuming no covariance between Γ and Δ, Angrist et al. (1996) developed an alternative approach when T and M are binary variables. They reasoned that four kinds of people exist: compliers, never takers, always takers, and defiers. *Compliers* are persons who would participate

($M = 1$) if offered a new program ($T = 1$) and not participate ($M = 0$) if assigned to a control group ($T = 0$). For compliers, the impact on M of being assigned to the program is $\Gamma = 1 - 0 = 1$. *Never takers* are persons who would never participate in the program, regardless of their treatment assignment. That means $M = 0$, regardless of treatment assignment, so that their impact on M of treatment assignment is $\Gamma = 0 - 0 = 0$. *Always takers* are persons who would always take up the program, regardless of their treatment assignment, so $M = 1$ either way, and for an always taker, $\Gamma = 1 - 1 = 0$. Defiers are persons who would refuse to take up M if assigned to the program (so $M = 0$ if $T = 1$) but who would participate if not assigned (so $M = 1$ if $T = 0$). Thus, for defiers, $\Gamma = -1$. In many situations, defiers are an implausible group and assuming that there are no defiers is often described as a *monotonicity* assumption. This means that assignment to the program cannot reduce the inclination of persons to participate in the program. Therefore, $\Gamma \geq 0$ (Angrist et al., 1996). Under this assumption:

$$\begin{aligned} E(B) = \beta &= E(\Delta\Gamma|\Gamma = 1)\Pr(\Gamma = 1) \\ &= E(\Delta|\Gamma = 1)\Pr(\Gamma = 1) \\ &= \delta_{\text{CACE}}\gamma. \end{aligned} \tag{24}$$

Here δ_{CACE} is the causal effect of program participation for persons with $\Gamma = 1$ (i.e., compliers), hence the label CACE. One problem for the interpretation of this effect is that it depends on who complies with treatment assignment, which in turn, depends on how effective a program is at inducing participation.

In sum, if the gain from program participation varies across participants (as in Figure 2), the population mean effect of program assignment (β) is no longer the simple product $\gamma\delta$, unless we invoke the assumption of no covariance between compliance and impacts. However, for a binary mediator, we can invoke the assumption of monotonicity (which is weaker).

Using Multisite Trials to Learn About Variation in CACE

Our aim now is to characterize the cross-site distribution of CACE effects. For this purpose, Raudenbush et al. (2012) introduce statistical methods for estimating the mean and the variance of CACEs across sites. We emphasize the important role in this process played by whether we focus on site-level or person-level estimands, as was the case for ITT impacts.

Defining and Estimating a Population Average CACE. Suppose we want to generalize to a population of sites and regard each site as equally representative of that population. We want to estimate the unweighted *true* average CACE (δ_{sites}), where:

$$\delta_{\text{sites}} = \sum_{j=1}^{J} \delta_j / J. \tag{25}$$

Here δ_j is the CACE for site j and δ_{sites} is the population average CACE. To estimate δ_j, an intuitive approach is to first estimate each site-specific CACE as $\hat{\delta}_j = \hat{\beta}_j / \hat{\gamma}_j$ and average these estimates across sites. Unfortunately, we find that this often does not work because a very low estimated compliance rate for even a single site (which is especially likely for small sites) can produce a value of $\hat{\delta}_j$ that is so large that it dominates the results for all other sites.[19]

As an alternative, we might begin with our unbiased estimate of the unweighted average ITT as:

$$\hat{\beta} = \sum_{j=1}^{J} \hat{B}_j / J. \tag{26}$$

Can we then divide this quantity by the estimated average compliance $\hat{\gamma}$ to obtain an average CACE? According to Raudenbush et al. (2012), in order to do so we must assume as an approximation, zero covariance between site mean compliance rates and site mean CACE impacts because:

$$E(\hat{\beta}) = \beta = E\left(\sum_{j=1}^{J} \frac{\gamma_j \delta_j}{J}\right) = \gamma_{\text{sites}}\delta_{\text{sites}} + \text{Cov}(\gamma_j \delta_j). \tag{27}$$

Suppose instead that we want to generalize to a population of persons so the CACE of interest weights each site's estimate by its population size (Equation 9), and we regard each person in our study as equally representative of that population, so that each site's sample size n_j is proportional to its population size N_j. In this case, our estimand (δ_{persons}) is:

$$\delta_{\text{persons}} = \left(\frac{\sum_{j=1}^{J^*} N_j \gamma_j \delta_j}{\sum_{j=1}^{J^*} N_j \gamma_j}\right). \tag{28}$$

Note there are $N_j\gamma_j$ compliers in site j and $\sum_{j=1}^{J^*} N_j\gamma_j$ compliers overall. So, each site's contribution to the overall CACE is weighted by the number of compliers in that site. If n_j is proportional to the site-specific population size N_j, we can substitute n_j for N_j in Equation 28.

We can therefore define the person-level population CACE as $\delta_{\text{persons}} = \beta_{\text{persons}}/\gamma_{\text{person}}$ without the assumption of no cross-site covariance between compliance and impact. To see this, note that our estimate of the overall ITT effect β has an expected value of $\beta = \delta\gamma$:

$$E(\hat{\beta}) = \lim_{J^*\to\infty}\left(\frac{\sum_{j=1}^{J^*} n_j B_j}{\sum_{j=1}^{J^*} n_j}\right) = \left(\frac{\sum_{j=1}^{J^*} n_j \gamma_j \delta_j}{\sum_{j=1}^{J^*} n_j}\right) = \left(\frac{\sum_{j=1}^{J^*} n_j \gamma_j \delta_j}{\sum_{j=1}^{J^*} n_j \gamma_j}\right) = \left(\frac{\sum_{j=1}^{J^*} n_j \gamma_j}{\sum_{j=1}^{J^*} n_j}\right) = \delta_{\text{persons}}\gamma_{\text{persons}}. \tag{29}$$

Studying a distribution of CACEs across sites. Raudenbush et al. (2012) describe several methods for estimating cross-site variance of CACEs, which are beyond the scope of this article. However, the key principles follow from the logic of the previous paragraphs: How we define our estimands is critical to shaping our approach for estimating a distribution of CACEs. Developing accessible methods for doing so is a focus of current methodological research.

Learning from a Distribution Of Program Impacts

We have discussed ways to study a cross-site distribution of program impacts. The idea now is to propose and test theories about when and why a program works, that is, to learn *from* a cross-site distribution of program effects in order to deepen our understanding of causal forces at work and how to manipulate them to improve program design and practice.

Moderation

Which types of persons benefit most from a program, and in what kinds of sites does the program work best? These important questions are about *moderation* of program impacts. We want to know whether a program works better for some types of persons than for others in order to target it efficiently or in order to investigate why the program does not work for certain types of persons. We would like to know which program sites are most effective, possibly to spur further investigation of practice in those sites or to frame general questions about why the program works when it does.

Questions about person-level and site-level moderators are almost always interdependent. Sites vary in the organizational conditions and practices that may be key to program success and in the composition of their client populations. Hence, claims about best practice at the site level might be misguided because especially effective sites might overrepresent persons who are most likely to benefit from the program being evaluated. As noted earlier, we define moderators of a program's impacts to be any characteristics of its clients or sites that influence the program's effectiveness but cannot be influenced by the program.

Person-level moderators. Evaluators commonly ask whether a program works better for boys than for girls, or for youth from high- versus low-income families, or for high- versus low-achieving students, or for persons of varying ethnicities. Such questions are often addressed through exploratory analyses conducted after average program impacts have been estimated. While such auxiliary analyses can enhance understanding, problems with this ad hoc, post hoc approach exist.

First, some subgroup findings may have limited relevance for policy or practice. For example, knowing that boys or ethnic minorities benefit most from a program might motivate further inquiry into why the program works for some clients but not for others—and this is a good thing. However, this knowledge does not necessarily imply that the program should make special effort to target particular subgroups.

Second, a search for subgroup impact variation can be stymied by the sheer number of subgroups to be examined. For example, the potentially large number of statistical tests of subgroup impact differences increases the likelihood of capitalizing on sample-specific differences that arise by chance and are therefore not replicable. Moreover, many subgroups are confounded with each other (i.e., they overlap). For example, ethnic minorities disproportionately comprise low-income persons, and boys have higher risk than girls for certain behavioral problems. Making theoretical sense of a large number of findings for such overlapping subgroups can be quite difficult.

Thus, we face a multiplicity of possible subgroups. No purely methodological fix to this problem exists, as the number of possible person-level moderators is too large to be sorted out by statistical hypothesis testing. What is needed is theory about who stands to benefit and why. Consider a program for increasing high school graduation rates. By construction, this program cannot appreciably increase graduation rates for students who would likely graduate without the program. At the opposite extreme, students with skills or prior grades that are so low that the program's resources are insufficient to appreciably improve their graduation prospects will tend not to benefit from the program. We have plenty of theory and evidence about which kids are most likely to drop out of school (Rumberger, 1995), so one can envision developing a theoretically informed model that predicts this probability in the absence of treatment. The evaluator might then stratify his or her sample based on this predicted probability or "prognostic score."

Stratifying on a prognostic score has several advantages. First, the prognostic score summarizes the predictive information in many different baseline characteristics, thereby greatly reducing the number of subgroup tests. Second, if program impacts depend strongly on a prognostic score, we confront interesting questions for policy and practice. One might envision, for example, targeting resources to persons with the greatest probability of benefiting. Third, stratifying on a prognostic

score might provide a more realistic assessment of the impact of the program than that provided by an estimate of its overall average effect. For example, a school drop-out prevention program can reduce dropouts only for students who are at some risk of dropping out. Suppose that at-risk students comprise 50% of one's sample. In that case, the average program effect on dropping out would be no more than half the size of the effect of the program on persons who could benefit from it.

We can augment a prognostic score analysis in ways that further understanding of impact variation. For example, with treatment group data, we could estimate a model that predicts post-program outcomes using individual baseline characteristics suggested by prior theory. Given randomization, the coefficients of this model for the treatment group should apply equally well to the control group, had they been assigned to the program. Thus, we can apply estimates of those coefficients to the baseline characteristics of control group members to predict how they might fare *with* access to the program. We could then use the same logic to obtain a prognostic score for how each program group member might fare *without* access to the program. In this way, we can estimate a *pair* of prognostic scores for each sample member and stratify them based on their pair of prognostic scores. By examining how program impacts vary across these strata *within sites*, we can efficiently summarize evidence about person-level moderators.[20]

Site-level moderators. Knowledge about site-level moderators is potentially of great importance for developing program theory, policy, and practice. We need to understand what organizational conditions are necessary if a new program will succeed. These conditions might include the availability of resources like staff skills and knowledge, the prevailing organizational climate in sites, or local ecological conditions such as neighborhood safety and unemployment rates.

Hence, just as we might wish to estimate program impacts for subgroups of persons, we might want to estimate program impacts for subgroups of sites. Once again, problems arise from the fact that many ways to define subgroups exist, and thus, there are many moderators to consider. Now the problem of "many moderators" is even more acute because there will always be far fewer sites per site-level subgroup than there are persons per person-level subgroup. Hence there is, much less precision for estimating impact differences across site-level subgroups than for estimating impact differences across person-level subgroups.[21] Consequently, the need for a priori theory to reduce the number of site-level moderators is even stronger than it is for person-level moderators.

Double stratification. As noted earlier, a major problem arises when studying moderators of program impacts; that is, site-level and person-level moderators are often mutually confounding. For example, sites with favorable organizational conditions might serve comparatively advantaged clients. Thus, what appears to be the influence of a site-level moderator on program impacts might actually be the influence of a person-level moderator or vice versa. One way to address this problem is "double stratification." For example, individual prognostic scores could be used to stratify sample members into two person-level subgroups—those at high risk of a negative outcome versus all others. In addition, program sites could be categorized according to a site-level moderator or set of moderators, (e.g., sites with high unemployment rates vs. all others and/or sites with high resource levels vs. all others). One could then split each site's sample into four groups: a high-risk treatment group and a high-risk control group plus a low-risk treatment group and a low-risk control group. In this case, some sites may have empty cells. For example, some sites might have no "low-risk" treatment group members or no low-risk control group members or both. However, for all sites that have high- and low-risk treatment *and* control group members, we can compare program impacts on high- and low-risk students controlling for a site-level moderator or set of moderators. Likewise, we can compare program impacts across values of site-level moderators controlling for participant risk.

Mediation

Why does a new program work—or not? Innovative programs are based on theories about how program operations generate short-term changes that produce long-term benefits. Such short-term changes are called mediators. We define mediators of program impacts to be those aspects of program implementation, staff practice, and short-term changes in participants' knowledge, skills, attitudes, or behavior that are outcomes of random assignment and predictors of long-term success. Mediators include shifts in organizational processes such as improved instruction or increased staff collaboration. These are often regarded as the *mechanisms* through which programs produce long-term benefits.

Methodological challenges. Analysis of mediational processes is popular in social science and program evaluation. However, drawing valid causal inferences about mediation is very challenging (for a detailed discussion of these challenges and alternative approaches to them, see the Keele article in this volume). For example, consider a study in which teachers are assigned at random to a professional development program with the aim of increasing instructional quality, which in turn is expected to improve student outcomes. Suppose that the program is successful in boosting student achievement. To what extent are the program-induced gains in student achievement explained—or ''mediated''—by program-induced improvement in instruction? This mediational analysis would assess the impact of the program on instructional quality. If teachers are assigned at random to the program or a control group, the difference between mean instructional quality for the treatment and control groups is an unbiased estimate of the causal effect of the program on instructional quality. Next, one seeks to assess the impact of instructional quality on student achievement. Establishing this causal link is especially challenging because teachers are not assigned at random to instructional quality. For example, teachers' pretreatment characteristics (experience, prior education, commitment, etc.) frequently predict their instructional quality. Such confounding can produce bias when studying the impact of instructional quality on youth outcomes.

A second problem arises when the impact of a mediator on the outcome has a different effect for treatment group members than for control group members. If this is the case, membership in the treatment group or control group *moderates* the causal effect of the mediator on the outcome. Conventional methods of path analysis thus do not work well (Holland, 1988; Pearl, 2001; Robins & Greenland, 1992). Presenters at the two national conferences referenced in the introduction of this article described three evolving statistical strategies for coping with these methodological challenges.

Multisite multimediator IV analysis. At the two conferences, Sean Reardon presented an approach that exploits site-to-site variation in the impact of a program on mediators. The rationale for this approach is intuitive. If M is a mediator and Y is an outcome of interest, we expect to see a large impact of a program on Y in sites where the program strongly affects M. If we fail to see such effects, we have evidence against the mediation theory. If we see effects, we have evidence of possible mediation. This idea extends nicely to the case of two mediators, call them M_1 and M_2. Suppose we see large effects of random assignment to the program on Y in sites where large effects of random assignment to the program on M_1 exist but not in sites where large effects of random assignment to the program on M_2 exist. Then, we would infer that M_1 is a more important mediator than is M_2. This intuition is the basis for Bloom, Hill, and Riccio's (2003) study of mediators in a series of large-scale multisite welfare-to-work experiments and Kling, Liebman, and Katz (2007) applied this approach to their study of MTO.

Reardon and Raudenbush (2013) derived the assumptions that must be met in order to infer that a specified mediator has a causal effect on a specified outcome. These assumptions are closely related

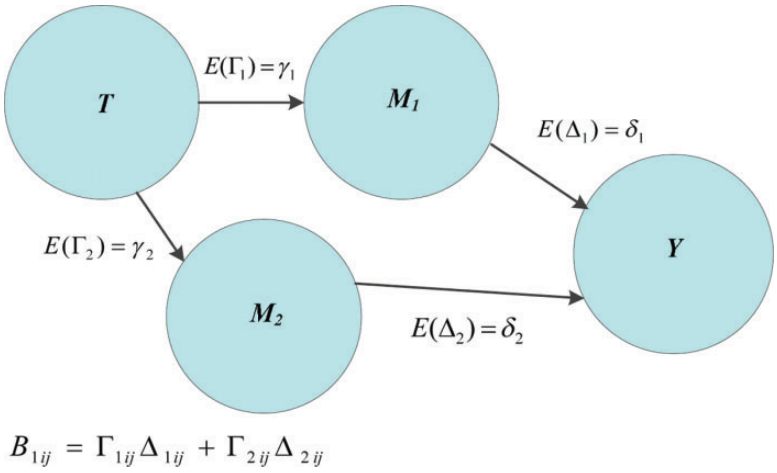

Figure 3. Multiple site, two mediators: Person-specific causal model.

to the assumptions we described earlier when the aim was to identify the impact of participating in a new program (CACE). Indeed, program participation can be regarded as a mediator of the effect of program assignment, as described in Figure 2. The multisite, multimediator model extends this basic idea to the case of two or more mediators, as shown in Figure 3. Now, our IV T induces a shift in two mediators, M_1 and M_2 and each of these, by hypothesis, influences the outcome Y. Readers familiar with IV methods will immediately raise a question. We now have one instrument and two causal variables, meaning that we will end up with one equation and two unknowns. How can this possibly work? Here the beauty of the multisite design comes into play. We can regard the treatment assignment indicator in each site as a separate IV. Thus, if there are J sites with a treatment group and control group for each site, we have J instruments, enabling us to identify the impact of our two or more mediators on the outcome under several important assumptions.

We can clearly recognize these assumptions, when we represent Figure 3 as a regression model. Let's call B_j the ITT effect in site j. Suppose that this effect works entirely through two mediators, M_1 and M_2. The impact of T on M_1 in site j is γ_{1j} and the impact of T on M_2 in site j is γ_{2j}. In terms of path analysis as shown in Figure 3, B_j is the total effect of T on Y in site j, and it works strictly through indirect effects on the two mediators. Hence, we can express the path model as

$$\begin{aligned} B_j &= \delta_{1j}\gamma_{1j} + \delta_{2j}\gamma_{2j} \\ &= \delta_1\gamma_{1j} + \delta_2\gamma_{2j} + e_j. \end{aligned} \tag{30}$$

Here δ_1 is the overall average impact of M_1 on Y controlling for M_2, δ_2 is the overall average impact of M_2 on Y controlling for M_1. Equation 30 is a simple regression model where the outcome B_j is the ITT effect on Y, and the predictors are γ_{1j} (the ITT effect on M_1) and γ_{2j} (the ITT effect on M_2).[22] The logic of this setup is that we can estimate B_j (the dependent variable in the regression model) as well as γ_{1j} and γ_{2j} (the two independent variables in the regression model) without bias based simply on the random assignment of participants to T. However, this gift comes at the price of several assumptions:

Assumption 1: We are assuming that T has no direct effect on Y (the exclusion restriction). This means that no unobserved mediators exist and is why there is no intercept in Equation 30. Note the absence of an intercept implies that when the impact of treatment effect on each mediator is

zero (that is $\gamma_{1j} = \gamma_{2j} = 0$), the predicted value of B_j will be zero. This prediction would be false in the presence of unobserved mediators.

Assumption 2: To regard Equation 30 as a regression model with identifiable parameters, no bias can be associated with the error term, $e_j = (\delta_{1j} - \delta_1)\gamma_{1j} + (\delta_{2j} - \delta_2)\gamma_{2j}$. We must assume that the site-average impact of T on each mediator is not related to the site-average impact of either mediator on the outcome. Specifically, this means we are assuming that γ_{1j} is unrelated to δ_{1j} and to δ_{2j} and that γ_{2j} is unrelated to δ_{1j} and to δ_{2j}.

Assumption 3: A nonzero impact of T on each mediator must exist in one or more sites.

Assumption 4: The impact of T on at least one of the two mediators must vary from site to site, and the impact of T on M_1 must not be too highly correlated with the impact of T on M_2.

Assumption 5: The mediators must operate "in parallel," meaning that one mediator is not a cause of the other. If this assumption fails, we need a sequence of regression equations to represent a sequential rather than parallel mediation process, and the assumptions become stronger.

We can readily check Assumptions 3 and 4 against observed data, so they do not pose a strong challenge. Assumption 5 is based on program theory. The other assumptions, however, cannot be checked against the data.

Reardon, Unlu, Zhu, and Bloom (2014) discuss conditions under which failures of these assumptions are most likely to cause bias for analyses of a single mediator. They also provide a bias correction that is applicable when Assumption 2 fails and the goal is to estimate a single mediator effect. We anticipate that future work will extend these innovations to the case of multiple mediators. This is important because Assumption 2 is potentially a strong assumption.

We conclude that the multisite, multimediator IV method opens up interesting new ways to exploit cross-site heterogeneity in order to study the impact of program mediators on participants' outcomes. However, this new and evolving method merits study to learn more about how failure of its assumptions influences its results.

Other strategies for mediation analysis in multisite trials. Finding flexible new strategies for mediation analysis is currently a topic of great interest in social science and public health (see recent books by Hong, 2015 and VanderWeele, 2015). Presenters at the aforementioned conferences reviewed two of the most potentially useful approaches: principal stratification and sequential randomization. A key feature of these approaches is that they do not require the exclusion restriction we relied on when describing the multisite, multimediator IV approach. A key limitation for our current discussion is that the application of these approaches to multisite trials is not yet well developed but is a topic of currently intense methodological research. Given the multisite theme of this article, we describe these approaches very briefly.

Principal stratification. One goal of principal stratification applied to the analysis of mediation is to estimate program impacts on persons whose mediator values are *not affected* by program assignment. These are "direct effects" of the program because they operate independently of the mediator or mediators of interest. The existence of a program impact on an outcome for persons who do not experience a program impact on a hypothesized mediator refutes the claim that the program's impact is generated entirely through that mediator. The idea is to stratify one's sample based on "potential mediator values" and to compare estimated program impacts for selected strata. Frangakis and Rubin (2002) label these strata as "principal strata." Two sample members belong to the same principal stratum if their *pair* of potential values for a given mediator is the same. In other words, they belong to the same principal stratum if the value of their mediator under assignment to a program is the same *and* if the value of their mediator under assignment to control status is the same.

The problem of course is that we cannot observe the two potential mediator values for any sample member, so the principal stratum membership is unknown. However, as presented by Lindsay Page in this volume, it is possible in some important cases to use baseline and follow-up data for sample members to estimate a model that *predicts* their two potential mediator values and thereby predicts their principal stratum membership.

Sequential randomization. An innovative strategy for mediation analysis, described by Guanglei Hong at the William T. Grant Foundation conference, conceives of the mediation process as a sequence of randomized experiments (Pearl, 2001; Robins & Greenland, 1992). Consider how this works in the case of a single binary mediator, where $M = 1$ if the mediator value is favorable and $M = 0$ if it is not favorable. The first experiment is directly observable: We assign participants at random to a new program ($T = 1$) or to its control group ($T = 0$). The second experiment is hypothetical: Program group members are assigned at random to the favorable value of the mediator with some probability. Control group members are also randomly assigned to the favorable mediator value but with a different probability. If we knew these two probabilities, we could make the needed causal inferences (Imai, Keele, & Yamamoto, 2010). The empirical challenge is to estimate these probabilities, which may depend on baseline characteristics of sample members and the study setting.

Let's call the mediator value to which a program group member is assigned $M(1)$ and the mediator value to which a control group member is assigned $M(0)$. In principal stratification, these two potential mediator values are treated as *fixed characteristics* of each sample member that depend on his background and the study setting. For analyses based on sequential randomization, the values of $M(1)$ and $M(0)$ are treated as *stochastic.* The probability that $M = 1$ depends on a participant's past and whether he or she is randomly assigned to the program group or control group. Under sequential randomization, an effective program is seen as increasing the chance of receiving a favorable mediator value.

Recall that principal stratification groups sample members in terms of their predicted pair of potential mediator values, $M(1)$ and $M(0)$, based on their background characteristics and future outcomes. In contrast, under the assumption of sequential randomization, we seek to group sample members based on their pair of *probabilities* of experiencing a favorable mediator value under assignment to the program group and under assignment to the control group. This approach enables the analyst to estimate (a) the indirect effect, which is the causal effect on the outcome of changing the mediator value without changing program assignment and (b) the direct effect, which is the causal effect of changing program assignment without changing the mediator value. The relative magnitudes of these two component effects indicate the degree to which the program effect was transmitted by the hypothesized mediator.

Perhaps the key challenge is that, while participants are randomly assigned to the treatment T, they are not randomly assigned to the mediator M. However, if a rich set of pretreatment characteristics (call them X) are measured, we may be willing to assume that, within a stratum of persons with similar values of X, assignment of the mediator is effectively "as if" random. This means that, within such strata, there are program participants whose mediator values vary by chance and control participants whose mediator values vary by chance. Methodologists have devised a number of clever strategies for estimating direct and indirect effects in this context (see Hong, 2015; VanderWeele, 2015).

One difficulty with this approach is stratifying sample members on a potentially long list of baseline characteristics (X). To deal with this issue, one can use a propensity score (Rosenbaum & Rubin, 1983) because stratifying sample members on a propensity score can balance stratum members on all variables used to predict the propensity score, at least in large samples.

Comparing alternative approaches to mediational analysis. The preceding approaches for studying mediation of program impacts—multisite IVs, principal stratification, and the approximation of sequential randomization—have different strengths and limitations. The first approach directly exploits multisite

RCTs to create a series of valid instruments and does not rely on pretreatment covariates to produced unbiased (or consistent) estimates. However, this approach requires all relevant mediators be observed and accounted for and is potentially subject to "omitted mediator bias." Moreover, one must assume that site-specific impacts of the treatment on the mediator are not associated with site-specific impacts of the mediators on the outcome. In contrast, the approximation of sequential randomization does not assume that all mediators are measured and modeled. Rather, like standard path analysis, this approach decomposes the effect of treatment assignment into indirect effects that work through specified mediators and a direct effect that works through additional mediators that are unobserved. In doing so, the approach relaxes parametric assumptions that are commonly used for path analysis. However, like path analysis, sequential randomization requires a rich set of pretreatment covariates to support the assumption that, conditional on these covariates, mediator values in the treatment group and in the control group are effectively assigned randomly but with different probabilities. The principal stratification approach does not require measuring and modeling all pretreatment confounders (as does sequential randomization) or the exclusion restriction (as does IVs). Instead, it requires covariates and follow-up outcomes that adequately predict the potential values of sample members' mediators. In addition, principal stratification is more useful for identifying a direct effect of program assignment and thereby *falsifying* a mediation theory than it is for estimating the parameters of a mediational process.

None of these approaches is perfect for all mediational analyses, and all mediational analyses (short of randomizing specified mediator values to treatment and control group members) require strong assumptions in order to estimate mediator effects. However, the assumptions required by these new strategies are less stringent than those required by conventional path analysis. Furthermore, despite the substantial difficulties of mediational analysis, we believe that it is essential for building a science of program design and development. However, selecting a method of mediational analysis for multisite trials is craft knowledge that is not yet fully understood or widely available.

Final Remarks

The presence of variation in program impacts upends conventional ways of analyzing and interpreting data from program evaluations, especially in multisite trials, which are very common in program evaluation. Among other things, impact variation makes it possible to define and estimate different types of average impacts. For example, we can define an average impact for a population of sites or an average impact for a population of persons, and with heterogeneity of program impacts, these parameters can differ.

However, any average becomes less informative as impact variation increases. Understanding this variation thus becomes more important, and new questions arise such as (a) By how much do impacts vary across individuals, subgroups of individuals and program sites? (b) What is the cross-site correlation between program impacts and control group mean outcomes? (c) What are the maximum and minimum site-specific program impacts? Searching for answers to these questions is learning *about* a distribution of program effects.

The existence of a distribution of program impacts also provides opportunities for testing theories about for whom, under what conditions, and why programs work. Toward this end, we can pose theories to guide future data collection for explaining impact variation within and across program sites. Theory building is learning *from* a distribution of program effects.

Statistical methods for discovering and explaining impact variation are developing rapidly, and we have provided a broad overview of new approaches. However, a great deal remains to be done, and we anticipate many new methodological breakthroughs during the next decade.

Declaration of Conflicting Interests

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

Funding

This article was supported by the Spencer Foundation (for Bloom) and the William T. Grant Foundation (for Raudenbush and Bloom). We thank the following individuals for their insights into the issues explored by the article: Adam Gamoran, Guanglei Hong, Lindsay Page, Sean Reardon, Michael Weiss and Kim Dumont.

Notes

1. Versions of this framework have been attributed to Neyman (1923/1990), Roy (1951), Heckman (1979), Rubin (1974, 1978), and Holland (1986).
2. Specifying one and only one potential outcome for each person under each treatment assignment assumes that potential outcomes cannot be influenced by the treatment assignment of others or by the mechanism by which the treatment is assigned. This is known in the literature as the Stable Unit Treatment Value Assumption or SUTVA (Rubin, 1986).
3. This requirement is a consequence of SUTVA (note 4). It implies that a person's potential outcomes do not depend on the treatment assignment of other persons. Hong and Raudenbush (2006) provide a strategy for relaxing this assumption when interactions in social settings make the assumption implausible.
4. Equation 5 implies that the only way for a program group outcome variance to differ from its control group outcome variance is for program impacts to vary across individuals.
5. The expression $\mathrm{Var}(B) + 2\mathrm{Cov}[Y(0),\ B]$ can be near zero even if the individual-level impact variance, $\mathrm{Var}(B)$, is positive and persons who would fare less well than average without the program benefit by more than average from it (i.e., when $\mathrm{Cov}[Y(0),\ B]$ is negative).
6. Equation 5 implies that the program group outcome variance can be less than the control group outcome variance only if $\mathrm{Cov}[Y(0),\ B]$ is negative.
7. If program effects vary, the program group outcome variance can exceed the control group outcome variance even if $\mathrm{Cov}[Y(0),\ B]$ is zero.
8. Many other design options are possible. For example, one might over-sample certain sites (perhaps those with small but scientifically important or policy-relevant subpopulations). Or one might over-sample particular subpopulations within sites. Yet one can still use β_{persons} or β_{sites} as an estimand of interest.
9. For this purpose, one might use the site sample size (n_j) to approximate the relative site population size (N_j).
10. Recall from our discussion of impact variation within sites that the program group outcome variance can differ from that for the control group. We ignore this complication here in order to focus on key issues.
11. The sampling variance is $V_i = \mathrm{Var}(\hat{B}_j - B_j) = \sigma^2/[n_j\bar{T}_j(1-\bar{T}_j)]$.
12. Perhaps the simplest way to do this is to specify the mixed model $Y_{ij} = \alpha_j + (\beta + b_j)T_{ij} + e_{ij}$, where α_j is the fixed site effect, $\beta + b_j$ is the random coefficient for the treatment indicator T_{ij}, and e_{ij} is a within-site random error. Some software packages allow specification of "site" as a fixed effect, while in others, it will be essential to represent the J site fixed effects with $J-1$ dummy variables and an intercept (or with J dummy variables and no intercept).
13. In contrast, suppose that we regard the persons in our convenience sample as representing a universe of similar persons who might experience the new program and we want to generalize to that universe of persons. Then, we might set $w_j = n_j$.
14. Equation 17 will be very closely approximated by estimating the model described in Note 15 using restricted maximum likelihood.
15. The estimated sampling covariance between the program impact and the control group mean in site j is $\hat{C}_{B0} = \hat{\mathrm{C}}\mathrm{ov}(\hat{B}_j, \hat{U}_{0j}) = \hat{\mathrm{C}}\mathrm{ov}(\bar{Y}_{1j} - \bar{Y}_{0j}, \bar{Y}_{0j}) - \hat{\sigma}^2/[n_j(1-\bar{T}_j)]$. This is obtained from the regression $Y_{ij} = U_{0j} + B_jT_{ij} + e_{ij}$ using ordinary least squares, where $\hat{\sigma}^2 = \sum_{j=1}^{J}\sum_{i=1}^{n_j}(Y_{ij} - \hat{U}_{0j} - \hat{B}_jT_{ij})^2/(N-J)$ and $N = \sum_{j=1}^{J} n_j$.

16. As the number of sites increases without bound, the estimator B_j^* will produce, on average, the minimum mean squared error of estimation of the true impact B_j.
17. For a more detailed discussion, see Raudenbush et al. (2012).
18. We use M to represent program participation because it mediates the impact of program assignment.
19. The problem here is that we are trying to divide by a value that is approaching zero.
20. When estimating a predictive model based on data for one of two groups and using it to predict outcomes for both groups, we must take care to avoid the problem of "overfitting" the model to the group for which it was estimated. See Abadie, Chingos, and West (2014) for a discussion of this problem and ways to avoid it.
21. This assumes that we are using a site-level random coefficients model to estimate and test impact differences across site-level strata.
22. Raudenbush et al. (2012) show how to identify the model (Equation 30) using two-stage least squares within the framework of a hierarchical linear model so we do not provide details here.

References

Abadie, A., Chingos, M., & West, M. (2014). *Endogenous stratification*. Retrieved from http://www.hks.harvard.edu/fs/aabadie/stratification.pdf

Angrist, J. D., Imbens, G., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, *91*, 444–455.

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*, 551–575.

Bloom, H. S., Raudenbush, S. W., Weiss, M., & Porter, K. (2014). *Using multi-site evaluations to study variation in effects of program assignment*. New York, NY: MDRC.

Bloom, H. S., & Weiland, C. (2015, March) *Quantifying variation in Head Start effects on young children's cognitive and socio-emotional skills using data from the National Head Start Impact Study*. New York, NY: MDRC.

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

Dempster, A. P., Rubin, D. B., & Tsutakawa, R. K. (1981). Estimation in covariance components models. *Journal of the American Statistical Association*, *76*, 341–353.

Fisher, R. A. (1925). *Statistical methods for research workers*. Edinburgh, Scotland: Oliver and Boyd.

Frangakis, C. E., & Rubin, D. B. (2002). Principal stratification in causal inference. *Biometrics*, *58*, 21–29.

Greenberg, D. H., & Shroder, M. (2004). *The digest of social experiments*. Washington, DC: The Urban Institute.

Hong, G., & Raudenbush, S. W. (2006). Evaluating kindergarten retention policy. *Journal of the American Statistical Association*, *101*, 901–910.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica: Journal of the Econometric Society*, *47*, 153–161.

Heckman, J. J., & Vytlacil, E. (1998). Instrumental variables methods for the correlated random coefficient model: Estimating the average rate of return to schooling when the return is correlated with schooling. *Journal of Human Resources*, *33*, 974–987.

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

Holland, P. W. (1988). Causal inference, path analysis, and recursive structural equations models. *Sociological Methodology*, *18*, 449–484.

Hong, G. (2015). *Causal inference in a social world: Moderation, mediation, and spillover*. Sussex, England: John Wiley.

Imai, K., Keele, L., & Yamamoto, T. (2010). Identification, inference and sensitivity analysis for causal mediation effects. *Statistical Science*, *25*, 51–71.

Katz, L. F., Kling, J. R., & Liebman, J. B. (2000). *Moving to opportunity in Boston: Early results of a randomized mobility experiment* (No. w7973). Cambridge, MA: National Bureau of Economic Research.

Kling, J. R., Liebman, J. B., & Katz, L. F. (2007). Experimental analysis of neighborhood effects. *Econometrica*, *75*, 83–119.

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

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

Morris, C.N. (1983)Parametric empirical Bayes inference: Theory and applications. *Journal of the American Statistical Association*, *78*, 47–65.

Neyman, J. S. (1990). On the applicability of probability theory to agricultural experiments: Essay on principles, Section 9. In D. M. Dabrowska & T. P. Speed (Eds. & Trans.), *Statistical Science* (Vol. *5*, pp. 465–480). (Original work published 1923)

Pearl, J. (2001, August). Direct and indirect effects. In J. Breese & D. Koller (Eds.), *Proceedings of the seventeenth conference on uncertainty in artificial intelligence* (pp. 411–420). San Francisco, CA: Morgan Kaufmann Publishers Inc.

Raudenbush, S. W. (2014). *Random coefficient models for multi-site randomized trials with inverse probability of treatment weighting*. Chicago, IL: Department of Sociology, University of Chicago.

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

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

Reardon, S. F., & Raudenbush, S. W. (2013). Under what assumptions do multi-site instrumental identify average causal effects? *Sociological Methods and Research*, *42*, 143–163.

Reardon, S. F., Unlu, F., Zhu, P., & Bloom, H. S. (2014). Bias and bias correction in multisite instrumental variables analysis of heterogeneous mediator effects. *Journal of Educational and Behavioral Statistics*, *39*, 53–86. doi:1076998613512525

Robins, J. M., & Greenland, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, *3*, 143–155.

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

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*, 41–55.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Eeducational Psychology*, *66*, 688.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics*, *7*, 34–58.

Rubin, D. B. (1986). Comment: Which ifs have causal answers. *Journal of the American Statistical Association*, *81*, 961–962.

Rumberger, R. W. (1995). Dropping out of middle school: A multilevel analysis of students and schools. *American Educational Research Journal*, *32*, 583–625.

Spybrook, J. (2013). Detecting intervention effects across context: An examination of the precision of cluster randomized trials. *The Journal of Experimental Education*, *82*, 334–357. doi:10.1080/00220973.2013.813364

VanderWeele, T. J. (2015). *Explanation in causal inference: Methods for mediation and interaction*. Oxford, England: Oxford University Press.

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