MDRC Working Papers on Research Methodology

Learning *About* and *From* Variation in Program Impacts Using Multisite Trials

Stephen W. Raudenbush
University of Chicago

Howard S. Bloom
MDRC

April 2015

Acknowledgments

This paper was supported by the Spencer Foundation and the William T. Grant Foundation as part of a jointly funded project on creating methods and using existing databases to detect, quantify, and predict variation in effects of educational and youth developmental programs. We thank Adam Gamoran and Kim Dumont for their helpful feedback on an earlier draft of the paper, and we thank our project colleagues Sean Reardon, Guanglei Hong, Lindsay Page, Michael Weiss, and Kristin Porter for their collaboration in the development of many of the ideas in the paper. The opinions expressed herein and any mistakes they might reflect are solely our responsibility, however.

Dissemination of MDRC publications is supported by the following funders that help finance MDRC's public policy outreach and expanding efforts to communicate the results and implications of our work to policymakers, practitioners, and others: The Annie E. Casey Foundation, Charles and Lynn Schusterman Family Foundation, The Edna McConnell Clark Foundation, Ford Foundation, The George Gund Foundation, Daniel and Corinne Goldman, The Harry and Jeanette Weinberg Foundation, Inc., The JBP Foundation, The Joyce Foundation, The Kresge Foundation, Laura and John Arnold Foundation, Sandler Foundation, and The Starr Foundation.

In addition, earnings from the MDRC Endowment help sustain our dissemination efforts. Contributors to the MDRC Endowment include Alcoa Foundation, The Ambrose Monell Foundation, Anheuser-Busch Foundation, Bristol-Myers Squibb Foundation, Charles Stewart Mott Foundation, Ford Foundation, The George Gund Foundation, The Grable Foundation, The Lizabeth and Frank Newman Charitable Foundation, The New York Times Company Foundation, Jan Nicholson, Paul H. O'Neill Charitable Foundation, John S. Reed, Sandler Foundation, and The Stupski Family Fund, as well as other individual contributors.

The findings and conclusions in this report do not necessarily represent the official positions or policies of the funders.

For information about MDRC and copies of our publications, see our website: www.mdrc.org.

Abstract

The present paper, which is intended for a diverse audience of evaluation researchers, applied social scientists, and research funders, provides a broad overview of the conceptual and statistical issues involved in using multisite randomized trials to learn *about* and *from* variation in program effects across *individuals*, across policy-relevant and theoretically relevant *subgroups* of individuals, and across program *sites*. Learning about variation in program effects involves detecting and quantifying this variation. Learning from variation in program effects involves studying the factors which predict or explain it. The paper is divided into four sections, plus a brief final discussion. The first section introduces the concepts and issues involved. Section 2 focuses on detecting and quantifying variation in effects of program *assignment*, which are often referred to as effects of intent to treat (ITT). Section 3 extends the discussion to variation in effects of program *participation*, which are often referred to as a complier average causal effect (CACE) or a local average treatment effect (LATE). Section 4 considers *moderators* of program effects (individual-level or site-level factors that influence the sign and magnitude of these effects) and *mediators* of program effects (individual-level or site-level "mechanisms" by which these effects are produced).

Contents

List of Exhibits

Figures

Section 1

Introduction

For over 50 years, local, state, and federal agencies and private foundations have sought to identify and support effective new programs to improve the life chances of young children, adolescents, and adults. These programs aim to help participants acquire new skills and educational credentials, find jobs, stay out of trouble with the law, and form and maintain healthy families. Committed teachers, administrators, service providers, and social scientists have designed these programs, often with the hope of helping the nation's most disadvantaged populations.

Research evidence plays a central role in this conception of social improvement. Innovators generate a lot of promising ideas, but not all ideas work out in practice. Some programs are hard to implement at a large scale, while others, when implemented, do not produce the benefits expected of them. Therefore, accurate and objective evidence about the impacts of new programs is essential for discovering truly effective practices. But such evidence is not so easy to acquire.

The Problem of Causal Inference About Program Impacts

To make a valid causal statement about the impact of a new reading program, dropout prevention program, or job-training initiative, it is not enough just to measure the gains made by program participants; one must also be able to estimate how participants would have fared had they *not* had access to the program. For this purpose, one needs a valid comparison group; that is, a group of individuals who are similar to those in the program at the outset of the study but who do not participate in the program. The impact estimation strategy is then to compare future outcomes of program group members to those of comparison group members under the assumption that comparison group outcomes tell us what "would have happened" to the program group without the program.

However, finding a good comparison group can be very difficult. Persons who decide to participate in a new program may be more motivated, skilled, or resourceful than persons in a study's comparison group. In other cases, persons assigned to a program may be more troubled or less skilled than those not assigned. In these situations, it is hard to know whether future differences in outcomes for the program and comparison groups reflect the impact of the program or simply reflect preexisting differences between the groups being compared. One strategy for addressing this problem is to measure the prior skills and motivation of program and comparison group members and use statistical methods to adjust for preexisting differences in these characteristics. However, this strategy requires the identification and measurement of all crucial

"confounding variables," that is, all characteristics of sample members that predict their program assignment *and* their future outcomes. The burden is thus on the evaluator to convince a skeptical audience that all crucial confounders were known in advance and accurately measured. This is a tall order.

To overcome this comparison group challenge for testing new medical treatments, scientists after World War II embraced the randomized controlled trial (RCT). The idea of randomly assigning units to alternative treatments or to either a specific treatment or a nontreatment "control group" had its origins in agricultural research during the early twentieth century, and medical researchers borrowed the accumulated knowledge about how to design and analyze such studies.

In addition, over the past 15 years, policymakers have funded a large number of RCTs to evaluate a wide range of new and existing programs that aim to improve the life chances of young children, adolescents, or adults. As a result, we have learned a great deal about the effectiveness of preschool education, charter schools, remedial math and reading interventions, afterschool services, teacher professional development, career academies, job training programs, social service programs, criminal justice programs, and more. The beauty of the RCT for this purpose is that if well implemented, it provides an unbiased estimate of the average impact of a new program on a target population. This is because random assignment to treatment or control status eliminates selection bias that might otherwise exist if persons assigned to the new program were more — or less — skilled, motivated, or otherwise advantaged or disadvantaged than persons in a nonexperimental comparison group.

A Critical Review of RCT Practice

At two recent national conferences, researchers and research funders met to critically review the design and analysis of RCTs. In October 2013, with funding from the William T. Grant Foundation, leading social scientists and research funders met in Chicago to consider ways to improve educational RCTs sponsored by the U.S. Institute of Education Sciences (IES), by the foundation itself, and by other funders of research in education and related fields.[1] In September 2014, the Administration for Children and Families (ACF) sponsored a related conference in Washington, DC, to consider how its programs and those of other federal, state, and local agencies might be more productively evaluated.[2]

[1]Conference on "Learning from Variation in Program Effects," sponsored by the William T. Grant Foundation (Chicago: October 7-9, 2013).

[2]Conference on "What Works, Under What Conditions, and How? Methods for Unpacking the 'Black Box' of Programs and Policies," sponsored by the Office of Planning, Research and Evaluation of the Agency

Analyses of most RCTs have focused mainly, if not exclusively, on a single question: What is the *average* impact of a new program on persons that it is intended to serve? This is a crucial question and one that is often not easy to answer even with an RCT, in part because of noncompliance with random assignment. Such noncompliance occurs when some persons randomly assigned to a program that is being evaluated do not participate in it, and some persons randomly assigned to the program's control group do participate in the program. Other problems that frequently occur with RCTs (as well as with other types of evaluation designs) include missing follow-up data for some sample members, missing data for some baseline characteristics for some sample members, and errors of measurement for key outcomes. Fortunately, methodologists have created ingenious strategies for coping with these challenges.

However, notwithstanding these methodological advances, participants in the two conferences noted that despite the importance of learning about *average* program impacts, this knowledge by itself is insufficient for the future development of public policy, professional practice, or program theory. What is also needed is an understanding of whether program impacts vary, by how much, and why. Thus conference participants made the case that we can benefit substantially from learning more *about* and *from* heterogeneity of program impacts.

Learning *About* Heterogeneity of Impacts

First, we need to know a great deal more *about* variation in program impacts. For example, we can learn a lot about the generalizability of findings from an RCT by detecting and quantifying variation in program impacts. And we can often do this without imposing (or assuming) a theory or model of who will benefit most or least from a program. Among other things, learning about variation in program impacts involves quantifying this variation, assessing the equity of this variation, and studying site-specific impacts.

Quantifying Variation

To what extent do participants vary in their response to a new program? To what extent does the impact of a new program vary across program sites? If individuals vary little in their response to a new program and if sites vary little in their average program impacts, the overall average impact is a truly valuable summary of evidence about program effectiveness. But if persons vary greatly in their response or if sites vary greatly in their average program impacts, that overall average is a much less useful guide for policymakers who might contemplate adopting the program or for practitioners who would like to know how to improve it. Hence, we need

---

for Children and Families of the U.S. Department of Health and Human Services (Washington, DC: September 3-4, 2014).

to learn more about *whether* and *to what extent* impacts vary across individuals, subgroups of individuals, and sites.

Assessing Equity

In multisite trials, we can study the correlation between the control-group mean and program impacts across sites. If we learn that program sites that serve individuals who would do especially poorly without the new program produce above-average impacts, we have evidence that the program will tend to reduce inequality. But if we learn that program sites that serve individuals who would do especially *well* without the program produce above-average impacts, this is evidence that the program will tend to increase inequality. We can assess program equity in this way without relying on strong theoretical assumptions. However, evaluators rarely ask this question, and currently available statistical methods do not answer it reliably.

Studying Site-Specific Impacts

To properly evaluate a program, we need to know what fraction of its participants benefit from it and what fraction, if any, fare less well with the program than they would have without it. For this purpose, we can capitalize on the fact that most RCTs in education and social program research are multisite trials. Thus outcome data for the control group at each site provide a valid estimate of how its program group would have fared without the program. Consequently, we can obtain a valid estimate of the program effect for each site. This information makes it possible to assess, among other things, the effectiveness of the most and least effective sites. And, in some cases, knowing how effective a program *can be* is as important as knowing how effective it is on average — particularly if one can find ways to learn from best practice.

Learning *From* Heterogeneity of Impacts

Having learned *about* the existence and magnitude of variation in program impacts, a great deal can also be learned *from* this variation. The idea here is that heterogeneity of impacts creates opportunities for testing theories about what social scientists call moderation and mediation of impacts.[3]

Moderation of Impacts

In theory, the impact of a program can vary because some types of persons are more likely than others to participate, some types of participants benefit more than others from the

[3]Weiss, Bloom, and Brock (2014) provide a general conceptual framework for studying the factors that influence and produce variation in program impacts.

program, staff at some program sites are more skilled than staff at other sites, or existing services available outside of the program are more widely available and/or more effective at some sites than at others. In this paper, we define as *moderators* any characteristics of clients and sites that (a) cannot be influenced by the intervention and (b) facilitate or inhibit program effectiveness.

Under our definition, any characteristic of a person or a site that is observable before random assignment is a potential moderator. However, we also can envision as potential moderators some characteristics that change over time. For example, the local unemployment rate, which changes over time, may influence participant's motivation to stay in school or attend job training and affect his chances of finding a job after attending training. If the intervention cannot plausibly influence the local unemployment rate, then we can regard the local unemployment rate as a potential moderator. In contrast, the effectiveness of staff practice observed after the intervention is implemented cannot, in our view, be a moderator of the effects of random assignment. This is because such practices can be influenced by the program being tested. Indeed, in many cases, this influence is part of how the program is intended to work. We refer to such variables as potential *mediators* of program effects.

To explore potential moderators, it is common practice for evaluators to conduct secondary analyses to determine whether certain client subgroups (defined by their gender, ethnicity, social background, risk of failure, etc.) benefit more than others from a program. It is much rarer to find an evaluation that is founded on a moderation *theory*: that is, a theory of who will benefit most or least from the program being studied and what organizational conditions are most important for its success. Yet posing and testing such theories using sound analytic methods would significantly increase the ability of evaluations to explain heterogeneity of impacts and thereby better inform future program practice and design.

Mediation of Impacts

Programs are effective if they are implemented well, if their implementation modifies the practices of program staff, and if these changes in practice help to generate skills, dispositions, and experiences that produce favorable outcomes for participants. We define as *mediators* those aspects of program implementation, staff practice, and short-term changes in participants' knowledge, skills, attitudes, or behavior that are (a) outcomes of random assignment and (b) predictors of participants' long-term success. These are often regarded as mechanisms through which programs produce long-term benefits. Our definition allows for the possibility that the association between a mediator and the outcome can depend on treatment group assignment. For example, in a job training program, it may be the case that a new program is better at fostering high levels of motivation to work than would arise in the control condition *and* that high levels of motivation are more predictive of later earnings among program group participants than among control group participants.

Program sites that are better at generating a program's mediators will, in theory, produce more favorable impacts on participants' outcomes. So heterogeneity of program impacts on mediators can, in principle, explain heterogeneity of program impacts on participants' outcomes. Nonetheless, although most programs are founded on a theory (which is implicit more often than it is explicit) regarding how program operations influence key mediators and how these mediators promote long-term outcomes, few rigorous large-scale evaluations have explicitly tested these theories, leaving heterogeneity of impacts largely unexplained.

Theories of moderation and mediation can thus help to explain heterogeneity of program impacts. Conversely, observed heterogeneity of program impacts can be used to test these theories. Thus knowing how sites vary in their program impacts creates an opportunity to learn about what works best and why; likewise, the power of a theory can be evaluated in terms of its capacity to account for observable heterogeneity of program impacts.

The Importance of Multisite Trials

In the pages below, we share key ideas that emerged from the two conferences noted above. Specifically, we summarize what is already known about how to learn *about* and *from* variation in program impacts, and we describe the new knowledge that is required in order to launch an ambitious agenda for research on these topics. But first, we need to say a little about the multisite RCT and its importance for this enterprise.

Multisite trials are RCTs in which sample members are randomly assigned to a new program or a control group within each of a number of sites. Sometimes sites are comparatively few in number. For example, the well-known Moving to Opportunity (MTO) experiment was conducted in five cities: Baltimore, Boston, Chicago, Los Angeles, and New York. Within each city, program applicants from local public housing projects were randomly assigned to receive either (a) a standard Section 8 housing voucher that could be used to rent private market housing (treatment one); (b) a limited Section 8 housing voucher that could be used to rent private market housing in neighborhoods with poverty rates below a specified level (treatment two); or (c) no housing voucher (control status; Katz, Kling, and Liebman, 2000). A key aspect of heterogeneity in this study is variation in program impacts *across* its five sites (Burdick-Will et al., 2011). In a much broader trial, the national Head Start Impact Study selected more than 350 oversubscribed Head Start centers and within each center randomly assigned a small number of eligible program applicants to the program or a control group (U.S. Department of Health and Human Services, 2010). Despite the small sample sizes per site for this RCT, one can learn a great deal about heterogeneity of Head Start impact by studying its cross-site variation (Bloom et al., 2014).

Although the methods of statistical analysis differ somewhat depending on the number of sites and participants per site, all multisite trials can be regarded as a "fleet" of randomized experiments. Hence they are uniquely suited to the study of impact heterogeneity. This is because sites often vary in interesting ways with respect to implementation of the program being tested and the kinds of participants being served. Moreover, the multisite trial is prevalent if not ubiquitous. For example, in her survey of 175 RCTs conducted by IES since 1994, Spybrook (2013) found that more than two-thirds were multisite trials. For these reasons, the present paper focuses on multisite trials. Many of the core ideas that it presents, however, can be adapted to single-site trials and to single- or multisite quasi-experiments.

Our Aims for This Paper

We begin by discussing what can be learned from multisite trials *about* heterogeneity of impacts. The idea here is to first detect, quantify, and describe impact heterogeneity without formulating or testing theories that might explain it. We thus focus first on methods for achieving the following analytic objectives:

- Estimating the average impact of a program across a population of persons or sites when the program's impacts vary
- Detecting variation in program impacts among persons within sites
- Detecting and quantifying variation in average program impacts between sites
- Estimating the average impact of a program at each site
- Estimating the cross-site distribution of site-specific average impacts

We next focus on how to learn from heterogeneity of impacts in order to achieve the following objectives:

- Posing and testing theories about why some individuals benefit more than others from a new program and why some program sites produce larger impacts than others (i.e., theories about impact moderation)
- Posing and testing theories about the mechanisms through which a program produces its impacts (i.e., theories about impact mediation)

In pursuing these aims, researchers often must confront the problem of noncompliance with random assignment noted earlier. For example, in MTO, only 47 percent of the sample members who were randomized to the limited Section 8 housing voucher actually used it (Katz,

Kling, and Liebman, 2000). Likewise, in lottery-based studies of charter schools, some winners of the lottery for a given charter school do not enroll in that school and some losers of this lottery end up enrolling in that school or in some other charter school.

In studies characterized by such noncompliance, researchers have come to distinguish between two types of causal effects or program impacts. The first is the impact of *assignment* to a program, which is typically called the intent-to-treat effect, or ITT (e.g., Yau and Little, 2001). The second is the impact of *participating* in a program for sample members who were induced by random assignment to participate. This is often called a Complier Average Causal Effect, or CACE (Yau and Little, 2001), or a Local Average Treatment Effect, or LATE (Angrist, Imbens, and Rubin, 1996). We shall use the term CACE. Knowledge about ITT and CACE are both important for advancing program theory, practice, and policy. However, estimating a cross-site mean and variance for the latter is more challenging and involves more assumptions than is the case for the former. Furthermore, more is known about how to study the former than the latter.

Section 2 of our paper considers how we can learn about variation in ITT impacts. Section 3 considers the problem of learning about variation in CACEs. Section 4 considers the problem of posing and testing theories in order to learn *from* variation in impacts.

Section 2

Learning *About* Variation in the Impacts of Program Assignment

To lay a conceptual and methodological foundation, we begin with the individual-level average and variance of ITT impacts *within* a single site. We then discuss how to use multisite RCTs to study impact variation *within and between* sites.

Studying ITT Impacts for a Single-Site RCT

We adopt the "potential outcomes" framework for causal inference, which is used widely in applied statistics.[4] We set $T = 1$ if a sample member is randomized to a new program (or treatment) and $T = 0$ if he is randomized to a control group. Each participant has two potential outcomes: $Y(1)$ if he is assigned to the program and $Y(0)$ if he is assigned to the control group.[5] By definition, the causal effect of a program for a person is the *difference* between his two potential outcomes:

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

It is not possible to calculate a program impact for an individual because we can observe only one of his two potential outcomes. If he is assigned to the program, we can observe $Y(1)$ but not $Y(0)$, and if he is assigned to the control group, we can observe $Y(0)$ but not $Y(1)$. We can, however, estimate the *average* program impact for a population of individuals from data from an RCT. This is possible because random assignment ensures that each sample member's potential outcomes are unrelated to assignment to the treatment or control group. Hence there is no "selection bias." The population-average causal effect of intent to treat (β_{ITT}) is thus

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

where E denotes an "expectation" or population average. In other words, the average causal effect of intent to treat for a population, β_{ITT}, equals the *difference* between the average outcome that would result if the entire population were assigned to the program, $E[Y(1)]$, and the aver-

[4]Versions of this framework have been attributed to Neyman (1923/1990), Roy (1951), Heckman (1979), Rubin (1974, 1978), and Holland (1986).

[5]Specifying these two potential outcomes for each person assumes that they cannot be influenced by the outcomes or program assignment of others, which is known in the literature as the Stable Unit Treatment Value Assumption, or SUTVA (Rubin, 1986).

age outcome that would result if the entire population were assigned to the control group, $E[Y(0)]$.[6]

Although we can estimate the average impact of assignment to a program for a single site, we cannot estimate the variance of these 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

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

which implies that

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

where $Cov[Y(0), B]$ is the individual-level *covariance* between control group outcomes and program effects. 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 $Var(B)$.

Further investigation (see Bryk and Raudenbush, 1988, and Bloom et al., 2014) reveals the following guidelines:

- If a program group and control group have different individual-level outcome variances, we can conclude that program impacts vary across individuals.[7]
- If a program group and control group do not have different individual-level outcome variances, we cannot conclude that program impacts do not vary across individuals.[8]
- If the program group variance is smaller than the comparison group variance, we can conclude that it produces larger-than-average impacts for persons

[6]Equation 2 assumes that scaling up the program to include all population members would not change the program's benefit for each participant. This assumption could fail, for example, if the entire population of unemployed persons were assigned to an effective job training program. In this case, each program participant would be competing in the job market against all other participants who also obtained the program's benefits, possibly reducing the labor-market advantage conferred by these benefits. This assumption is a consequence of SUTVA (note 5).

[7]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.

[8]The expression $Var(B) + 2Cov[Y(0), B]$ can be near zero even if the individual-level impact variance, $Var(B)$, is positive. This can occur when persons who would fare less well than the average without the program benefit by more than the average from it (i.e., when $Cov[Y(0), B]$ is negative).

who would fare worse than average without the program (i.e., program effects are compensatory).[9]

- If the program group variance is larger than the comparison group variance, we cannot conclude that the program produces larger-than-average impacts for persons who would fare better than average without the program.[10]

In summary, then, a single-site RCT can provide full information about the average impact of program assignment at a single site and limited information about the heterogeneity of this impact across individuals at that site.

Studying ITT Impacts Across Multiple Sites

As noted earlier, a multisite trial is fleet of independent RCTs, which makes it possible to estimate the impact of assignment to a new program at each site. In principle, this means that we can address a series of important questions about impact heterogeneity by comparing impact estimates *across sites*. For example, we can ask whether impacts vary across sites and how large this variation is, and whether sites serving high-risk persons produce effects that are larger or smaller than those produced by sites serving lower-risk persons. We can even envision displaying the cross-site distribution of impacts, making it possible to identify the most and least effective sites. We can achieve these goals by capitalizing on random assignment within each site without the need to impose strong assumptions.

In practice, however, Spybrook's (2013) review suggests that past multisite trials have rarely pursued these questions. One possible explanation for this is that researchers regard site samples to be too small to draw strong conclusions about cross-site impact variation. Researchers may also be concerned about "capitalizing on chance" when testing the statistical significance of impact estimates for multiple sites. However, statistical advances over the past several decades make it possible to efficiently analyze impacts that vary across sites, producing credible summaries of evidence about site-specific impacts even when site samples are small.

In what follows, we consider how to conduct multisite analyses of impact variation in a way that expands on our discussion of single-site analyses. Recall that for single-site studies, we considered the average impact of a program, the variance of program impacts across individuals, and the covariance between individual impacts and their "untreated counterfactual outcomes" (the outcomes they would experience if not assigned to the program). We now consider

[9]Equation 5 implies that the program group outcome variance can be smaller than the control group outcome variance only if $Cov[Y(0), B]$ is negative.

[10]If program effects vary, the program group outcome variance can exceed the control group outcome variance even if $Cov[Y(0), B]$ is zero.

analogous site-level quantities: the cross-site average impact, the cross-site variance (or standard deviation) of impacts, and the cross-site covariance between impacts and untreated counterfactual outcomes.

We shall see that the tasks of defining the parameters of interest in a study, designing the study, and analyzing data from the study become somewhat complex when we anticipate that program impacts vary within and across sites. We consider these issues first in the familiar case of studying the population-average impact.

The Population-Average ITT Impact

The Problem of *Defining* the Population-Average Impact When Impacts Are Heterogeneous

When program impacts vary across persons and/or sites, there are different ways to define the population-average impact. On the one hand, we might define the population of interest to be a population of *sites*. For example, we might consider the population of all local Head Start centers in the United States to be the focus of our analysis and therefore might want to know the average impact for that population of sites, counting each site equally regardless of its number of program-eligible children.

On the other hand, if we were interested in generalizing findings to the population of eligible Head Start *children*, we would want to address a different question: What is the average program impact for the U.S. population of program-eligible children? Statisticians often define a parameter of interest in a study to be a "target of inference" or "estimand." Ideally, researchers should be explicit about their estimands of interest before designing a study.

Suppose that, prior to designing a study, we have information about the number of sites, call it J^*, in the population of all sites of interest, and that we also have information about the number of eligible persons, call it N_j in each site j, there being $\sum_{j=1}^{J^*} N_j$ persons in the entire population. As in Equation 1 above, each person i in each site j possesses a potential outcome $Y_{ij}(1)$ if assigned to the program and a second potential outcome $Y_{ij}(0)$ if assigned to the control group. Thus, we can define, for each person, the causal effect of assignment to the program (relative to the control) as

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

The site-average causal effect is then

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

If we wish to generalize to a population of sites, we can define our estimand of interest as the unweighted "mean of site means," that is,

$$\beta_{unweighted} = \frac{\sum_{j=1}^{J^*} B_j}{J^*} \,. \tag{8}$$

In Equation 8, each site counts equally in contributing to the estimand, regardless of how many persons are located in that site. In contrast, if we wish to generalize to the population of persons, we can define our estimand as the average taken over all persons, that is,

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

If the impact of the program is invariant across all sites, or if the number of participants per site is constant, the unweighted average estimand in Equation 8 and the weighted average estimand in Equation 9 will be identical. But neither of these conditions seems plausible. Hence, our two estimands could be quite different, particularly if programs in sites that serve large populations of persons are more or less effective, on average, than are programs in sites with small populations.

The Problem of Designing a Multisite Trial When Impacts Are Heterogeneous

The choice of estimand will strongly influence the optimal design of a study. To see why, let us suppose for simplicity that the cost of sampling children within sites is invariant across sites, the cost of studying program group members equals the cost of studying control group members, and the variance of the outcome within experimental groups does not vary between experimental groups or across sites.

If the estimand of interest is the unweighted mean of site means (Equation 8), it would be optimal first to draw a simple random sample of sites from the population of sites, then to draw a simple random sample of n persons within each site, and then to assign those persons with equal probability to the program or control group. This would be a perfectly balanced design (with $n/2$ persons in each experimental group from each site).

However, if the estimand of interest is the weighted mean (9), a good option would be to again draw a simple random sample of sites from the population of sites but then to set the sample size for each site to be proportional to the number of program-eligible persons located in the site.[11]

Many other design options are possible. One might oversample certain sites, e.g., those that serve small but scientifically interesting subpopulations. And one might oversample particular subpopulations within sites. Yet one might still define Equation 8 or, alternatively, Equation 9 as an estimand of interest.

Unfortunately, evaluators rarely have the luxury of implementing probability samples of sites or of persons within sites and instead select samples of convenience. Yet evaluators typically conceive the sites in their study to represent a larger universe of similar sites that might take up the program, and evaluators want their findings to apply not only to the specific persons in their sample but rather to a larger universe of similar persons who might benefit should the program be found effective. In this setting, care must still be taken during the design phase regarding choice of the estimand. One might contemplate generalizing findings to a larger universe of sites and hence weighting the contribution of each site equally when defining the estimand of interest (as in Equation 8), or generalizing to a larger universe of persons represented by the persons in the sample, hence adopting a weighted average as the estimand (like Equation 9) and substituting the known site sample size n_j for the desired site population size N_j. In this way, each sampled person contributes equally to the overall mean impact.

The Problem of *Estimating* the Mean Impact in a Multisite Trial When Impacts are Heterogeneous

Having carefully defined the estimand of interest and having designed the study accordingly, we now face the question of how to estimate the desired mean impact given the data that are collected. Now we must confront the fact the sample size n_j per site might vary significantly from what the design intended. Moreover, the fraction of persons assigned to the program (as opposed to the control group) will typically vary from site to site even if the research design held this fraction constant across sites. The fraction assigned to the program is known in the statistical literature as a "propensity score" (Rosenbaum and Rubin, 1983). In a multisite trial, the propensity score 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

[11] Alternatively, we could stratify sites by the size of their eligible population N_j, sample sites with probability proportional to their size N_j, and then draw equal-size samples within each site.

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 some 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 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 $\overline{Y}_{1j}$) for its program group members. Similarly, we can estimate U_{0j} for site *j* from the sample mean outcome (call it $\overline{Y}_{0j}$) for its control group members. This resulting estimate of the average program impact for site *j* is a simple difference of means $\hat{B}_j = \overline{Y}_{1j} - \overline{Y}_{0j}$ 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 the discussion simple, we now confine our attention to the case where the unweighted "mean of site means" defined by Equation 8 is the estimand of interest. The logic of our inquiry would remain the same had we chosen to focus on Equation 9. This logic entails clarifying the bias and precision of conventional methods when used to learn about the estimand of interest.

The "Site Fixed-Effects" Estimator

A common analytic strategy for estimating the average ITT effect in a multisite trial is the site fixed-effects estimator. One writes down a standard regression model where the outcome is Y_{ij}, the key predictor is treatment group assignment T_{ij}, and the between-site variation is removed by the inclusion of site fixed effects. We can write this model as

$$Y_{ij} = \beta T_{ij} + \alpha_j + e_{ij}. \qquad (10)$$

α_j is a site-specific fixed intercept and e_{ij} is a random error having zero mean, and, for simplicity, we will assume that e_{ij} has a constant variance σ^2.[12] The resulting estimator is readily shown to be a weighted average of site-specific impact estimates $\hat{B}_j = \overline{Y}_{1j} - \overline{Y}_{0j}$:

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

where

$$w_j = n_j \overline{T}_j (1 - \overline{T}_j).$$

Here n_j is the sample size for site *j* and $\overline{T}_j$ is the propensity score (the proportion of sample members assigned to the program) in site *j*. Interestingly, the weight w_j for site *j* is inversely proportional to the reciprocal of the sampling variance of the site-specific estimate $\hat{B}_j$.[13] This estimator is optimal (it is unbiased and has minimum variance) when site-specific impacts are homogeneous.

However, things change when site impacts are heterogeneous. Now the fixed-effects estimator will be biased for the unweighted population mean of site means (Equation 8) whenever the "true" site-specific impact B_j is correlated with w_j (see Raudenbush, 2014). This implies that if the sample size n_j or the propensity score $\overline{T}_j$ is statistically associated with B_j, we have a risk of bias.

Using Simple Averages

It is natural to think that we can greatly simplify this problem by using one of several straightforward averages. For the case at hand, in which we are generalizing to a population of sites, consider the very simple estimator

[12]Recall from our discussion of heterogeneity of impact within sites that we will often find that the within-site variance of program group and control group members will differ, but we ignore this complication here in order to focus on key issues.

[13]The sampling variance is $Var(\hat{B}_j \mid B_j) = \sigma^2 / [n_j \overline{T}_j (1 - \overline{T}_j)]$.

$$\hat{\beta}_{unweighted} = \frac{\sum_{j=1}^{J} \hat{B}_j}{J}. \quad (12)$$

This "unweighted" sample average is sure to be unbiased when we want to count all sites equally! The problem, however, is that it could be very imprecise because we give sites with very small samples the same importance as sites with very large samples.

In sum, in the case where we wish to generalize to a population of sites, the site fixed-effects estimator (Equation 11) is unbiased and precise when site-specific impacts are homogeneous but potentially biased when they are heterogeneous. In contrast, the unweighted average (Equation 12) is always unbiased but potentially very imprecise when sample sizes vary substantially.

A Simple Random Coefficient Estimator

Having to choose between the site fixed-effects estimator in Equation 11 and the unweighted site average in Equation 12 creates a forced choice of the kind that many statisticians find objectionable. Is there not a flexible alternative that puts our analysis on a continuum between these two extremes and is sensible across a range of cross-site variation in impacts and sample sizes? The answer is a qualified "yes."

To see how, consider a hierarchical linear model (Lindley and Smith, 1972; Dempster, Rubin, and Tsutakawa, 1981; Raudenbush and Bryk, 2002), or HLM, which specifies site impacts that vary randomly around a population grand mean (β) with a variance τ^2. If τ^2 were known for the population of sites and V_j were known for each site, we would have another estimator with site weights equal to the reciprocal of the *total* variance of their site impact estimate $(\tau^2 + V_j)^{-1}$. We therefore would have an estimator having the same form as the fixed-effects estimator but with site-specific weights:

$$w_j = (\tau^2 + V_j)^{-1} = \left[\tau^2 + \frac{\sigma_2}{n_j \bar{T}_j (1 - \bar{T}_j)} \right]^{-1}. \quad (13)$$

When site-specific impacts are homogeneous ($\tau^2 = 0$), this weight is the same as that for the fixed-effects estimator ($w_j = \frac{1}{V_j} = n_j \bar{T}_j (1 - \bar{T}_j)$) in Equation 11, which is optimal for homogeneous impacts and heterogeneous site sample sizes. But 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. In this way, incorporating the "heterogeneity parameter" τ^2 into our weights creates a continuum of estimators that lie between

the two extremes posed by the site fixed-effects estimator and the unweighted estimator. We can conjecture that, by putting us in a reasonable place on this continuum, the HLM estimator is superior to either of its two extreme alternatives in terms of root mean squared error.

However, recall that our answer about the preceding dilemma was a "qualified" yes and there are two important qualifications. First, our reasoning in the previous paragraph 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 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 site fixed-effects estimator and the unweighted estimator. Still, we will not land outside this continuum, so it is possible that in many cases this estimator has acceptable properties.

The second qualification is that, like the fixed-effects estimator, the HLM estimator is biased when its weights are correlated across sites with true impacts. This bias will tend to be smaller than that for the fixed-effects estimator, but neither bias vanishes as the number of sites increases. We anticipate that it will be possible to place a bound on potential bias in many applications by including information about the association between the weights w_j and the true impacts B_j, and even to correct for this bias, but this is a topic for future study.

The Cross-Site Variance of ITT Impacts and the Correlation Between Program Impacts and Control Group Means

Defining the Cross-Site Variance

We have made the argument that, for multisite trials, one should estimate the cross-site variance, or standard deviation, of mean program impacts as well as the cross-site mean. But how do we do this?

First, just as in the case of the overall average impact, we need to take care to define our estimand. An intuitively appealing definition of this cross-site variance is the simple average squared difference between site-specific impacts B_j and the unweighted cross-site average impact β. This variance, which is the cross-site variance in site mean program impacts, can be written as

$$\tau^2 = \frac{\sum_{j=1}^{J^*}(B_j - \beta)^2}{J^*}. \qquad (14)$$

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

$$\tau^2 = \frac{\sum_{j=1}^{J^*} N_j (B_j - \beta)^2}{\sum_{j=1}^{J^*} N_j}, \tag{15}$$

where β is now defined as the weighted mean in Equation 9. It may seem counterintuitive to define a variance as a weighted average of the difference between site-specific mean program effects and a population mean program effect. However, as shown in Appendix A, Equation 15 can be very useful in enabling us to estimate the *portion* of the total variation in program impacts across *individuals* in a population that reflects the differences in mean impact of the sites in which they are located.

Estimating the Cross-Site Variance of ITT Impacts

Few studies that we have reviewed have attempted to estimate the cross-site variance of ITT impacts, and we have not found methodological literature that provides guidance for how to estimate Equation 14 or 15 from sample data. Clearly, the optimal method will depend on how the study is designed, just as in our discussion of estimating the mean ITT impact. However, we reason that a broad class of estimators will have the form:

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

where w_j is a weight for each site's contribution to the variance estimate. Negative estimates will be set to zero. The idea here is that $(\hat{B}_j - \hat{\beta})^2 = [\hat{B}_j - B_j + B_j - \hat{\beta}]^2$ gives us, for each site, an unbiased estimate of the sampling variance $V_j = Var(\hat{B}_j - B_j)$ and the cross-site variance $\tau^2 = Var(B_j - \beta)$. So we subtract the estimated sampling variance $\hat{V}_j$ from $(\hat{B}_j - \hat{\beta})^2$ and call this difference the site-specific estimate of τ^2. We then compute a weighted average of these site-specific estimates using a weight w_j.

Suppose now that we have a convenience sample, yet we regard our sites as representing an interesting, if undefined, universe of similar sites and our variance estimand is the unweighted one (Equation 14). Then we would be inclined to set $w_j = 1$ in computing the estimate defined by Equation 16. In contrast, suppose that again we have a convenience sample, but we

regard the persons in our study 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 $w_j = n_j$ in calculating our estimate (Equation 16).

How reasonable are such approaches? Here we will consider the case of the unweighted estimand (Equation 14) and hence set $w_j = 1$ when calculating our estimate. This estimate would be "consistent," that is, it will always converge to the correct value as the number of sites in the sample becomes ever larger. However, the unweighted approach may be very imprecise, particularly if small sites produce outlying estimates $(\hat{B}_j - \hat{\beta})^2 - \hat{V}_j$, because these outliers will be given weight equal to that of far more precise estimates coming from larger sites. This would be particularly problematic if the number of sites is modest.

An alternative is a hierarchical linear model analysis based on maximum likelihood. Such an approach uses iteratively reestimated least squares to obtain, at iteration *m*+1,

$$\hat{\tau}^{(m+1)} = \frac{\sum_{j=1}^{J} w_j^{(m)}[(\hat{B}_j - \beta^m)^2 - \hat{V}_j^m]}{\sum_{j=1}^{J} w_j^m} \qquad (17)$$

where $w_j^{(m)} = (\hat{\tau}^{(m)} + V_j^m)^{-2}$. (See Appendix B.) Here the weight is inversely proportional to the reciprocal of the *square* of the variance of the site-specific estimate $\hat{B}_j$. This is optimal when we assume no correlation between the weight and site-specific impact estimates because it appropriately weights down outliers coming 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 with the unweighted estimator. We need to learn more about the bias-precision trade-off that can arise in practice from this approach.

Estimating the Cross-Site Covariance or Correlation Between Treatment Effects and Control Group Mean Outcomes

What is the cross-site correlation between program impacts and control group mean outcomes? This is another question that is rarely asked but potentially quite informative. If sites with high control group mean outcomes produce larger effects than do sites with low control group means (the correlation between these two quantities is positive), we can conclude that the program will tend to increase cross-site outcome inequality. Alternatively, if this correlation is negative, the program will tend to reduce outcome inequality across sites. The impact of this correlation on the overall distribution of outcomes across all children is described in Appendix A. It can be estimated without imposing strong theory or assumptions. Yet conventional meth-

ods do not, in general, provided an unbiased estimate, as is the case for the cross-site mean and variance of program effects.

Once again, selecting the estimand is important, and we can define one that is unweighted or one that is weighted by N_j. To see how we might estimate the covariance, suppose that we could actually compute the true mean impact (β) as well as the site-specific impact B_j, and suppose we also knew the true mean untreated counterfactual outcome (μ_0) for the population of sites as well as the true site-specific untreated oucome U_{0j}. Then, for each site, we could compute the product $(B_j - \beta)(U_{0j} - \mu_0)$, 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 error covariance, $\hat{C}_j$, to obtain a weighted average:

$$\hat{\tau}_{B0} = \frac{\sum_{j=1}^{J} w_j[(\hat{B}_j - \hat{\beta})(\hat{U}_{0j} - \hat{\mu}_0) - \hat{C}_j]}{\sum_{j=1}^{J} w_j}, \qquad (18)$$

where τ_{B0} is the cross-site *covariance* between mean program impacts and mean counterfactual untreated outcomes. Again, we have choices; for example, we could set $w_j = 1$ or $w_j = n_j$, or instead we could use maximum likelihood estimation of a hierarchical linear model (although the latter is more complicated and beyond the scope of the present discussion). Thus, more needs to be learned about how these methods work in practice.

Studying Site-Specific Impacts

Knowing the cross-site mean and variance of program impacts is very useful, but more refined questions also arise. For example, looking across sites, how large are the largest impacts and which sites produce them? How small are the smallest impacts and which sites produce them? In what fraction of sites is the impact negative? While such questions are rarely asked in practice, statistical methods exist for answering them. Here graphical tools are important to reflect the uncertainty about the answers obtained and to check the underlying assumptions of the analyses involved.

If we could observe the site-specific impact B_j for each site, we could display its cross-site frequency distribution and determine, for example, the 10th, 25th, 75th, or 90th percentile values. The problem with doing so 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

amount of cross-site impact variation that exists and provide a highly misleading sense of the percentile distribution of true impacts. This problem arises because the cross-site distribution of conventional OLS impact *estimate* $\hat{B}_j$ reflects two sources of variation: (1) cross-site variation in true impacts and (2) cross-site variation in impact estimation error. This problem can also cause us to greatly exaggerate how effective or ineffective the program might be at the most and least effective sites. Furthermore, the problem can cause us to misrepresent the rank order of impacts for different sites, particularly if sites vary markedly in their sample sizes. In this case, sites with the smallest samples would tend to have the largest sampling error and thus the most extremely large or small values of $\hat{B}_j$, even if true values of B_j vary little.

Perhaps the most popular method for addressing this problem is to compute, for each site, an "empirical Bayes" estimator of the form:

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

This empirical Bayes impact estimate, B_j^*, is a weighted average of the site-specific OLS impact estimate, $\hat{B}_j$, and the overall mean impact estimate, $\hat{\beta}$. The weight accorded the site-specific estimate is its reliability:

$$\lambda_j = \tau^2/(\tau^2 + V_j)\,. \tag{20}$$

Sites with large samples will tend to produce $\hat{B}_j$ values have a small sampling variance, V_j, and thus have high reliability. For those sites, $\hat{B}_j$ will receive a large weight. For small-sample sites with large V_j, reliabilities will be smaller, and our estimate of the true site impact will "shrink" toward the grand mean, $\hat{\beta}$. There is considerable reason to believe that these empirical Bayes "shrinkage estimators" will, on average, better predict true site-specific impacts under cross-validation (see Morris, 1983, for a review).

A key problem here, however, is that the shrinkage estimator depends on knowing the true value of τ^2, the variance of the true impacts. In practice, τ^2 must be estimated from data. In studies with few sites or very small samples per site, the estimate of τ^2 can be quite imprecise and will tend to be biased toward zero. Imprecise estimates of τ^2 will give us imprecise estimates of the reliability (Equation 20), leading us to "shrink" site-specific estimates $\hat{B}_j$ too much or too little. How can we diagnose this problem for a particular data set? Rubin (1981) provides graphical tools for checking the sensitivity of inferences about site-specific impacts to uncertainty about τ^2.

The site-specific empirical Bayes estimators B_j^* are, in a sense, optimal for each site, given reasonably large sample sizes of sites and participants.[14] However, Louis (1984) noted that a histogram of empirical Bayes estimators will *understate* the variability of true impacts B_j. For purposes of generating a histogram that approximates the distribution of these true impacts, the empirical Bayes approach will have "over-shrunk" the site-specific estimates $\hat{B}_j$. Therefore, if we wish to approximate the histogram, we can use "constrained" empirical Bayes estimators (Bloom et al., 2014).

[14] 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 for a specific site.

Section 3

Learning About Variation in the Impacts of Program Participation

So far we have talked about the impact of random assignment of participants to a new program, known as an ITT effect. If everyone assigned to the new program participates in it and if no one assigned to the control group participates in the new program, we have an ideal situation known as "perfect compliance" with random assignment. Unfortunately, in large-scale field trials, perfect compliance rarely occurs. Instead, we have partial compliance caused by two forms of noncompliant behavior. First, some participants who are assigned to the program will fail to participate. For example, in the Moving to Opportunity experiment (Kling, Liebman, and Katz, 2007), families living in public housing were assigned at random to receive a voucher that could be used to pay rent in a low-poverty neighborhood. However, only 47 percent of the families who were assigned to receive the voucher actually used it. Second, participants assigned to the control group end up in the new program. For example, in studies of charter schools, lottery winners are invited to attend the school. Lottery losers, who are often placed on a waiting list, may actually enroll in that charter school anyway or enroll in some other charter school. Whenever we have partial compliance, the ITT effect is not the same as the impact of participation in the program.

The ITT effect is typically of interest to policy: We would like to know, for each site, the average impact of the program on the persons for whom it was intended — that is, those assigned to the program. But we would also like to know the impact of actually participating in the program. For this second purpose, a problem of selection bias arises, even in the context of an RCT. This is because study participants (or their parents, counselors or teachers) shape the decision about whether to comply with random assignment. To cope with this selection bias when estimating the impact of program participation, methodologists have widely adopted the method of instrumental variables (see Angrist, Imbens, and Rubin, 1996, and Heckman and Vytlacil, 1998). For this approach, random assignment is conceived as an instrumental variable (IV) that induces a subset of youth to participate, and we can estimate the average impact of participation on those so induced (the "compliers") under several comparatively weak assumptions (Angrist, Imbens, and Rubin, 1996).

Let's take a look at how the IV method works in a single-site study with homogeneous impacts. Next, we will see how the analysis becomes more complex — and more interesting — when we allow treatment effects to vary across participants within such a study. We will then consider how to exploit a multisite trial not only to estimate the average impact of program participation but also to learn about heterogeneity of these effects across sites.

The IV Method in a Single-Site Trial with Homogeneous Impacts

It is easy to understand the conventional IV method by considering a simple causal model as in Figure 1. We begin by assigning participants at random to the new program ($T = 1$) or to the control group ($T = 0$). We expect random assignment to influence program participation, defined as $M = 1$ if a study subject participates in the new program and $M = 0$ if not. The impact of random assignment T on program participation M 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 Y is denoted as δ.

Figure 1

Single-Site Homogeneous Impacts

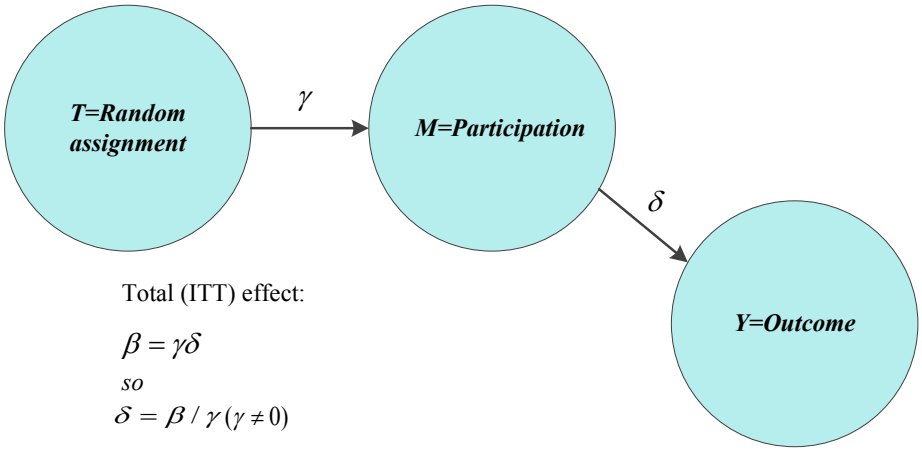


Figure 1 is a standard path model except there is no arrow between T and Y. This is known as the "exclusion restriction" — we have excluded the direct causal path between T and Y. This reflects a key assumption of the IV method: Any impact of assignment T on outcome Y works indirectly through program participation M. In the language of path analysis, participation M "fully mediates" the ITT effect, that is, the effect of T on Y, which we call β. That means that the ITT effect is just the "indirect" effect of T on Y that operates through M, and is therefore

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

The beauty of Equation 21 is that we can estimate δ (the impact of program participation M on outcome Y) without ever using M to predict Y. That is important because, as mentioned above, a model that uses M to predict Y is subject to selection bias whenever pretreatment personal char-

acteristics that predict the decision to participate in the program also predict the outcome. Instead, IV uses a two-stage approach. We can estimate γ (the impact of *T* on *M*) and β (the impact of *T* on *Y*) without bias because *T* is randomly assigned. We can then divide our estimate of β by our estimate of γ to obtain an approximately unbiased (consistent) estimate of δ :

$$impact\ of\ \ program\ \ participation = \delta = \frac{\beta}{\gamma} = \frac{ITT\ \ effect\ of\ \ T\ \ on\ Y}{ITT\ \ effect\ of\ \ T\ \ on\ M}, \gamma > 0. \qquad (22)$$

A key assumption of Equation 22 is that assignment to the program must increase the probability of participation, that is $\gamma > 0$. This is easily checked and it would be rare to find an experiment in which assignment to the program had little or no effect on participation, nor would such a case be of much interest.

The IV Method in a Single-Site Trial with Heterogeneous Impacts

Unfortunately, the simplicity of the conventional IV model (Figure 1) depends on a very strong assumption: that all participants respond the same way to treatment assignment. Although this assumption is frequently invoked implicitly in IV analyses, there is good reason to expect that for many applications the assumption will be false. People may vary in their motivation to participate, and some may benefit more than others from participating.

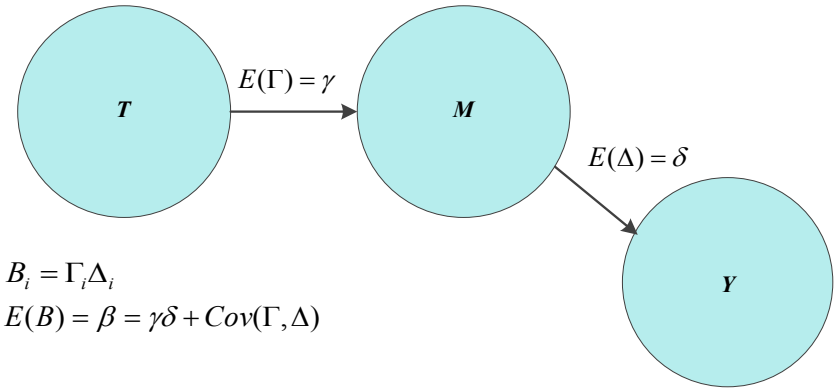
To represent heterogeneity in response to treatment assignment, we can construct a person-specific path diagram, as in Figure 2. Each participant has a unique causal effect of *T* on *M* denoted by Γ . We refer to Γ as "compliance," as it measures whether an individual's value of *M* changes in response to assignment to *T*. The population-average value of Γ is $E(\Gamma) = \gamma$. In the same spirit, the person-specific causal effect of *M* on *Y* is Δ . The average effect in the population is $E(\Delta) = \delta$. The "total effect" of *T* on *Y* for our participant is the product of the two causal effects Γ and Δ , as shown in Figure 2. Defining this total effect for our participant as B , we see then that $B = \Gamma\Delta$. The population average effect of *T* on *Y* is

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

The problem with Equation 23 is that the average effect of intent to treat, β , depends not only on the product of the two causal effects $\gamma\delta$ but also on $Cov(\Gamma, \Delta)$, the covariance between Γ and Δ . This covariance term implies that the population average effect of treatment assignment will be larger than $\gamma\delta$ when people who comply with the program (and thus have positive values of Γ) tend to benefit more than others from it (and thus have positive values o*f*

Figure 2

Single-Site Heterogeneous Impacts:
Person-Specific Causal Model



Δ). In contrast, if for some reason persons who would benefit more than average from program participation were less likely than average to participate, the average impact of program participation would be less than $\gamma\delta$.

How can we then accomplish our aim, which is to estimate the average impact of program participation, δ, when the treatment effect is heterogeneous? One option is simply to assume as an approximation that $Cov(\Gamma,\Delta) = 0$, i.e., that there is no (or negligible) covariance between compliance Γ and effect Δ. In this case, Equation 23 becomes equivalent to the conventional model in Equation 22, and we can identify the average treatment effect of *M* on *Y* as $\delta = \beta/\gamma, \quad \gamma \neq 0$. However, the "no compliance-effect covariance" assumption may be implausible in many cases, particularly in cases where individuals have some knowledge of how much they will benefit from *M* (i.e., they have at least partial knowledge of their Δ), and if that knowledge influences their compliance with assignment to *T* (their Γ) (Roy, 1951). Thus, the no-compliance-effect covariance assumption will seem to apply when either (a) participants (or other agents such as students or physicians making assignments to *M*) have no foreknowledge of likely benefits from participation, or (b) the benefits from participation are a constant, taking us back to the conventional model.

Rather than assuming no covariance between Γ and Δ, Angrist, Imbens, and Rubin (1996) developed an alternative approach. If *T* and *M* are binary, the authors reasoned that there must be four different kinds of people: 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). Thus for compliers, the impact on *M* of being assigned to the novel program is $\Gamma = 1 - 0 = 1$. *Never-takers* are persons who would fail

to participate in the new 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 this group $\Gamma = 1 - 1 = 0$. 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 assigned not to (so $M = 1$ if $T = 0$). Thus, for defiers, $\Gamma = -1$.

We might assume that there are no defiers. This is known in the literature as the *monotonicity* assumption, meaning that being assigned to the program cannot reduce the likelihood of program participation, hence $\Gamma \geq 0$ (Angrist, Imbens, and Rubin, 1996). Under this assumption, we have

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

Here δ_{CACE} is the causal effect of participation on persons for whom $\Gamma = 1$, the compliers. Hence the label "complier average causal effect." One problem for the interpretation of this effect is that the magnitude of δ_{CACE} depends on who complies with treatment assignment, and this will depend on how effective the program is in inducing participation. A program director who is very skilled at encouraging participation in the program in one study may generate a different δ_{CACE} than will a program director in another study who is less skilled at doing so, even if the population average impact of program participation is the same for the two studies.

In sum, if the gain from program participation varies among participants (as in Figure 2), the population average total effect of being assigned to the treatment is no longer a simple product $\gamma\delta$, unless we invoke the rather strong assumption of no covariance between compliance and impact. However, in the case of binary *M*, we can invoke the monotonicity assumption, in which case $\delta = \delta_{CACE}$, the complier average causal effect. The monotonicity assumption is generally much weaker than the no-covariance assumption.

Using Multisite Trials to Learn About Variation in CACE

Our aim now is to characterize the distribution of the CACE across sites in the same way that we did for ITT effects. First, we would like to obtain an estimate of the average CACE across all sites. Second, we would like to know whether and to what extent the site-average CACE impact varies. Third, we would like to know how large (and small) the CACE might be. And we would like to know whether the impact of program participation is high or low for compliers

who would fare poorly if assigned to the comparison group. This would tell us whether sites serving persons at the highest risk have the most (or the least) to gain from program participation.

Raudenbush, Reardon, and Nomi (2012) introduced statistical methods for estimating the mean and the variance of the CACE across sites. Rather than repeating the details of how to conduct this analysis, we refer the interested reader to that article. However, we would like here to emphasize a subtle but important point regarding the assumptions that must be met to justify the approach the authors describe. These assumptions depend upon how one defines the goals of the study. As is true for ITT, we need to define our estimands for CACE.

Recall that the conventional analysis of the ITT effect is a site-specific fixed-effects model, and that this approach works well so long as the site-specific impacts are homogeneous. However, as soon as we allow for heterogeneous treatment effects, we have to make a decision about how to define our population average. Although many definitions are possible, we contrasted two: (a) we define the population as the universe of all sites and seek to weight each site's true mean effect equally; or (b) we define the population as the universe of all persons from all sites and thus seek to weight each person's true program effect equally. Which definition we choose will influence how we decide to estimate the average impact as well as the variance of the impacts.

The logic is similar for the CACE. The conventional estimator is a two-stage least squares estimator with site-fixed effects, an approach that works well so long as the CACE is invariant across sites as displayed in Figure 1. As soon as the CACE is heterogeneous, as in Figure 2, we have to decide how to define the population average — and also the variance of the CACE.

Defining and Estimating the Overall Average CACE

Suppose we want to generalize to a population of sites and regard each site as equally representative of that population. So we want to estimate the unweighted *true* average CACE, that is,

$$\delta = \sum_{j=1}^{J^*} \delta_j / J^*. \tag{25}$$

Here δ_j is the CACE in site *j* and δ is the overall average CACE. To estimate δ, an intuitive approach is to first estimate the site-specific CACE as $\hat{\delta}_j = \hat{\beta}_j / \hat{\gamma}_j$, where γ_j is the compliance rate in site *j* and $\hat{\gamma}_j$ is a sample estimate. We could then compute an unweighted

average of $\hat{\delta}_j$ across all sample sites and then define this average as our estimate of δ. Unfortunately, we find that this would not work well in many of the multisite data sets we have analyzed so far because there will typically be many sites having an insufficient sample size to obtain a stable estimate of γ_j. This is particularly important for the CACE because in small sites we may obtain *by chance* a very low rate of compliance with randomization , $\hat{\gamma}_j$. When we use this quantity as a denominator to estimate δ_j for a site, we obtain an estimate with an extremely large magnitude that can seriously perturb the overall average.

As an alternative, we might begin with our unbiased estimate, 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? It turns out that we can do so only with a strong assumption, as described by Raudenbush, Reardon, and Nomi (2012). Specifically, we have to assume "no covariance between site-mean compliance and site-mean impact." We can see this by noting that Equation 22 can be written as

$$\begin{aligned} E(\hat{\beta}) = \beta &= E\left(\sum_{j=1}^{J} \gamma_j \delta_j / J \right) \\ &= \gamma\delta + Cov(\gamma_j \delta_j). \end{aligned} \tag{27}$$

The presence of the covariance term in Equation 27 means that we cannot simply divide both sides by γ to obtain the average impact of participating, that is, δ — unless we assume the covariance to be zero. Such an assumption implies that the fraction of people who comply with treatment assignment in site *j* is uncorrelated with the impact of participating in that site. This assumption would be false if effectively managed sites are good at convincing people to participate and also at generating positive effects.

However, suppose instead that we want to generalize to a population of persons so that the CACE of interest is the one that weights each site's estimate by the population size (Equation 9), and that we regard each person in our study as equally representative of that population, that is, the sample size n_j is proportional to the site-specific population size N_j. Corresponding to this, our estimand for the average CACE is

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

Note here that 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 contributions to the overall CACE is weighted by the number of compliers in that site. If n_j is proportional to the site-specific complier population size N_j, we can substitute n_j for N_j in Equation 28.

In this scenario, we can define the overall CACE as $\delta = \beta / \gamma$ without resorting to the strong assumption about no covariance between compliance and impact. To see this, note that our estimate of the overall ITT effect β has an expected value of $\beta = \delta\gamma$:

$$\begin{aligned} E(\hat{\beta}) = E\left(\frac{\sum_{j=1}^{J} n_j B_j}{\sum_{j=1}^{J} n_j} \right) &= E\left(\frac{\sum_{j=1}^{J} n_j \gamma_j \delta_j}{\sum_{j=1}^{J} n_j} \right) \\ &= E\left(\frac{\sum_{j=1}^{J} n_j \gamma_j \delta_j}{\sum_{j=1}^{J} n_j \gamma_j} \right) \left(\frac{\sum_{j=1}^{J} n_j \gamma_j}{\sum_{j=1}^{J} n_j} \right) \\ &= \delta\gamma \end{aligned} \tag{29}$$

The first term in the middle line of Equation 29 is the individual-level complier population mean effect of program participation (δ); the second term is the individual-level population mean compliance rate. It seems intuitive to want to estimate the average impact of the program for the entire complier population (Equation 29). And we can do so with comparatively mild assumptions.

Estimating Cross-Site Variation in CACEs

Raudenbush, Nomi, and Reardon (2012) describe several methods for estimating the cross-site variance of CACEs. Tackling this problem would lead us into a technical discussion beyond the scope of the present paper. However, the key principles follow from the logic of the previous paragraphs: How we define the variance will be critical in shaping our approach to

estimating it. We may define this variance as an average squared deviation in a population of persons or in a population of sites, or in some other way that reflects a specific sampling design, and this decision will guide the approach to estimation.

Estimating the Correlation Between the Site-Specific CACE and the Mean Outcome for Compliers in the Control Group

Recall that, in the case of the ITT effect, we were interested in whether sites that produce large ITT effects are those that serve participants who would fare badly — or well — in the absence of the new program. This correlation would tell us something about whether implementation of the program tends to decrease or increase equality of outcomes. Now we would like to know whether sites that produce large values of CACE are those in which compliers would have done poorly — or well — in the absence of the program. We can also envision asking whether noncompliers — "always takers" and "never takers" — do better or worse than compliers in the absence of the new program. This would tell us something about the kinds of participants who comply and thus benefit from the program. Answering these questions goes beyond the scope of the current paper but will likely be of great interest going forward.

Estimating Site-Specific CACEs

A final topic for future research concerns how best to estimate site-specific CACEs along with their quantiles — e.g., 25th and 75th percentiles. As mentioned, we might estimate site-specific ITT effects on the outcome and divide by the fraction of compliers, that is, compute $\hat{\delta}_j = \hat{\beta}_j / \hat{\gamma}_j$. However, we noted that these estimates will tend to be unstable in small samples. We believe that empirical Bayes shrinkage estimators, described above with respect to the ITT effect, will also be useful for CACE; this is a topic of considerable interest for future research.

Section 4

Learning *From* Variation in Program Impacts

So far we have discussed ways to study the cross-site distribution of program impacts. In so doing, we have focused on the cross-site average impact, the cross-site variance of impacts, the cross-site association between impacts and control group outcomes, and especially large or small site-specific impacts. Learning *about* this impact variation sets the stage for explaining it. The idea now is to propose and test theories about when and why a program works, that is, to learn *from* impact variation in order to deepen our understanding of the causal forces at work and how best to manipulate them to improve program practice.

Moderation

Which types of persons benefit most from a program and in what kinds of sites does the program work best? These two important questions are about *moderation* of program effects. For example, we would like 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 undertake further study about why the program does not work for certain types of persons. We would also like to know which program sites are most effective, possibly to spur intense investigation of practice in those sites or to frame general questions about why the program works when it does.

When addressing questions about such person-level and site-level moderators, it is important to recognize that they are almost always interdependent. Sites vary not only in the organizational conditions and practices that may be a key to program success but also in the composition of their client population. Hence, claims about best practice at the site level may be misguided because especially effective sites may overrepresent persons who are most likely to benefit from the program being evaluated.

In what follows, we define moderators of a program's impacts to be any characteristics of its clients or sites that (a) facilitate or inhibit the program's effectiveness and (b) 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. These questions are often addressed through exploratory analyses conducted after average program effects have been estimated. While such auxiliary analyses can enhance understanding, there are problems with this frequently *ad hoc, post hoc* approach.

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 that is a good thing. However, this knowledge does not necessarily imply that the program should make special efforts to target those 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 significant 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. For example, ethnic minorities disproportionately comprise low-income persons, and boys have higher risk than girls for certain behavioral difficulties. Making sense of a large number of findings for such overlapping subgroups can be quite difficult.

Third, subgroup membership may vary substantially across sites and thus be confounded with geographic or ecological features of sites like their urbanicity, neighborhood disadvantage, or school quality. These site-level differences may be the drivers of variation in program impacts which could be mistakenly attributed to differences in person-level risk factors. Conversely, person-level risk factors may be the drivers of variation in program impacts, which could be mistakenly attributed to differences in site features.

All of these concerns cry out for *a priori specification* of a theory about who stands to benefit from a given program and why. Suppose, for example, that a program aims to increase high-school graduation rates. It cannot appreciably increase graduation rates for students who are virtually certain to graduate without the program. In addition, there may be some students whose skills or prior grades are so low that the program cannot help them to graduate.

We now have plenty of evidence about which kinds of kids are most likely to drop out of school (Rumberger, 1995), so that one can envision developing a model that predicts the probability of dropping out in the absence of treatment. The evaluator might then stratify his sample based on this predicted probability or "prognostic score" and test for cross-strata impact differences.

Stratifying on a prognostic score has several major advantages. First, a prognostic score summarizes the predictive information in many different baseline characteristics, thereby greatly reducing the number of subgroup tests to be conducted. 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, or, in some other way, differentiating program practice with respect to subgroups with different prognostic scores. 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 av-

erage effect. A school dropout prevention program can reduce dropouts only for students who are at some risk of dropping out. Suppose that at-risk students constitute 50 percent 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 also can augment a prognostic score analysis in ways that further our understanding of impact variation. For example, with program group data we could estimate a model that predicts postprogram outcomes using individual baseline characteristics suggested by prior theory. Given randomization, the coefficients of this model for the program 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 would have fared with access to the program. We could then use the same logic to obtain a prognostic score for how each program group member would have fared 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 score pairs. By examining how program impacts vary across these strata *within sites* that represent both strata (described below), we can efficiently summarize evidence about person-level moderators.[15]

Site-Level Moderators

Knowledge about site-level moderators is potentially of great importance for developing program theory, policy, and practice. For example, we need to understand what organizational conditions must be in place if a new program is to be successful. 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 wish to estimate program impacts for subgroups of persons, we will want to estimate program impacts for subgroups of sites. Once again problems arise from the fact that there are many ways to define subgroups and thus there are many moderators to consider. But 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 will be much less precision for estimating impact differences across site-level subgroups than for estimating impact differences across person-level subgroups.[16] Consequently,

[15]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 et al. (2014) for a discussion of this problem and ways to avoid it.

[16]This assumes that we are using a site-level random-coefficients model to estimate and test impact differences across site-level strata.

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 that arises when studying program moderators is that site-level and person-level moderators are often confounded with each other. For example, sites with favorable organizational conditions might serve comparatively advantaged clients. Thus, what appears to be the influence on program effects of a site-level moderator might actually be the influence of person-level moderators, or vice versa. To address this problem, we could use a "double stratification" strategy. For example, sample members might be stratified into two subgroups — persons at high risk of failure versus all others — according to their person-level prognostic scores; and sites might be categorized according to a specific site-level moderator or several such moderators (e.g., sites with high unemployment rates versus all others and/or sites with high resource levels versus all others). It is then possible to split each site's sample into four groups: a high-risk program group and a high-risk control group; and a low-risk program group and a low-risk control group. In this case, some sites may have empty cells. Specifically, some sites have no low-risk treatment group members or no low-risk control group members or both. However, for all sites that have *both* 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. Different versions of this strategy are possible, depending on one's sample structure.

Mediation

Why does a new program work — or fail to work? Innovative programs are based on theories about how program operations generate short-term changes that, if sustained, produce long-term benefits. The short-term changes are called mediators. Mediators include shifts in organizational processes such as improved instruction or more effective staff collaboration. Such improved processes are hypothesized to produce changes in mediators measured at the level of the person, including, for example, young people's attitudes, behaviors, or skills that promote favorable long-term outcomes such as educational attainment, employment, earnings, or effective parenting. We therefore define mediators of program effects to be those aspects of program implementation, staff practice, and short-term changes in participants' knowledge, skills, attitudes, or behavior that are (a) outcomes of random assignment and (b) predictors of participants' long-term success. These are often regarded as the *mechanisms* through which programs produce long-term benefits.

Program sites that effectively generate relevant mediator values will, in theory, produce favorable impacts on participants' outcomes. So heterogeneity of program impacts on program mediators can, in principle, explain heterogeneity of program impacts on participants' outcomes. Nonetheless, although most programs are founded on a theory (which is implicit more often than it is explicit), few rigorous large-scale evaluations have systematically tested these theories.

If a program fails to produce long-term benefits, it may be because it failed to affect key mediators. For example, teachers who participated in a new professional development program may not have improved their instruction, so their students did not become more engaged in school, hence no change in educational attainment ensued. Alternatively, the program may have influenced its mediators as expected, but the hypothesized relationship between the program's mediators and outcomes failed to emerge. Understanding the source of program failure is one key way to learn how to develop better programs.

In other cases, the program may have produced positive long-term effects, but not entirely through the expected mediational processes. Understanding which mediators are crucial is important for designing new programs and for evaluating whether a new implementation of a program is achieving its short-term goals.

Methodological Challenges

Analysis of mediational processes is popular in social science and program evaluation. However, drawing valid causal inferences about mediation is very challenging. To see why, consider a study in which teachers are assigned at random to a professional development program with the aim of increasing instructional quality and student outcomes. Suppose that, on average, the program is successful in boosting student achievement. One would like to know whether the gains in student achievement are explained — or "mediated" — by measured increases in the quality of instruction. The mediational analysis would first ask whether the program boosted instructional quality. If teachers are assigned at random to the program, this part of the analysis is "protected" by randomization; the mean difference between instructional quality in the program group and the control group is an unbiased estimate of the causal effect of the program on instructional quality. Next, one seeks to know whether instructional quality affects student achievement. Establishing this causal link is especially challenging, however, because teachers are not assigned at random to instructional quality. It will typically be the case, for example, that teachers' pretreatment characteristics (experience, prior education, commitment, etc.) predict instructional quality. Such pretreatment 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 for program group members is different from its impact for control group members. If this is the case, mem-

bership in the program group or control group *moderates* the causal effect of the mediator on the outcome. Conventional methods of path analysis thus do not work (Holland, 1988; Robins and Greenland, 1992; Pearl, 2001). Presenters at the two recent national conferences described in the introduction of this article described three evolving statistical strategies for coping with these methodological challenges.

Multisite Multimediator Instrumental Variables Analysis

Sean Reardon presented at the two conferences an approach that exploits site-to-site variation in the impact of a program on mediators. The rationale for this approach is quite intuitive. If *M* is a mediator and *Y* is an outcome of interest, we would 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 do 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 there are large effects of random assignment to the program on M_1 but not in sites where there are large effects of random assignment to the program on M_2. Then we'd be inclined to 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.

Kling, Liebman, and Katz (2007) applied this approach to the Moving to Opportunity Study (MTO) described earlier. As noted, MTO randomly assigned eligible public housing residents in five cities to receive a standard Section 8 housing voucher, to receive a Section 8 housing voucher that could only be used in neighborhoods with poverty levels below a specified level, or to a control group that did not receive a voucher. A key hypothesized mediator of MTO was neighborhood poverty. The researchers reasoned that the opportunity to move would reduce neighborhood poverty, and available theory suggests that neighborhood poverty undermines the well-being of parents and children. Sites in which assignment to a voucher produced not only a high level of voucher use but also a reduction in neighborhood poverty were the sites that tended to produce large effects on outcomes, so the authors concluded that voucher use and neighborhood poverty were mediators of the impact of program assignment.

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 to the assumptions we described earlier when the aim was to identify the impact of participating in a new program (CACE or LATE). Indeed, program participation can be regarded as a mediator of the effect of program assignment, as described in Figure 2. In that figure, we see that assignment to the program increases the probability of participating in it; and participating in the program influences the outcome. As described earlier, a key assumption is that there is no "direct effect" of random assignment on the outcome (no arrow between random

assignment and the outcome). The idea here is that random assignment can affect the outcome only by inducing program participation. Recall that treatment assignment *T* is an instrumental variable that induces a change in the causal variable *M*, which in turn causes a change in *Y*.

The multisite, multimediator model extends this basic idea to the case of two or more mediators, as shown in Figure 3. Now our instrumental variable *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 instrumental variable 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 instrumental variable. 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 defined by Reardon and Raudenbush (2013).

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

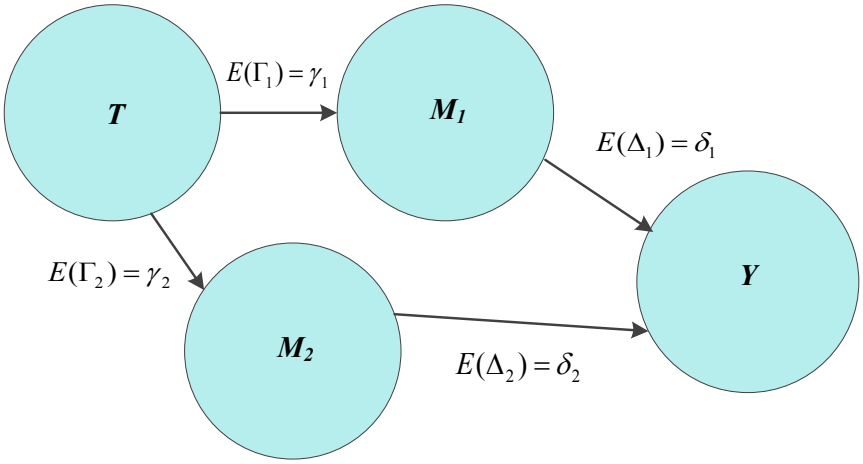
$$\begin{aligned} B_j &= \gamma_{1j}\delta_{1j} + \gamma_{2j}\delta_{2j} \\ &= \gamma_{1j}\delta_1 + \gamma_{2j}\delta_2 + e_j. \end{aligned} \tag{30}$$

Here δ_1 is the overall average impact of M_1 on *Y* controlling for M_2, and 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). The beauty of this setup is that we can estimate all three ITT effects without bias based simply on the random assignment of participants to *T*. So we do not have to worry about the possible pretreatment confounding that would otherwise arise from the fact that participants are not randomly selected into values of the mediators. However, this gift comes at the price of other assumptions:

1. We are assuming that there is no direct effect of *T* on *Y* (the exclusion restriction). This means that there are no unobserved mediators. That's why there is no intercept in Equation 30.

Figure 3

Multiple Sites, Two Mediators:
Person-Specific Causal Model



$$B_{1ij} = \Gamma_{1ij}\Delta_{1ij} + \Gamma_{2ij}\Delta_{2ij}$$

2. To regard Equation 30 as a regression model with identifiable parameters, there must be no bias associated with the error term $e_j = \gamma_{1j}(\delta_{1j} - \delta_1) + \gamma_{2j}(\delta_{2j} - \delta_2)$. This requires us to make assumptions similar to those described earlier in the case of studying the impact of program participation. Specifically, we can assume that the impact of *T* on each mediator is not related to the impact of either mediator on the outcome, or we have to define the impacts δ_{1j}, δ_{2j} as complier average causal effects (CACEs). This definition will apply when we assume "monotonicity," that is, that assignment to the program cannot reduce the value of either mediator. Reardon and Raudenbush (2013) describe these assumptions in detail.

3. There must be a nonzero impact of *T* on each mediator in one or more sites.

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.

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 a parallel mediation process, and the assumptions become stronger.

We can readily check assumptions (3) and (4) against the observed data, so these do not pose a strong challenge. Assumption (5) is based on program theory. The other assumptions, however, are quite strong and 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 in the case 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 future work that will extend these innovations to the case of two or more mediators.

We can conclude that the multisite, multimediator instrumental variable method opens up interesting new ways to exploit cross-site heterogeneity in order to draw conclusions about the impact of mediators on outcomes. However, this is a new method and we need to learn more about how failure of its assumptions influences its results.

Principal Stratification

The method of principal stratification (Frangakis and Rubin, 2002) can be used to estimate program impacts for endogenous subgroups of sample members that are defined in terms of post-random-assignment potential outcomes, like dropping out of school if randomized to a study's control group or being exposed in school to "world of work activities" if randomized to a study's treatment group. 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. Program impacts for such persons are "direct effects" 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 here 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 they have the same *pair* of potential values for a given mediator. 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. Previously we reasoned that if treatment assignment had no impact on program participation (a potential mediator), it could have no impact on the outcome, and we called this the "exclusion restriction." When adapting principal stratification to study mediation, we relax this restriction.

To illustrate this idea, consider a study in which teachers are randomly assigned to a new professional development program ($T = 1$) or to a control condition ($T = 0$) with the aim of improving student test scores, Y. Our mediator of interest is instructional quality ($M = 1$ if high and $M = 0$ if not high) and our outcome of interest (Y) is student test scores. Suppose we knew in advance (a) how each teacher's instructional quality would respond if he were assigned to the program and (b) how each teacher's instructional quality would respond if he were assigned to the control group. This is the *pair* of potential mediator values for each teacher. If we stratified teachers so that all teachers in the same stratum had the same pair of potential mediator values, we could estimate the program impact on student achievement for each stratum using random assignment of stratum members to the program or control group. We'd be interested in knowing whether the program produced an appreciable impact in strata for which the two potential mediator values were the same — for example, whether instructional quality was low under assignment to the program *and* under assignment to the control group. If this result were observed it would contradict our theory of mediation through instructional quality.

The problem, of course, is that we cannot observe the two potential mediator values for a teacher, so his principal stratum membership is unknown. However, as presented at the two conferences by Lindsay Page, 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. For further information about this approach, see the article by Page et al. (2014).

Note, however, that the use of principal stratification for mediation analysis contrasts sharply with the use of multisite instrumental variables analysis for this purpose. As noted, instrumental variables analysis uses the "exclusion restriction" (that program assignment only works through specified mediators and thus has no remaining "direct effect") to estimate the *indirect effects* of program assignment as they operate through the specified mediators. In contrast, the central use of principal stratification analysis for studying mediation is to estimate *direct effects* within strata for which there can be no indirect effects. Thus principal stratification might be regarded as a strategy for *falsifying* a mediation theory rather than a strategy for estimating a mediation process.

Sequential Randomization

A third innovative strategy for mediation analysis, described by Guanglei Hong at the Chicago conference, conceives of the mediation process as a *sequence* of randomized experiments (Robins and Greenland, 1992; Pearl, 2001). 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 the conventional one: 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, and 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, depending on his background and the study setting. For analyses based on sequential ignorability, the values of $M(1)$ and $M(0)$ are treated as *stochastic*. The probability that $M(1) = 1$ depends on a participant's past, the current situation, and whether he is randomly assigned to the program group or control group. However, given this probability, the actual value of $M(1)$ for a given sample member depends only on chance. 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 or sequential ignorability, 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. We denote these probabilities as:

> $\Pr[M(1) = 1|X]$ is the probability that a sample member will receive a favorable mediator value if assigned to the program, given his background characteristics (X).
>
> $\Pr[M(0) = 1|X]$ is the probability that a sample member will receive a favorable mediator value if assigned to the control group, given his background characteristics (X).

Under the program theory, we expect $\Pr[M(1) = 1|X] > \Pr[M(0) = 1|X]$. That is, being assigned to the program should *increase* the probability of receiving a favorable mediator value. However, assignment to the program might reduce this probability for some types of sample members.

If assigned to the program ($T = 1$), a sample member will, in effect, randomly receive a mediator value $M(1)$. Thus her outcome will be $Y[1,M(1)]$. If assigned to the control group, the sample member will randomly receive a mediator value $M(0)$. Thus her outcome will be

$Y[0,M(0)]$. We can now define the "total effect" of random assignment to the program on the outcome for a specific sample member as

$$\text{Total Effect: } B = Y[1,M(1)] - Y[0,M(0)]. \tag{31}$$

To define mediation, proponents of this approach ask how the sample member would have fared if she were assigned to the program but instead of receiving the program-assignment mediator value, $M(1)$, she received the control-group mediator value, $M(0)$. Her *direct* causal effect of assignment to the program (the part of the program assignment effect that is not produced through the mediator) is thus

$$\text{Direct Effect} = Y[1,M(0)] - Y[0,M(0)]. \tag{32}$$

Note that here the mediator value is held constant at $M(0)$, and we ask how the outcome changes only as a function of assignment to the program or control group. We can reason that if the total effect is large and positive and the direct effect is much smaller, the change in the mediator induced by program assignment is contributing substantially to the total impact. In contrast, if the direct effect is equal to the total effect, we can conclude that the hypothesized mediator played no role in transmitting the program effect.

This reasoning enables one to decompose the total program effect on sample members' outcomes into an indirect effect and a direct effect by subtracting *and* adding the quantity $Y[1,M(0)]$ from the total effect, such that

$$\begin{aligned} \text{Total Effect: } B &= Y[1,M(1)] - Y[0,M(0)] \\ &= Y[1,M(1)] - Y[1,M(0)] \text{ (Indirect effect)} \\ &+ Y[1,M(0)] - Y[0,M(0)] \text{ (Direct effect).} \end{aligned} \tag{33}$$

From Equation 33 we can see that (a) the indirect effect is the causal effect on the outcome of changing the mediator value without changing program assignment, and (b) the direct effect 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.

The term that is added and subtracted in Equation 33, $Y[1,M(0)]$, represents a counterfactual quantity that cannot be observed, because we cannot know what the mediator value for someone assigned to a program group would be if instead he were assigned to the control group. The key assumption used to identify this quantity (sequential ignorability) is that once sample members are stratified based on their observable baseline characteristics (X) and their program or control group membership (T), the probability of a favorable mediator value is the

same for all program group members in a stratum and the same for all control group members in the stratum. These probabilities can be estimated from sample data.

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 and Rubin, 1983), because stratifying sample members on a propensity score can balance them on all variables used to predict the propensity score, at least in large samples. This method proceeds as follows.

First, use the control group's data on X to predict the probability of receiving a favorable mediator value under the control condition, $\Pr[M(0) = 1|T = 0,X]$. Logistic regression analysis can be used for this purpose, producing an estimated coefficient for each predictor. These coefficients can then be applied to the background characteristics of program group members to estimate their probability of a favorable mediator value if assigned to control status, $\Pr[M(0) = 1|T = 1,X]$, which is a crucial counterfactual quantity. This procedure works in studies that randomly assign people to T because, in those studies, potential mediator values under program or control group assignment are unrelated to treatment assignment:

$$\Pr[M(0) = 1|T = 0,X] = \Pr[M(0) = 1|T = 1,X] = \Pr[M(0) = 1|X]. \quad (34)$$

Similarly, we can use the program group's data to predict the probability of receiving a favorable mediator value if assigned to the program $\Pr[M(1) = 1|T = 1,X]$, yielding a second set of estimated coefficients. We can then apply those coefficient estimates to the background characteristics of control group members to estimate $\Pr[M(1) = 1|T = 0,X]$. Using these two sets of estimated coefficients, we can generate, for every sample member, the predicted probability of a favorable mediator value under assignment to treatment or control status.

But given this information, how do we carry out the mediation analysis? This is a topic of active research among statisticians, and space prohibits us from describing the varied approaches that have been proposed to date. Suffice it to say that there seem to be three major approaches. One approach uses multiple imputation (Imai, Keele, and Yamamoto, 2010) to generate predictions of how the program group in each stratum would have fared if it had the distribution of mediator values observed for control group members from that stratum. A second approach develops a series of regression models to estimate direct and indirect effects (VanderWeele, 2015), which in some ways is akin to path analysis. A third, nonparametric approach uses weighting (Hong, forthcoming) to produce three groups: (a) the control group, (b) the original program group, and (c) a reweighted program group that represents how the program group would have fared had it received the distribution of mediator values exhibited by the control group for the same stratum.

Comparing Approaches to Mediational Analysis

The preceding approaches for studying mediation of program impacts — multisite instrumental variables, principal stratification, and the approximation of sequential randomization — have different strengths and weaknesses. The first approach exploits the availability of multisite RCTs to create a series of valid instruments and does not rely on pretreatment covariates to remove selection bias. However, in order to produce unbiased (or consistent) estimates of mediator effects, this approach requires that all relevant mediators be observed and accounted for. Thus it is potentially subject to "omitted mediator bias." 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, the sequential ignorability approach requires the identification and measurement of a rich set of pretreatment covariates to support the assumption that the observed covariates are sufficient to remove selection bias in estimates of the impact of the mediator on the outcome. The principal stratification approach does not require measuring and modeling all pretreatment confounders (as does sequential randomization), or the exclusion restriction (as does instrumental variables). Instead it requires covariates and follow-up outcomes that adequately predict the potential values of sample members' mediators. Moreover, principal stratification is more useful for identifying a direct effect of program assignment and thereby *falsifying* a mediation theory than it is for estimating parameters of a mediation process.

None of these approaches is perfect for all mediational analyses, and all mediational analyses (short of randomizing specified mediators) require strong assumptions in order to estimate mediator effects. Still, 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. Knowing how to choose a method of mediational analysis under the varied conditions that exist in the practice of evaluation research, however, is craft knowledge not yet fully or widely available.

Section 5

Final Remarks

Variation in program impacts upends conventional ways of analyzing and interpreting data from program evaluations. Among other things, it makes it possible to define (and estimate) different types of average impacts. For example, we can define a mean impact for the *typical site* from a population of sites or a mean impact for the *typical person* from a population of persons. These two parameters can differ and each can be valid for different purposes.

However, any average becomes less informative as impact variation increases. Furthermore, understanding this variation 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 are the maximum and minimum site-specific program impacts? and (c) What is the cross-site correlation between program impacts and control group mean outcomes? We term the search for answers to these questions learning "about" impact variation.

Variation in program impacts also provides opportunities for testing theories about why interventions do or do not work, for whom they work when they work, and under what conditions they work. Toward this end, we can pose theories to guide future data collection for explaining impact variation within and across program sites. We term such theory building learning "from" impact variation.

Statistical methods for discovering and explaining impact variation are developing rapidly, and we have tried to provide 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.

To promote this enterprise, the Spencer Foundation and the William T. Grant Foundation have joined together to sponsor a concerted three-year effort to develop new methodological approaches for studying impact variation and to undertake major empirical analyses of impact variation using existing data from multisite trials in education and youth development. We are privileged to be leading this unique effort. In addition, the Institute for Education Sciences of the U.S. Department of Education is funding work by one of the present authors (Bloom) and his colleagues to use existing RCT data sets to explore empirically the amount of impact variation that exists and to examine some important implications of this variation.

This collaborative project will bring together teams of researchers from the firms that conducted most of these RCTs (MDRC, Mathematica Policy Research, and Abt Associates Inc.) and methodologists from the University of Chicago, Stanford University, Harvard Univer-

sity, and the University of Pittsburgh. These teams will assess the utility of current methods for answering questions about impact variation (including moderation and mediation), develop and test new methods where necessary, and produce usable tools to help others conduct these types of analyses. In addition, the team will consider sample size requirements and develop new methods for assessing statistical power to help aid the design of future multisite trials. The core goals of the project are to build a capacity for studying impact variation and to bring to light important new substantive findings from existing RCTs. It is hoped that doing so will help to build a new research agenda for improving child, youth, and adult outcomes by better understanding which interventions work best, for whom, and under what conditions.

Appendix A

Characterizing the Impact of an Intervention on the Mean and Variance of Outcomes and Program Impacts for a Population of Individuals

It is often of interest to describe how an intervention affects the mean and variance of the outcome in a population of participants. The table below provides expressions for these parameters when the population consists of participants nested within each of J^* sites. Note that the between-site means and covariance parameters are weighted averages where the weight accorded each site is its population size. Here the potential outcomes for person *i* in site *j* are $Y_{ij}(0)$ if assigned to the control group and $Y_{ij}(1)$ if assigned to the program group. The person-specific causal effect is $B_{ij} = Y_{ij}(1) - Y_{ij}(0)$. All of the items in the table are estimable from data produced by a multisite randomized trial.

	Control Group	Program Group
Mean	μ_0	$\mu_0 + \beta$
Variance	$\sigma_0^2 + \tau_0^2$	$\sigma_1^2 + \tau_0^2 + \tau_B^2 + 2\tau_{0B}$

where

$$\mu_0 = \sum_{j=1}^{J^*}\sum_{i=1}^{N_j} Y_{ij}(0) / \sum_{j=1}^{J^*} N_j = \sum_{j=1}^{J^*} N_j U_{0j} / \sum_{j=1}^{J^*} N_j$$

$$\beta = \sum_{j=1}^{J^*}\sum_{i=1}^{N_j} B_{ij} / \sum_{j=1}^{J^*} N_j = \sum_{j=1}^{J^*} N_j B_j / \sum_{j=1}^{J^*} N_j$$

$$\sigma_t^2 = \sum_{j=1}^{J^*}\sum_{i=1}^{N_j} (Y_{ij}(t) - U_{tj})^2 / \sum_{j=1}^{J^*} N_j, \ t \in \{0,1\}$$

$$\tau_0^2 = \sum_{j=1}^{J^*} N_j \left(U_{0j} - \mu_0\right)^2 / \sum_{j=1}^{J^*} N_j$$

$$\tau_\beta^2 = \sum_{j=1}^{J^*} N_j (B_j - \beta)^2 \sum_{j=1}^{J^*} N_j$$

$$\tau_{\beta 0} = \sum_{j=1}^{J^*} N_j (B_j - \beta)(U_{0j} - \mu_0) / \sum_{j=1}^{J^*} N_j.$$

Appendix B

Derivation of the HLM Estimator of the Cross-Site Mean and Variance of Program Impacts

A simple hierarchical linear model (HLM) for the cross-site mean and variance of program impacts begins with the idea that the site sample difference of mean outcomes $\hat{B}_j = \overline{Y}_{1j} - \overline{Y}_{0j}$ is an unbiased, (approximately) normally distributed estimator of the true site-specific impact B_j with variance $V_j = n_j^{-1}[\sigma_1^2 / \overline{T}_j) + \sigma_0^2 /(1-\overline{T}_j)]$. So the first-stage model says:

$$\hat{B}_j \sim N(B_j, V_j). \tag{B.1}$$

However, the true impacts B_j vary over sites according an exchangeable normal distribution:

$$B_j \sim N(\beta, \tau^2). \tag{B.2}$$

Hence the marginal distribution of the sample mean difference is

$$\hat{B}_j \sim N(\beta, \tau^2 + V_j). \tag{B.3}$$

One can readily derive the score vector and Fisher information matrix for β and τ^2 conditional on V_j as shown by Raudenbush (1994), producing the pair of estimating equations at iteration *m*+1:

$$\begin{aligned}
\beta^{(m+1)} &= \frac{\sum_{j=1}^{J} (\tau^{2(m)} + V_j)^{-1} \hat{B}_j}{\sum_{j=1}^{J} (\tau^{2(m)} + V_j)^{-1}} \\
\tau^{2(m+1)} &= \frac{\sum_{j=1}^{J} (\tau^{2(m)} + V_j)^{-2} [(\hat{B}_j - \beta^{(m)})^2 - V_j]}{\sum_{j=1}^{J} (\tau^{2(m)} + V_j)^{-2}}
\end{aligned} \tag{B.4}$$

References

Abadie, A., Chingos, M., and West, W. (2014). Endogenous stratification. NBER working paper.

Angrist, J. D., Imbens, G., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, *91*(434), 444-455.

Bloom, H. S., Hill, C. J., and 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., and Porter, K. (2014). Using multi-site evaluations to study variation in effects of program assignment. Manuscript under review.

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

Burdick-Will, J., Ludwig, J., Raudenbush, S. W., Sampson, R. J., Sanbonmatsu, L., and Sharkey, P. (2011). Converging evidence for neighborhood effects on children's test scores: An experimental, quasi-experimental, and observational comparison. In G. J. Duncan and R. J. Murnane (Ed.), *Whither opportunity* (pp. 255-276). New York: Russell Sage.

Dempster, A. P., Rubin, D. B., and Tsutakawa, R. K. (1981). Estimation in covariance components models. *Journal of the American Statistical Association*, *76*(374), 341-353.

Frangakis, C. E., and Rubin, D. B. (2002). Principal stratification in causal inference. *Biometrics*, *58*(1), 21-29.

Heckman, J., and 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*, 974-987.

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

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

Holland, P. W. (1988). Causal inference, path analysis, and recursive structural equations models. *Sociological Methodology*, *18*(1), 449-484.

Hong, G. (forthcoming). *Causality in a social world: Moderation, mediation and spillover*. New York: Wiley-Blackwell.

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

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

Kling, J. R., Liebman, J. B., and Katz, L. F. (2007). Experimental analysis of neighborhood effects. *Econometrica*, *75*(1), 83-119.

Lindley, D. V., and Smith, A. F. (1972). Bayes estimates for the linear model. *Journal of the Royal Statistical Society. Series B (Methodological)*, 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*(386), 393-398.

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

Neyman, J. S. (1990). On the applicability of probability theory to agricultural experiments: Essay on principles, section 9 (D. M. Dabrowska and T. P. Speed, Eds. and Trans.), *Statistical Science, 5*(4), 465-480. (Original work, in Polish, published 1923, in *Roczniki Nauk Rolniczych Tom X* (*Annals of Agricultural Sciences*), 1-51.

Page, L. C., Feller, A., Grindal, T., Miratrix, L., and Somers, M. A. (2014). Principal stratification: A tool for understanding variation in program effects across endogenous subgroups. Manuscript under review.

Pearl, J. (2001). Direct and indirect effects. In *Proceedings of the seventeenth conference on uncertainty in artificial intelligence* (pp. 411-420). San Francisco: Morgan Kaufmann Publishers Inc.

Raudenbush, S. W. (1994). Random effects models. In H. Cooper and L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 301-321). New York: Russell Sage Foundation.

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

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

Raudenbush, S. W., Reardon, S., and Nomi, T. (2012). Statistical analysis for multi-site trials using instrumental variables. *Journal of Research and Educational Effectiveness, 5*(3), 303-332.

Reardon, S. F., and Raudenbush, S. W. (2013). Under what assumptions do multi-site instrumental identify average causal effects? *Sociological Methods and Research, 42*(2), 143-163.

Reardon, S. F., Unlu, F., Zhu, P., and Bloom, H. S. (2014). Bias and bias correction in multisite instrumental variables analysis of heterogeneous mediator effects. *Journal of Educational and Behavioral Statistics*, *39*(1), 53-86.

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

Rosenbaum, P. R., and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41-55.

Roy, A. D. (1951). Some thoughts on the distribution of earnings. *Oxford economic papers*, *3*(2), 135-146.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*(5), 688.

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

Rubin, D. B. (1981). Estimation in parallel randomized experiments. *Journal of Educational and Behavioral Statistics*, *6*(4), 377-401.

Rubin, D. B. (1986). Comment: Which ifs have causal answers. *Journal of the American Statistical Association*, *81*(396), 961-962.

Rumberger, R. W. (1995). Dropping out of middle school: A multilevel analysis of students and schools. *American Educational Research Journal*, *32*(3), 583-625.

Spybrook, J. (2013). Detecting intervention effects across context: An examination of the precision of cluster randomized trials. *Journal of Experimental Education, 82*(3), 334-357. doi: 10.1080/00220973.2013.813364.

U.S. Department of Health and Human Services. (2010). Head Start impact study: Final report, 3-39.

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

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

Yau, L. H., and Little, R. J. (2001). Inference for the complier-average causal effect from longitudinal data subject to noncompliance and missing data, with application to a job training assessment for the unemployed. *Journal of the American Statistical Association*, *96*(456), 1232-1244.

About MDRC

MDRC is a nonprofit, nonpartisan social and education policy research organization dedicated to learning what works to improve the well-being of low-income people. Through its research and the active communication of its findings, MDRC seeks to enhance the effectiveness of social and education policies and programs.

Founded in 1974 and located in New York City and Oakland, California, MDRC is best known for mounting rigorous, large-scale, real-world tests of new and existing policies and programs. Its projects are a mix of demonstrations (field tests of promising new program approaches) and evaluations of ongoing government and community initiatives. MDRC's staff bring an unusual combination of research and organizational experience to their work, providing expertise on the latest in qualitative and quantitative methods and on program design, development, implementation, and management. MDRC seeks to learn not just whether a program is effective but also how and why the program's effects occur. In addition, it tries to place each project's findings in the broader context of related research — in order to build knowledge about what works across the social and education policy fields. MDRC's findings, lessons, and best practices are proactively shared with a broad audience in the policy and practitioner community as well as with the general public and the media.

Over the years, MDRC has brought its unique approach to an ever-growing range of policy areas and target populations. Once known primarily for evaluations of state welfare-to-work programs, today MDRC is also studying public school reforms, employment programs for ex-offenders and people with disabilities, and programs to help low-income students succeed in college. MDRC's projects are organized into five areas:

- Promoting Family Well-Being and Children's Development
- Improving Public Education
- Raising Academic Achievement and Persistence in College
- Supporting Low-Wage Workers and Communities
- Overcoming Barriers to Employment

Working in almost every state, all of the nation's largest cities, and Canada and the United Kingdom, MDRC conducts its projects in partnership with national, state, and local governments, public school systems, community organizations, and numerous private philanthropies.