ANNUAL REVIEWS

Annual Review of Statistics and Its Application

Randomized Experiments in Education, with Implications for Multilevel Causal Inference

Stephen W. Raudenbush[1] and Daniel Schwartz[2]

[1]Department of Sociology, Harris School of Public Policy, and Committee on Education, University of Chicago, Chicago, Illinois 60637, USA; email: sraudenb@uchicago.edu

[2]Department of Public Health Sciences, University of Chicago, Chicago, Illinois 60637, USA

Annu. Rev. Stat. Appl. 2020. 7:177–208

The *Annual Review of Statistics and Its Application* is online at statistics.annualreviews.org

https://doi.org/10.1146/annurev-statistics-031219-041205

Keywords

multilevel data, causal inference, experimental design, heterogeneous treatment effects, hierarchical linear models, educational statistics

Abstract

Education research has experienced a methodological renaissance over the past two decades, with a new focus on large-scale randomized experiments. This wave of experiments has made education research an even more exciting area for statisticians, unearthing many lessons and challenges in experimental design, causal inference, and statistics more broadly. Importantly, educational research and practice almost always occur in a multilevel setting, which makes the statistics relevant to other fields with this structure, including social policy, health services research, and clinical trials in medicine. In this article we first briefly review the history that led to this new era in education research and describe the design features that dominate the modern large-scale educational experiments. We then highlight some of the key statistical challenges in this area, including endogeneity of design, heterogeneity of treatment effects, noncompliance with treatment assignment, mediation, generalizability, and spillover. Though a secondary focus, we also touch on promising trial designs that answer more nuanced questions, such as the SMART design for studying dynamic treatment regimes and factorial designs for optimizing the components of an existing treatment.

1. INTRODUCTION

Educational practice has two key features that shape research and experiments. First, formal schooling typically occurs in a multilevel setting, reflecting the hierarchical social structure of the schooling system. Second, heterogeneity abounds among both individuals and organizations. Children obviously vary, and on top of this they are grouped into classes mostly by age rather than by knowledge. Naturally, teachers and administrators also vary both in basic skill and in how faithfully they implement hypothetically standardized interventions.

These features generate many statistical challenges and, at the same time, raise interesting substantive questions that might otherwise be overlooked. In the following, we introduce the statistical challenges under review. These topics may also interest statisticians working in one of the many fields that share a multilevel setting and/or substantial heterogeneity, such as criminology, social welfare, job training, and medicine. Similarly, these issues also arise in observational studies, which, of course, must additionally contend with selection bias.

1.1. Endogeneity of Study Design

In multisite field experiments, the design is often not entirely controlled by the investigator. Site sizes and proportions treated may be correlated with the average treatment effect (ATE) in each site (for instance, if smaller schools are better at implementing the treatment). Though typically not considered, this can create a difficult bias-variance tradeoff for some important targets of inference.

1.2. Heterogeneity of Treatment Effects

Researchers suspect that educational interventions will affect children differently for a host of reasons, but many popular methods in this area assume constant treatment effects. Heterogeneity of impacts has broad implications for defining estimands, properties of estimators, and optimal trial design. By generating an ensemble of unbiased treatment effect estimates across sites, multisite trials offer unique opportunities to study heterogeneity, with implications for generalizability.

1.3. Noncompliance with Treatment Assignment

Some schools assigned to a new program may not implement the program, and even in schools that do implement the program, some teachers may decline to participate. Noncompliance makes the overall average effect of treatment participation challenging to identify, so many analysts instead estimate the average effect of treatment participation in a latent subpopulation of compliers. In multisite trials, compliance will vary from person to person and likely (on average) from site to site, giving rise to extra heterogeneity in the effects of treatment assignment and complicating interpretation.

1.4. Mediation

Mediators are proximal outcomes of an intervention that, in turn, shape longer-term outcomes. Experimenters often study such mediators to reveal mechanisms through which a treatment produces effects. Multisite trials generate new opportunities to study such causal mechanisms: The impact of treatment on mediators may vary across sites, and how this variation affects outcomes carries some information about mediators. However, the causal process itself may vary from site to site.

1.5. Generalizability

The ultimate scientific goal underpinning most trials is to accurately generalize results in some way, often to a broader population that the observed sample may not directly reflect. By linking data from educational experiments to population survey or census data, statisticians have begun to tackle this challenge. Researchers in this field have concluded that many trials could be better designed in support useful generalizations.

1.6. Spillover

Experimenters typically assume no interference between units, meaning that the treatment assignment of one unit does not affect the potential outcomes of any other unit. In education, where interventions occur in social milieu such as classrooms and schools, this assumption may be untenable. Some new work considers how to design experiments to uncover spillover effects and how to detect spillover even when it is not of primary substantive interest.

1.7. Novel Designs

A dynamic treatment regime is a multistage treatment that uses up-to-date information about individuals to personalize treatment at each stage, and it can be studied by the SMART (sequential multiple assignment randomized trial) experimental design; in education, this design is highly relevant to instruction in many forms. A second design of growing interest is the classical factorial design, which offers opportunities to study how particular components of a new intervention contribute to treatment effectiveness.

Before discussing each of these statistical challenges, we briefly review the history of how education research came to its modern era of large-scale randomized experiments and describe the most common experimental designs. We also introduce and motivate a theoretical model for the basic multisite trial and describe key estimands.

2. HISTORY AND MAJOR DESIGNS

A huge industry of publishers, nonprofit organizations, and universities sells text books, professional training programs, tests, and technological innovations to schools. Within this vast market, reformers have advocated the adoption of new curricula, new modes of teacher training, increased accountability, reductions in class size, school-based management, new assessments of student skill, and the adoption of school-wide programs for teaching reading and mathematics. Cook (2002) concluded that, prior to 2002, few of these reform efforts had been rigorously evaluated. He described a culture among educational program evaluators that favored surveys and small-scale, in-depth qualitative case studies as opposed to randomized experiments. In this culture, how a reform operated and how practitioners and students perceived its influence were more important than estimating the average impact on students or cost-effectiveness of the reform. Many evaluators, including most faculty in education, regarded randomized trials as infeasible or unethical, and some believed that random assignment would create artificial conditions unrepresentative of the daily practice of schooling.

2.1. A Turn Toward Experimentation

In 1999 the American Academy of Arts and Sciences sponsored a conference on the state of research in education. Chairing the meeting were Frederick Mosteller and Howard Hiatt, two men

who had helped lead the movement in the 1950s to establish the randomized trial as a foundation for causal inference in medicine, as well as Robert Boruch, a longtime advocate of social experimentation. They asked why there were there so few randomized trials in education and whether it was time to launch a new epoch of educational research that would parallel the history of medical research. The conference led to an important volume advocating more randomized trials in education (Mosteller & Boruch 2002).

One important stimulant for this initiative was the Tennessee class size experiment (Finn & Achilles 1990), which was motivated by a stalemate in the Tennessee legislature. The lawmakers could not agree on whether or not to outlaw large classes, but they did agree to study that question. Helen Pate Bain, then an associate professor at Tennessee State University and well-known advocate for education reform, argued for a randomized trial (Boyd-Zaharias 1999). Past studies of class size had mixed results, and some studies seemed to suggest that larger classes were actually more effective than small classes, almost surely because more effective teachers tend to attract more students. So Tennessee funded a study in which kindergarten students and teachers were both randomly assigned to classes large and small. Finn & Achilles (1990, p. 557) reported that "the results are definitive": Reducing class size could significantly increase student learning in reading and mathematics. The findings appeared uniformly positive across 79 diverse schools, 325 teachers, and 5,786 students. Mosteller (1995) celebrated this finding for a broad audience at the 1999 conference and asserted that this was among the best studies in the history of education. As an interesting by-product, Krueger & Whitmore (2001) found that low-income and minority students benefitted most from class size reduction. They also found that those randomly assigned to smaller kindergarten classes were, on average, more likely to attend college.

In 2001 Congress passed the No Child Left Behind (NCLB) law. It is well known that NCLB unleashed a regime of school accountability based on high-stakes testing. Less well known is the fact that the law also mandated the formation of the Institute of Education Sciences (IES) with the purpose of creating a new scientific basis for educational research. In 2002 Russell Whitehurst became the founding director of IES. With Whitehurst's commitment to experimental evaluation, and supported by a large increase in the budget for educational research, the IES funded more than 175 large-scale randomized controlled trials (RCTs) during the first decade of its existence (Spybrook 2014, Spybrook et al. 2016). Other government agencies and foundations have also lent support to movement toward randomized trials, and subsequent leaders of IES have continued to emphasize the importance of random assignment in program evaluation.

2.2. Research Designs

Cook (2002) discusses how many education researchers, influenced by limited exposure to experimental design and a funding landscape that did not encourage them to prioritize RCTs, were resistant to using randomized trials in part because for many types of interventions, students within the same classroom cannot not naturally (or, arguably, ethically) be randomized to different treatments. They could not see how randomized experiments could be appropriate in broad swaths of education. The field has come far in the past two decades, relying mainly on the following experimental designs to fit into the context of schooling.

2.2.1. Multisite randomized trials. Many important experiments in education involve randomization within sites; in the classical experimental design literature, this is a randomized block design, and we use the term multisite randomized trial to emphasize that the blocking sites are often of substantive interest. The National Head Start Impact Study, funded by the

Administration for Children, Youth and Families (Puma et al. 2010), is a prominent example. From a list of all Head Start centers in the United States, experimenters randomly selected more than 300. At each local center, low-income families applied for their children's admission to Head Start. Applications outnumbered available places, so offers of admission were based on a random lottery. Here we regard the centers as sites, and randomization is within sites. In our usage, a site is always a unit in which randomization occurs.

2.2.2. Cluster-randomized trials. As mentioned, for many educational interventions, a design that randomizes individual students to treatments with no restrictions is clearly unacceptable. For example, schoolwide instructional interventions apply to all children in a school (Borman et al. 2007, 2008). The sensible plan is to assign entire schools to treatments. However, the launch of the IES program of widespread experimentation almost foundered on the shoals of inadequate statistical power for such studies.

The problem was a widespread misunderstanding of the sample size requirements of the cluster-randomized trial. This history paralleled the early history of city-wide health promotion experiments (Donner et al. 1981, Fortmann et al. 1995, Murray 1995) in which entire cities were assigned at random to treatment or control. These studies included many hundreds of thousands of individuals at the cost of hundreds of millions of dollars, but the designers were apparently unaware that even when the between-city variance component is very small, randomization by city requires a fairly large number of cities in order to achieve adequate statistical power. Fortunately this experience led a small number of scientists to examine optimal sample sizes for cluster-randomized trials and to produce books and software that could enable future researchers to avoid the error (Klar & Donner 1997, Murray 1995, Raudenbush 1997). These ideas and tools took center stage at a 2004 conference sponsored by the William T. Grant Foundation on cluster-randomized trials in education attended by 50 leading funders and government officials. At this conference, attendees explored the design of hypothetical trials by applying user-friendly software. Many educational evaluators were shocked to discover that the statistical power of a cluster-randomized trial depends strongly on the number of clusters. But recruiting schools and teachers and sustaining their involvement is expensive; if every field trial required a large number of schools, the entire project was in danger.

To cope with this threat, educational experimenters generated two strategies (Bloom et al. 2007, Raudenbush et al. 2007). The first was to match or block clusters on demographic variables, geographic location, prior educational outcomes, or other factors believed related to the outcome, and then, within blocks, to randomize clusters to treatment. The second was to identify and control for covariates measured at the level of the cluster that were, by hypothesis, strongly predictive of the outcome. Both strategies showed promise for increasing power, but the second strategy proved remarkably effective when the outcome of interest was a measure of academic achievement such as a reading or math test, a common feature of educational evaluations. The reason is that school mean test scores collected before and after intervention may be correlated as high as $r = 0.90$ (Bloom et al. 2007, Hedges & Hedberg 2013); in this case using the covariate boosts power by roughly the same amount as doubling the number of clusters. Experience designing RCTs in education thus led evaluators to increasingly rely on some combination of prerandomization blocking and/or covariance adjustment to address the challenge of statistical power for the cluster-randomized trial.

2.2.3. Multisite cluster-randomized trials. Spybrook (2014) reported that in the first 175 randomized trials funded by IES, the single most common design was a multisite cluster-randomized

trial. This is a three-level design in which sites (e.g., schools) are blocks within which clusters (e.g., classrooms) are assigned at random; outcomes vary randomly among students who are nested within clusters.

2.2.4. Three-level person randomized trials. An increasingly common design randomly assigns each student to treatments within a block that is itself nested within a site, another three-level design. A prominent example is the school lottery study (Angrist et al. 2016, Clark et al. 2015, Hassrick et al. 2017). In such a study, parents apply for their children to be admitted to a new school. Applications exceed the number of available places, so a randomized lottery decides who will be offered admission. Evaluators follow lottery winners and losers to gauge the impact of random assignment. A separate lottery is held each year for each grade within each school. Thus, over several years, each school produces a collection of lotteries. We regard the lotteries as blocks nested within the school conceived as a site. Hence, each school generates a collection of average causal effects, one for each lottery, and an average over these lottery effects within each school constitutes the average effect of random assignment to the school. How to summarize effects across schools can be a tricky problem.

3. THEORETICAL MODEL

The generic goals of a randomized experiment in education (as in most fields) are to estimate some kind of ATE and to somehow characterize the heterogeneity of treatment effects. But before we can define these estimands with real clarity, we need to specify a theoretical model for the phenomenon under study. Below, we introduce the core modeling decisions in the basic setting of two-level multilevel educational experiments with a binary treatment and continuous outcome and then describe basic estimands. We use the term theoretical model to emphasize that the model thought to generate the data may not be the model used for estimation. Details and generalizations follow in later sections.

3.1. Potential Outcomes and Causal Effects

To begin, we assume a two-level structure in each student i is nested within site j, where a site can be a school, preschool center, or classroom. Three-level multisite trials can be regarded as specific cases of this design. We also assume intact sites (Hong & Raudenbush 2006), meaning that students inhabit one and only one site, and we assume that sites are statistically independent.

Define $T_{ij} = 1$ if student $i \in \{1, \ldots, n_j\}$ in site $j \in \{1, \ldots, J\}$ is assigned to a new treatment and $T_{ij} = 0$ if not. In principle, a student's potential outcome may depend on the teacher who implements the treatment. Moreover, a student's potential outcome may also depend on the treatment assignment of other students in the same site. We discuss this possibility in Section 9. For now, we adopt the stable unit treatment assignment assumption (SUTVA) (Rubin 1986), which holds that there is only one version of the treatment and that each student's potential outcomes are independent of the treatment assignment of other students. Thus, student i in site j possesses two potential outcomes $Y_{ij}(t)$ for $t \in \{0, 1\}$ for some interval scale or binary outcome variable Y, and one causal effect,

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

We never observe B_{ij} because if $T_{ij} = 1$, then $Y_{ij}(0)$ is missing, and if $T_{ij} = 0$, then $Y_{ij}(1)$ is missing. The observed outcome for child i in cluster j is then

$$\begin{aligned} Y_{ij} &= T_{ij}(1) + (1 - T_{ij})Y_{ij}(0) = Y_{ij}(0) + T_{ij}B_{ij} \\ &= \mu_{0j} + \beta_j T_{ij} + \varepsilon_{ij}. \end{aligned} \quad 2.$$

Here $\beta_j = \mu_{1j} - \mu_{0j}$ is the ATE in site j; $\mu_{tj} = \mathrm{E}[Y_{ij}(t)|\mu_{tj}]$ is the average potential outcome under treatment t for students in site j; and $\varepsilon_{ij} = T_{ij}\varepsilon_{1ij} + (1 - T_{ij})\varepsilon_{0ij}$ is a random zero-mean disturbance where $\varepsilon_{tij} = Y_{ij}(t) - \mu_{tj}$, with potentially heteroscedastic variance σ^2_{ij}. Heckman et al. (2010) terms this the correlated random coefficient, emphasizing the potential correlation between the person-specific intercept, $Y_{ij}(0)$, and the treatment effect, B_{ij}. An early discussion of this idea within education appears in Bryk & Weisberg (1976).

Now looking across sites, we write $\mu_{0j} = \mu_0 + u_{0j}$ and $\beta_j = \beta + b_j$, where u_{0j}, b_j are zero-mean random effects having variances τ_{00} and τ_{bb} and covariance τ_{0b}. This generates the familiar linear mixed model equation,

$$Y_{ij} = \mu_0 + \beta T_{ij} + u_{0j} + b_j T_{ij} + \varepsilon_{ij}. \quad 3.$$

The random terms in this model need not be normally, or parametrically, distributed. We focus our discussion on two parameters, the mean β and variance τ_{bb} of the common distribution of the site-specific treatment effects β_j. This is far from the only interesting approach to describing variability across sites; for others, see the sidebar titled A Note on Randomness.

3.2. Estimands

One common aim is to study the distribution of treatment effects over a population of sites. If we regard each sampled site as equally representative of this population, we might define our key estimands as

$$\mathrm{E}_{\text{sites}}(\beta_j) \equiv \beta_{\text{sites}}, \qquad \mathrm{Var}_{\text{sites}}(b_j) = \mathrm{E}(b_j^2) \equiv \tau_{bb\ \text{sites}}, \quad 4.$$

the mean and variance of the site-specific ATE distribution defined over a population of sites.

A NOTE ON RANDOMNESS

One of the most fundamental issues we must confront is where randomness enters our model; in particular, should site-specific causal effects be modeled as random or fixed? The random sampling route is natural if we want to generalize results beyond the observed sample to a large population. Most often, the sample for the trial is a volunteer sample or a convenience sample, yet the experimenter regards the sample as generated from an infinitely large, if not clearly defined, superpopulation. We reason that this is what interests most designers of education trials, which is why our model above adopts this point of view. Alternatively, one might view site effects as random in a Bayesian sense, regarding site-specific effects as exchangeable to reflect our subjective uncertainty. The notion that the site effects are fixed is consistent with viewing the sample as a finite population. This is sensible if we confine our interest to the observed sample. This approach is popular in some of the work done by contract research organizations (Schochet 2015), but due to space limitations, we do not treat it thoroughly. For more discussion on these opposing points of view, see Section 8, where we also discuss recent work linking convenience samples to larger, well-defined populations.

In contrast, suppose that the aim is to generalize to a population of students and that sites vary in how many students they serve. Then we might define parameters differently. For example, the person average treatment effect might be defined

$$\mathrm{E}_{\text{persons}}(B_{ij}) = \beta_{\text{persons}} = \mathrm{E}_{\text{sites}}(\omega_j \beta_j), \quad 5.$$

where $\mathrm{E}_{\text{persons}}$ denotes expectation over the distribution of person-specific random variables in a population of people. Here ω_j is a weight, scaled to have a mean of one, that is proportional to the size of the student subpopulation served by site j. In the population of persons, the variance components take on different meaning. The variance of person-specific causal effects is

$$\mathrm{Var}(B_{ij}) = \mathrm{E}_{\text{persons}}(B_{ij} - \beta_j)^2 + \mathrm{E}_{\text{sites}}[\omega_j(\beta_j - \beta_{\text{persons}})^2]. \quad 6.$$

The within-site variance component $\mathrm{E}_{\text{persons}}(B_{ij} - \beta_j)^2$ cannot be identified without heroic assumptions because it depends on the within-site covariance between potential outcomes $Y_{ij}(0)$ and $Y_{ij}(1)$, which are never jointly observed. However, the between-site variance $\tau_{bb\ \text{sites}}$ is identified because the multisite trial contains information about all the site-specific control group and experimental means.

3.3. Choice of Estimands

The choice of estimands can significantly affect the optimal design of an experiment. For example, if the population of interest is composed of sites, a simple random sample of sites combined with a simple random sample of students within sites may be optimal, depending on costs. If the population of interest is students, it will make sense to sample sites with probability proportional to size, again depending on costs. However, based on our reading of recent RCTs in education, we argue that both populations will often be of great interest. Policy makers will typically want to know the average impact of an intervention over the target population of students. However, information about the distribution of treatment effects across sites will often be of great interest for identifying especially effective or ineffective sites (Rubin 1981) and for learning about variation in the effectiveness of educational organizations. Moreover, we show below that the analyst can learn a great deal about noncompliance and mediation by studying variation in treatment effects across a population of sites. Therefore, designs that allow us to explore different populations may be of great interest, if feasible. For example, in a three-level design, we might be interested in a population of students, and a population of teachers, and a population of schools.

4. ESTIMATION AND ENDOGENEITY OF DESIGN

4.1. Data

We define n_j as the sample size for site j, and the proportion of the sample assigned to treatment in site j is $\bar{T}_j$. Under SUTVA, intact schools, and random assignment within each site, the sample mean difference $\hat{\beta}_j = \bar{Y}_{1j} - \bar{Y}_{0j}$ is unbiased for the site-specific ATE $\beta_j = \mathrm{E}(B_{ij}|\text{Site} = j)$, and its sampling variance is $\mathrm{Var}(\hat{\beta}_j|\beta_j) = V_j$. We assume $V_j = c/[n_j\bar{T}_j(1 - \bar{T}_j)]$. Many analysts have assumed a constant within-site variance, $\sigma^2_{ij} = \sigma^2$ for all i and j, in which case $c = \sigma^2$. Because treatment effects plausibly vary across students, Bloom et al. (2017) recommend specification of separate variances σ^2_1 and σ^2_0 for treatment and control students, respectively. In this case $c = \sigma^2 + (\sigma^2_1 - \sigma^2_0)(1 - 2\bar{T})$, where $\sigma^2 = \bar{T}\sigma^2_1 + (1 - \bar{T})\sigma^2_0$ with $\bar{T} = \sum_{j=1}^{J}\sum_{i=1}^{n_j} T_{ij}/N$. Under either choice, the sampling precision of $\hat{\beta}_j$ as an estimator of β_j is

$$P_j = V_j^{-1} \propto n_j\bar{T}_j(1 - \bar{T}_j). \quad 7.$$

To focus on key ideas, we assume V_j to be known, as c is estimated with great accuracy in educational RCTs. Thus, the basic data for our purpose here consist of $\hat{\beta}_j, V_j$ for $j \in \{1, \ldots, n_j\}$, $j \in \{1, \ldots, J\}$. Given the endogeneity of design, we assume V_j to be random.

4.2. Endogeneity of Design

As mentioned, when a new school or educational program opens, it is often the case that more people apply for admission than can be accommodated. If the number of applicants exceeds the number of available places, it is common practice, and in some cases legally required, to hold a random lottery to decide who should be offered admission. These lotteries have generated many opportunities for researchers to experimentally test novel approaches to school organization and practice (Angrist et al. 2016, Clark et al. 2015, Dobbie & Fryer 2013), in large part because they support both fairness to applicants and the priorities of statistical inference. Moreover, entire school districts have recently adopted new admissions rules that enable parents to list preferred schools for their children (Bloom & Unterman 2014), holding open the possibility of experimentally testing the impact of many regular public schools.

A concern, however, is that the number of persons who apply, equivalent to the sample size, n_j, may reflect the popularity—and thus, indirectly, the effectiveness—of the new program. Weighting site-specific data by site-specific precision P_j may then bias estimates of the site ATE β_{sites} by up-weighting the most effective schools. The number who apply may alternatively reflect the availability of good local alternatives, in which case, a large applicant pool might indicate a comparatively disadvantaged local population. In either case, the combination of the number who apply and the number of available seats determines the fraction, $\bar{T}_j$, that are offered admission. Therefore, $\bar{T}_j$ may also be endogenous. Hence, the sampling precision $P_j \propto n_j \bar{T}_j(1 - \bar{T}_j)$ may, in many cases, be regarded as an endogenous variable, that is, as a nonignorable random variable rather than a fixed aspect of the design, as is conventional in experimental research.

Even in studies that do not use lotteries, it may be the case that sites with varied size, more or fewer resources, or varied preferences may vary with respect to sampling precision P_j, opening up the possibility that precisions covary with site effectiveness. This is a common reality in education research and other fields (e.g., multihospital clinical trials), so we keep the effects of endogenous designs in mind as we discuss basic estimators for the estimands of a multisite trial.

4.3. Estimating the Site Average Treatment Effect

Under endogeneity of design, the familiar ATE estimators have different properties than usual, leading to a nontrivial bias-variance tradeoff (Raudenbush & Schwartz 2019). We highlight the basic results for the site ATE below.

4.3.1. Unweighted estimator. If each site equally represents the population of sites, Schochet (2015) recommends an unweighted estimator

$$\hat{\beta}_{\text{uw}} = J^{-1} \sum_{j=1}^{J} \hat{\beta}_j. \qquad 8.$$

This is clearly consistent for β_{sites}, though it treats all sites as equally informative, so it will typically have larger sampling variance than weighted estimators when precision, P_j, varies significantly

from site to site. Under our theoretical model,

$$\mathrm{Var}_{\mathrm{sites}}(\hat{\beta}_{\mathrm{uw}}) = \tau_{bb\ \mathrm{sites}} + E_{\mathrm{sites}}(V_j). \quad 9.$$

4.3.2. Ordinary least squares with site fixed effects. Probably the single most commonly used estimator regresses the outcome Y on treatment T with site fixed effects, yielding

$$\hat{\beta}_{\mathrm{FE}} = \frac{\sum_{j=1}^{J}(T_{ij} - \bar{T}_j)Y_{ij}}{\sum_{j=1}^{J}(T_{ij} - \bar{T}_j)^2} = J^{-1}\sum_{j=1}^{J}\frac{P_j}{\bar{P}}\hat{\beta}_j, \quad 10.$$

where $\bar{P} = \sum P_j/J$. We see that $\hat{\beta}_{\mathrm{FE}}$ weights each site's estimate $\hat{\beta}_j$ proportional to its sampling precision, P_j. The fixed effects model is our general model (Equation 3) with $\mu_0 + u_{0j}$ set to a fixed constant from site to site and b_j set to zero so that $\tau_{bb\ \mathrm{sites}} = 0$. If this model is correct and the within-site variances are homogeneous, the familiar Gauss-Markov theory guarantees that $\hat{\beta}_{\mathrm{FE}}$ is best linear unbiased with variance

$$\mathrm{Var}_{\mathrm{sites}}(\hat{\beta}_{\mathrm{uw}}) = E\Big(\sum_{j=1}^{J} P_j\Big)^{-1}. \quad 11.$$

Under these assumptions, $\hat{\beta}_{\mathrm{FE}}$ can be much more precise than $\hat{\beta}_{\mathrm{uw}}$, since then

$$\frac{\mathrm{Var}_{\mathrm{sites}}(\hat{\beta}_{\mathrm{uw}})}{\mathrm{Var}_{\mathrm{sites}}(\hat{\beta}_{\mathrm{FE}})} = \frac{E[\bar{V}_{\mathrm{Arithmetic}}]}{E[\bar{V}_{\mathrm{Harmonic}}]}, \quad 12.$$

where $\bar{V}_{\mathrm{Arithmetic}} = \sum V_j/J$ is the arithmetic mean of the sampling variances and $\bar{V}_{\mathrm{Harmonic}} = J/\sum P_j$ is the harmonic mean of the sampling variances (recall that the harmonic mean of positive numbers is always smaller than the arithmetic mean). However, if the treatment effects vary and are correlated with precision $\hat{\beta}_{\mathrm{FE}}$ is inconsistent, with finite-sample bias

$$\mathrm{E}_{\mathrm{sites}}(\hat{\beta}_{\mathrm{FE}}) - \beta_{\mathrm{sites}} = \mathrm{Cov}_{\mathrm{sites}}\Big(\frac{P_j}{\bar{P}}, \beta_j\Big). \quad 13.$$

4.3.3. Fixed intercepts, random coefficients. The fixed intercepts, random coefficients (FIRC) approach proposed by Bloom et al. (2017) is similar to fixed effects in setting $\mu_0 + u_{0j}$ to a fixed constant from site to site in order to minimize covariance assumptions. However, unlike with fixed effects, b_j is allowed to vary with $\tau_{bb\ \mathrm{sites}} \geq 0$, yielding

$$\hat{\beta}_{\mathrm{FIRC}} = J^{-1}\sum_{j=1}^{J}\frac{w_{j\,\mathrm{FIRC}}}{\bar{w}_{\mathrm{FIRC}}}\hat{\beta}_j, \quad 14.$$

where $w_{\mathrm{FIRC}\,j} = (V_j + \hat{\tau}_{bb\ \mathrm{sites}})^{-1} = P_j/(1 + P_j\hat{\tau}_{bb\ \mathrm{sites}})$ and $\bar{w}_{\mathrm{FIRC}} = \sum w_{\mathrm{FIRC}\,j}/J$. Note that the degree of precision-weighting depends on the (estimated) cross-site treatment effect variance: As $\hat{\tau}_{bb\ \mathrm{sites}} \to 0$, $\hat{\beta}_{\mathrm{FIRC}} \to \hat{\beta}_{\mathrm{FE}}$, and as $\hat{\tau}_{bb\ \mathrm{sites}} \to \infty$, $\hat{\beta}_{\mathrm{FIRC}} \to \hat{\beta}_{\mathrm{uw}}$ (Raudenbush & Bloom 2015). Thus, $\hat{\beta}_{\mathrm{FIRC}}$ lies on an interval between $\hat{\beta}_{\mathrm{FE}}$ and $\hat{\beta}_{\mathrm{uw}}$, tending toward $\hat{\beta}_{\mathrm{FE}}$ when treatment effects are compressed and toward $\hat{\beta}_{\mathrm{uw}}$ when they are dispersed. For known $\tau_{bb\ \mathrm{sites}}$, $\hat{\beta}_{\mathrm{FIRC}}$ is best linear unbiased when $b_j \perp P_j$ and nearly efficient when the within-site and between-site random effects are normally distributed with V_j correctly specified. Under these assumptions, $\hat{\beta}_{\mathrm{FIRC}}$ is potentially far more efficient than is $\hat{\beta}_{\mathrm{uw}}$. For large J,

$$\frac{\mathrm{Var}_{\mathrm{sites}}(\hat{\beta}_{\mathrm{uw}})}{\mathrm{Var}_{\mathrm{sites}}(\hat{\beta}_{\mathrm{FIRC}})} = \frac{\mathrm{E}[\bar{D}_{\mathrm{Arithmetic}}]}{\mathrm{E}[\bar{D}_{\mathrm{Harmonic}}]}, \quad 15.$$

where $\bar{D}_{\text{Arithmetic}}$ is the arithmetic mean of $D_j = V_j + \tau_{bb\ \text{sites}}$ and $\bar{D}_{\text{Harmonic}}$ is the harmonic mean.

However, if the assumption $b_j \perp P_j$ fails, $\hat{\beta}_{\text{FIRC}}$ is inconsistent with bias

$$\text{E}_{\text{sites}}(\hat{\beta}_{\text{FIRC}}) - \beta_{\text{sites}} = \text{Cov}_{\text{sites}}\left(\frac{w_{\text{FIRC}\,j}}{\bar{w}_{\text{FIRC}}}, \beta_j\right). \quad 16.$$

Raudenbush & Schwartz (2019) prove that the bias of $\hat{\beta}_{\text{FIRC}}$ is never larger that of $\hat{\beta}_{\text{FE}}$. Nevertheless, the inconsistency of $\hat{\beta}_{\text{FIRC}}$ under these conditions is troubling.

4.3.4. Open questions. Endogeneity of precision combined with treatment effect heterogeneity generates a bias-variance tradeoff that is worthy of more study. It stands to reason that a hybrid estimator will work better than those under consideration here, but such a hybrid has not appeared in the literature on educational field trials.

4.4. Estimating the Person Average Treatment Effects: Unweighted Persons Estimator

Recall that when the target is a population of students, the estimand is

$$\beta_{\text{persons}} = \text{E}_{\text{persons}}(B_{ij}) = \text{E}_{\text{sites}}(\omega_j \beta_j), \quad 17.$$

where ω_j is proportional to the size of site j. If sites are sampled with probability proportional to size, the desired weight (scaled to have a mean of 1.0) is $\omega_j = \frac{n_j}{\bar{n}}$, where $\bar{n} = \sum n_j/J$, yielding

$$\hat{\beta}_{\text{persons}} = J^{-1} \sum_{j=1}^{J} \frac{n_j}{\bar{n}} \hat{\beta}_j. \quad 18.$$

For small $\tau_{bb\ \text{persons}}$, $\hat{\beta}_{\text{persons}}$ is likely to be quite precise, though its variance will exceed that of $\hat{\beta}_{\text{FE}}$. If treatments effects vary and are correlated with precision, $\hat{\beta}_{\text{FE}}$ is inconsistent with bias

$$\text{Cov}_{\text{persons}}\left(\frac{w_{1j}}{\bar{w}_1}, B_{ij}\right) = \text{E}_{\text{sites}}\left[\omega_j\left(\frac{w_{1j}}{\bar{w}_1} - 1, \beta_j\right)\right], \quad 19.$$

where $w_{1j} = \bar{T}_j(1 - \bar{T}_j)$ and $\bar{w}_1 = \sum w_{1j}/J$. Under the same scenario, the bias of FIRC is a bit more complex and has not been studied.

4.5. Estimating the Variance Components

Raudenbush & Bloom (2015) show that an unbiased estimator of the treatment effect variance has the form

$$\hat{\tau}_{bb\ \text{pop}} = J^{-1} \sum \frac{w_{\text{pop}\,j}}{\bar{w}_{\text{pop}}}[(\hat{\beta}_j - \hat{\beta}_{\text{pop}})^2 - V_j], \quad 20.$$

where, for the designs mentioned above, $w_{\text{pop}\,j} = 1$ if pop = sites, and $w_{\text{pop}\,j} = \frac{n_j}{\bar{n}}$ if pop = persons, and $\hat{\beta}_{\text{pop}}$ is the unbiased ATE estimate for the population given by Equation 4 or Equation 5. If negative estimates are set to zero, the estimates are no longer unbiased but are J-consistent so long as $\tau_{bb\ \text{pop}} > 0$. Similar to the site ATE story, the consistent estimator for $\tau_{bb\ \text{sites}}$ may be quite inefficient if precisions vary substantially from site to site. The maximum likelihood estimator under the FIRC model substitutes $\frac{w^2_{\text{FIRC}\,j}}{\sum w^2_{\text{FIRC}\,j}/J}$ as the weight in Equation 18, and this estimate will

tend be more efficient than those given by Equation 20 if precisions are independent of treatment effects. The logic of estimation of the remaining variance components is similar. Relatively little research has examined variance components estimation when precisions are nonignorable. Given the widespread interest in linear mixed models, this is a topic of considerable interest.

4.6. An Empirical Example

Using the methods just described, Raudenbush & Schwartz (2019) reanalyze data from the National Head Start Impact Study, including 3,392 children randomly assigned by lottery within each of 316 sites. Site-specific sample sizes vary widely, with many small sites. Across five outcomes (reading, math, oral language, receptive vocabulary, and aggressive behavior), point estimates of the ATE and decisions based on a nominal significance level of $\alpha = 0.05$ were quite similar, with two exceptions: For math and aggressive behavior, standard errors using the unweighted estimator were more than half the size of the point estimates. The unweighted estimator β_{sites}, while consistent, appears woefully inefficient when site sizes are highly variable and often small. The performance of the unweighted estimator of $\tau_{bb\ \text{sites}}$ is particularly variable in this case. Raudenbush & Schwartz (2019) emphasize the need for more research on estimation of β_{sites} and $\tau_{bb\ \text{sites}}$ when site sizes are highly variable, which often arises in large-scale field trials that use lotteries to accomplish random assignment.

5. STUDYING HETEROGENEITY OF TREATMENT EFFECTS

Statisticians have primarily pursued two main questions describing and estimating heterogeneity of treatment effects: "How much?" and "Where?" In this section, we discuss both in turn. A third, "Why?", is treated in Section 7.

5.1. Quantifying Between-Site Variation in Treatment Effects

We saw in Section 4 that multisite trials enable the analyst to estimate the between-site component of the variance of the treatment effects. A recent survey of multisite trials in education and job training programs suggests that the between-site variance may tend to be small in studies where the intervention itself has a small ATE (Weiss et al. 2017). This type of heterogeneity is discussed at length by Raudenbush & Bloom (2015), who also present basic random effects estimators.

5.2. Visualizing Between-Site Variation in Treatment Effects

Suppose we display a histogram of sample mean differences, that is, $\hat{\beta}_j = \bar{Y}_{1j} - \bar{Y}_{0j}, j = 1, \ldots, J$. Unless all site sizes are large, this histogram will be too wide because of the noise of $\hat{\beta}_j$ as an estimate of β_j. Bayes or empirical Bayes estimates, which shrink unreliable estimates toward the mean, will be too narrow, but Louis (1984) shows how to recalculate these estimates to ensure that the dispersion in the histogram is consistent with the estimate of $\tau_{bb\ \text{sites}}$. Bloom et al. (2017) provide an example in education. We note that the shape of such a histogram will be influenced by parametric assumptions regarding the distribution of the unobservable values of β_j. Shrinkage estimators of site-specific effects can be used to rank sites by effectiveness and to identify especially effective (or harmful) sites for further study (Shen & Louis 1998, Paddock et al. 2006).

5.3. Quantifying Within-Site Variation in Treatment Effects

We noted earlier that the variance of treatment effects within sites is not identified because the data contain no information about the covariance between the two potential outcomes of any unit.

However, Ding et al. (2016) show in a single-site setting that at least there always exists a valid test of zero treatment effect variance in the finite sample, and thus in any population containing that sample. Their approach relies on a Fisher randomization test, using a clever trick (Berger & Boos 1994) to handle the unknown ATE, which in this setting is a nuisance parameter. Under a similar framework the same authors give sharp bounds for the treatment effect variance using a mathematical property of the empirical quantile function (Ding et al. 2019). Both of these methods can also be used to study the idiosyncratic variance that remains in treatment effects after conditioning on covariates that moderate the effects, as discussed in the next section.

5.4. Identifying Moderators of a Treatment Effect

Moderation occurs when the conditional average treatment effect (CATE) differs from the ATE. Moderation answers the question "for whom does the treatment work differently?" Generally, moderation refers to CATEs that condition on some observed pretreatment covariate like race or gender, though we might also expand our conception to include conditioning on latent characteristics as in principal stratification (Feller et al. 2016). We also see in Section 7 that statisticians have expanded the concept of mediation to apply to interaction effects between treatment assignment and a mediator. Potential moderators may be chosen by substantive theory (Angrist et al. 2013, Shadish et al. 2002) or by statistical methods for variable selection and high-dimensional data (Green & Kern 2012, Guo et al. 2017, Imai & Ratkovic 2013, Wager & Athey 2018). In either case, researchers should be wary of multiple testing since searches for moderation may devolve into fishing expeditions (Wang & Ware 2013).

5.5. Principal Stratification

Frangakis & Rubin (2002) propose a novel approach to studying the moderating effect of membership in a principal stratum, a latent class of persons. To illustrate, Feller et al. (2016) provide an application of this approach using the National Head Start Impact Study. They define three principal strata according to how students would respond to treatment assignment. Stratum 1 are those who would attend Head Start if assigned to Head Start but who would attend an alternative day care center if assigned to control. Stratum 2 includes those who would go to Head Start if so assigned, but who would otherwise stay home (with a parent, neighbor, or relative). Stratum 3 consists of those who would go to an alternative center regardless of treatment assignment, and Stratum 4 are those who would stay home regardless of treatment assignment. Other logically possible strata are assumed empty. For example, those who would not attend Head Start if assigned to Head Start but who would attend Head Start if not assigned to Head Start are called defiers and are assumed not to exist. Similarly, random treatment assignment is assumed not to affect the choice between attending an alternative center or staying home. A crucial feature of this methodology is that stratum membership is treated as a pretreatment covariate, though one we cannot observe for all units. We discuss a special case of this methodology in detail in Section 6.

5.6. Nonuniqueness of Moderator Models

Using principal stratification, Feller et al. (2016) find that the children who benefitted most are those who would attend Head Start if assigned but who would otherwise stay home. In contrast, those who would attend Head Start if assigned but who would otherwise attend an alternative preschool center benefitted little. Using the same data but a different methodology, Bitler et al. (2014) found that children with low skills as measured by pretreatment cognitive assessments

benefitted most from assignment to Head Start. By analyzing impacts across demographic subgroups, Bloom & Weiland (2015) find that low-income Hispanic children benefitted most. In fact, it is possible that all three findings are correct, if the low-skill children were low-income Hispanic children who would have stayed home if not assigned to Head Start. More generally, moderation models do not give unique explanations. See Section 7 for a review of strategies for testing explanatory theories.

6. NONCOMPLIANCE

So far, we have studied the impact of random assignment to treatment T. If all units actually receive their assigned treatment, the effect of assignment is simply the effect of the treatment. This is a benign state of affairs that statisticians have labeled full compliance with treatment assignment. Full compliance is not typically at play in large-scale educational field trials, and under partial compliance the effect of random assignment is called the intention to treat (ITT) effect. In the Tennessee study of class size reduction, some students assigned to large classes ended up in small classes (Krueger & Whitmore 2001). In the National Head Start Impact Study, about 25% of the children randomly offered a place in Head Start did not attend, and about 15% of those not offered a place actually did attend (Bloom & Weiland 2015). In lottery studies of new schools, a typical finding is that about 75% of lottery winners and 25% of lottery losers attend the school (Hassrick et al. 2017). This occurs because the lottery losers are placed on a waiting list and may be offered a place if lottery winners decline.

6.1. Noncompliance in a Single-Site Study

The method of instrumental variables (IV) has been widely used to study the impact of program participation in randomized experiments when compliance with randomization is imperfect: Treatment assignment is an instrument used to identify the impact of actually experiencing the program (Angrist et al. 1996, Heckman & Robb 1985). We first review how the IV method is now conventionally used in single-site studies. Next, we consider a much smaller literature that generalizes these results to the case of multisite trials. We see that new complications arise, requiring careful consideration of assumptions that underlie alternative analytic strategies.

6.1.1. Homogeneous treatment effects. **Figure 1*a*** displays a model of the association treatment assignment, $T \in \{0, 1\}$, program participation $M \in \{0, 1\}$, and Y. The effect of assignment on participation is regarded here as a constant, γ. The effect of participation on the outcome is regarded as the constant δ. Note that there is no direct path between T and Y (the direct effect θ is then set to 0). This is known as an exclusion restriction, reflecting the key assumption that a student's assignment to treatment can affect the outcome only if that student participates in the program.

We regress M on T to obtain an estimate of γ, and we regress Y on T to obtain an estimate of β. Both estimates, protected by randomization, are unbiased, so the ratio $\hat{\delta} = \hat{\beta}/\hat{\gamma}$, called the Wald estimator after Wald (1940), is consistent for δ so long as $\gamma \neq 0$. In practice, the null hypothesis $H_0 : \gamma = 0$ must be rejected with sufficient confidence to avoid what is called finite-sample bias (Bound et al. 1995). Analysts often use $F(1, df) > 10$, where df is the denominator degrees of freedom for the central F distribution, as a criterion that renders finite-sample bias small.

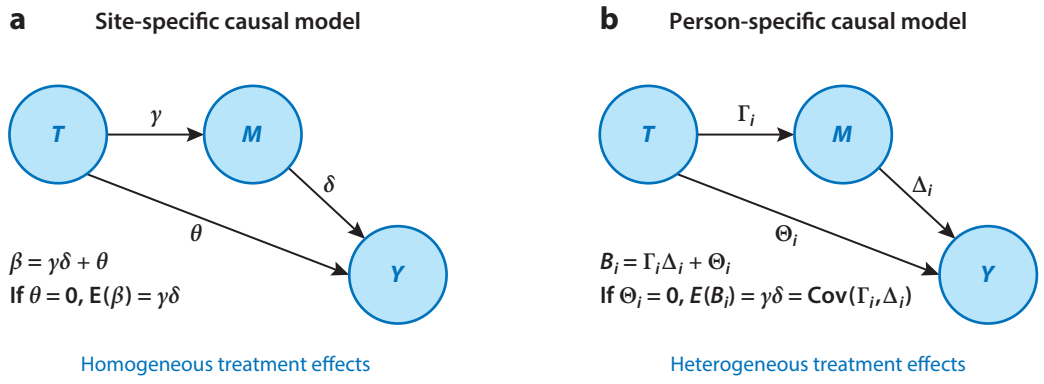

Figure 1

A path model with (*a*) homogeneous treatment effects and (*b*) heterogeneous treatment effects.

6.1.2. Heterogeneous treatment effects. While **Figure 1*a*** represents a causal model for a population under the assumption of homogeneous treatment effects, **Figure 1*b*** displays a person-specific causal model using potential outcomes (Raudenbush et al. 2012). Define the person-specific causal effect of assignment on participation as $\Gamma_i \equiv M_i(1) - M_i(0)$ and the person-specific effect of participation $\Delta_i \equiv Y_i(m = 1) - Y_i(m = 0)$. In principle, the potential outcome of assignment depends not only on assignment itself but also on whether assignment generates participation. Thus, we can write the potential outcomes $Y_i(t) = Y_i(t, M_i(t))$ (Angrist et al. 1996). However, under the exclusion restriction, once we know whether student i participates, knowing that student's treatment assignment has no bearing on that student's outcome. Thus, under the exclusion restriction, $Y_i(t) = Y_i(t, M_i(t)) = Y_i(M_i(t))$ and the ITT effect is

$$\begin{aligned} B_i &\equiv Y_i(M_i(1)) - Y_i(M_i(0)) \\ &= Y_i(0) + M_i(1)\Delta_i - [Y_i(0) + M_i(0)\Delta_i] \\ &= [M_i(1) - M_i(0)]\Delta_i \\ &= \Gamma_i\Delta_i. \end{aligned} \qquad 21.$$

The second step in Equation 21 follows from linearity, which is trivially met here because the predictor $M_i(t)$ is binary but can be contentious when $M_i(t)$ is continuous. We can define the average ITT effect through the equation

$$\begin{aligned} \mathrm{E}(B_i) \equiv \beta = \mathrm{E}(\Gamma_i\Delta_i) &= \mathrm{E}(\Gamma_i) * \mathrm{E}(\Delta_i) + \mathrm{Cov}(\Gamma_i, \Delta_i) \\ &\equiv \gamma\delta + \sigma_{\Gamma\Lambda}. \end{aligned} \qquad 22.$$

We see from Equation 22 that average effect of treatment assignment will be large when any of three terms is sufficiently large: the average compliance γ, the average benefit of treatment δ, or the covariance $\sigma_{\Gamma\Lambda}$. This covariance will be large when program staff are able to induce students who stand to benefit most from the program to comply with treatment assignment, or when these students are otherwise more likely to comply. But recall that the conventional IV estimand is (for $\gamma \neq 0$)

$$\beta/\gamma = \delta + \sigma_{\Gamma\Lambda}/\gamma. \qquad 23.$$

So, the conventional instrumental variable estimator will only be consistent if $\sigma_{\Gamma\Lambda} = 0$, the strong assumption of no covariance between compliance and effect, which in most education applications we would like to avoid.

6.1.3. Monotonicity. Angrist et al. (1996) propose replacing the assumption $\sigma_{\Gamma\Lambda} = 0$ with a weaker assumption known as monotonicity, namely, that $\Gamma_i \geq 0$ for all units i. This assumption requires that treatment assignment discourages no one from participating in the treatment. The price we pay for this weaker assumption is that our interpretation of δ is more constrained.

To motivate this concept, and following Frangakis & Rubin (2002), we define principal strata as subsets of students defined by their potential participation under random assignment. The compliers are those who would participate $[M_i(1) = 1]$ if offered the program and not participate $[M_i(0) = 0]$ if assigned to control. For compliers, therefore, the impact of being assigned to the program is $\Gamma_i = 1$. Noncompliers include never takers, those who would not participate under either treatment assignment $[M_i(1) = M_i(0) = 0]$, and always takers, those who would participate regardless of treatment assignment $[M_i(1) = M_i(0) = 1]$. Thus, for noncompliers, $\Gamma_i = 0$. A fourth, logically possible stratum would include defiers, who would take up the program if assigned to control but not if assigned to the program. For defiers, $\Gamma_i = -1$. The monotonicity assumption rules out the existence of this stratum. Decomposition of the ITT effect by stratum generates $\mathrm{E}(\Delta_i|\Gamma_i = 1) \equiv \delta_{\mathrm{CACE}}$, or simply CACE, the complier average causal effect:

$$\begin{aligned}\beta = \mathrm{E}(B_i) = \mathrm{E}(\Gamma_i\Delta_i) &= 1 * \mathrm{E}(\Delta_i|\Gamma_i = 1)\Pr(\Gamma_i = 1) + 0 * \mathrm{E}(\Delta_i|\Gamma_i = 0)\Pr(\Gamma_i = 0) \\ &= \mathrm{E}(\Delta_i|\Gamma_i = 1)\gamma \equiv \delta_{\mathrm{CACE}}\gamma.\end{aligned} \quad 24.$$

Hence, we can identify $\delta_{\mathrm{CACE}} = \beta/\gamma,\ \gamma > 0$, the average causal effect for the subpopulation whose participation is influenced by random assignment.

A problem for interpretation is that the magnitude of δ_{CACE} may depend on how effective the program is at inducing participation (Heckman & Vytlacil 2001). A program director who is very skilled at encouraging participation 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 participation, δ_{ATE}, is the same in the two studies. This ambiguity pervades applications of IV in multisite trials, where staff and participants vary across sites. A beautiful feature of the multisite trial is its capacity to evaluate heterogeneity in compliance and therefore to explore the seriousness of this potential ambiguity for interpretation.

6.1.4. Heckman correction. Heckman (1979) introduced an unbiased estimator of the average effect of program participation (not just among compliers) under alternative assumptions, which is now known as the Heckman correction. This approach uses a simultaneous equation model: One equation is a probit regression of program participation on assignment, and the other equation is a linear regression of the outcome on participation. The two equations have correlated errors. The key assumption is that the outcome errors and the latent propensity to participate (from the probit) are bivariate normal in distribution. This can be a strong and brittle assumption, leading to poor properties for apparently small violations. Zhelonkin et al. (2016) propose modifications to increase robustness (see also Kline & Walters 2019).

6.2. Extension to Multisite Trials

In a multisite trial, the process described above is replicated within each site. We have person-specific effects of treatment assignment, $B_{ij} = \Gamma_{ij}\Delta_{ij}$, and site average effects of treatment assignment, $\beta_j = \gamma_j\delta_j$ (under a within-site no covariance assumption), where γ_j is the fraction of persons who comply with treatment assignment in site j and δ_j is the CACE in site j. As before, we might identify CACE within each site through Wald estimators, but this requires a relatively large sample and high compliance in each site of the trial, and these conditions have not held in

educational RCTs to date. Raudenbush et al. (2012) consider methods for studying the mean and variance of δ_j using a site-level no-compliance-effect covariance assumption. In highly original work, Walters (2015) responds to this problem by extending the traditional Heckman correction with site-specific pairs of simultaneous equations, which have coefficients with random effects that induce shrinkage across sites. This method relies on potentially strong parametric assumptions.

6.2.1. Comment on interpretation. If we have full compliance in all sites, treatment effects will vary if one or both of two conditions hold: (*a*) sites vary in subpopulations and subpopulations respond variably to the same treatment, and (*b*) the treatment is implemented with variable effectiveness across sites. Under noncompliance, CACE can differ not only because of *a* and *b* but also because (*c*) some sites achieve higher compliance than others or (*d*) different sites encourage different subsets of the population to comply. Thus, the complier population is really an endogenous outcome of the interplay between site practices and heterogeneous subpopulations. If a site can achieve a large CACE by encouraging only the most promising persons to comply, that site will look better than average on CACE. Stated more generally, if high compliance predicts low CACE, the low-compliance sites will look better than average, particularly compared with δ_{persons}. If high compliance predicts high CACE, high-compliance sites will look better than average, particularly when compared with δ_{sites}.

6.3. Open Problems

From the standpoint of the multisite trial, it seems that in replacing the no-covariance assumption with the monotonicity assumption and thereby changing the definition of the impact of attendance from ATE to CACE, we have not really solved the problem of selection bias that noncompliance generates (Heckman & Vytlacil 2001). In essence, the multisite trial can reveal the challenges of summarizing evidence from a replicated experiment characterized by noncompliance. Unless compliance rates are uniformly high, some modeling based on additional assumptions appears essential to achieve clear scientific interpretation. More research is needed on alternative modeling strategies and required assumptions.

7. MEDIATING MECHANISMS

In 2002, the IES created the What Works Clearinghouse, an agency that sorts through claims of educational effectiveness and certifies particular interventions as effective (Confrey 2006). In recent years, however, IES has increasingly pressed for answers to harder questions, not about whether an intervention works, but rather about why. For example, suppose that a training program helps students learn only if it improves a teacher's measurable instructional practice (see Allen et al. 2011). Knowing this is crucial for those who are adopting an experimentally tested intervention at a new site. If measured teacher practice is not changing at the new site, the training is not working as expected there, and one needs to modify or discontinue the training. This sounds simple, but nailing down mediational mechanisms is hard.

7.1. Conventional Mediation in Single-Level Studies

Many thousands of studies have explored mediating mechanisms using path analysis, an approach originated by Wright (1921), extended by Duncan (1966), and codified for application by Baron & Kenny (1986) in one of the most widely cited articles in the history of psychology. Hong (2015,

chapter 10) provides a detailed review of this approach as well as modern criticism and alternative approaches.

A stylized representation of this approach is displayed in **Figure 1*a***. The impact of T on M is represented by the regression coefficient γ; the effect of M on Y is the regression coefficient δ. The total effect of T on Y is then the regression coefficient $\beta = \gamma\delta + \theta$, where $\gamma\delta$ is the indirect effect of T that operates through the mediator M, and θ is the direct effect that operates through unspecified mediators. If $\gamma\delta$ is large and θ is small, M is said to largely mediate the impact of T on Y.

7.2. Assumptions Underlying the Conventional Model

Holland (1988) was the first to apply the counterfactual account of causality (Haavelmo 1943, Holland 1986, Neyman 1935, Rubin 1978) to derive the assumptions required for this conventional method of mediation analysis. A useful extension is provided by Bullock et al. (2010). Hong (2015, chapter 10) reviews these critiques and evaluates a series of methodological innovations intended to relax the strong assumptions underlying this model. Assuming T is randomly assigned, the following assumptions must be met if the conventional model is to identify the causal pathway: (*a*) linearity of the association between M and Y within levels of T; (*b*) additivity of the impact of T and M on Y, meaning that the impact of the mediator cannot depend on treatment assignment; (*c*) ignorable assignment of M; (*d*) unobserved mediators lurking within θ are uncorrelated with M given T; and (*e*) no covariance between the person-specific impact of T on M and the impact of M on Y. To understand this last assumption, we find it useful to represent the mediation process through a person-specific model with heterogeneous effects (see **Figure 1*b***). We see that the person-specific indirect effect $\Gamma_i\Delta_i$ has expectation

$$\mathrm{E}(\Gamma_i\Delta_i) = \gamma\delta + \mathrm{Cov}(\Gamma_i, \Delta_i). \qquad 25.$$

The conventional model requires setting this covariance to 0. To see why this is a strong assumption, let us consider the following example (Nomi & Allensworth 2009): Assignment to intensive high-school math instruction (T) increases advanced mathematics course-taking later in high-school (M), which in turn increases college enrollment (Y). To assume $\mathrm{Cov}(\Gamma_i, \Delta_i) = 0$ is to assume that students who respond to treatment assignment by taking more advanced courses (that is, who have large values of Γ_i) are not especially likely to benefit in terms of college enrollment from taking advanced math courses (that is, to have large values of Δ_i). This seems implausible and motivates further modeling.

7.3. Attempts to Relax the Assumptions of the Conventional Model

Suppose now that we conceive of T (receiving intensive math instruction early) and M (taking advanced math courses later on) as two treatments, both binary. Rather than regarding a student's potential mediator values as a pretreatment covariate as in principal stratification, we view advanced course taking as a second treatment to which a student is effectively randomly assigned. By hypothesis, random assignment to T increases the probability of random assignment to M, which increases the outcome Y. Random assignment to M, however, does not occur in practice. Instead, the analyst regards subsets of students who have the same distribution of pretreatment covariates, X, as being, in effect, randomly assigned to M. The information in X, which may have high dimension, is summarized by the propensity score (Rosenbaum & Rubin 1983).

Recall that in principal stratification with binary T and binary M, potential mediator values were fixed a priori, so that each participant possessed two potential outcomes. In contrast, under

sequential random assignment, each participant possesses four potential outcomes. Define $M(1)$ and $M(0)$ as two random variables, each of which can take on two values (1 or 0). We now have the decomposition (Pearl 2001)

$$\begin{aligned}\mathrm{E}[Y(1,M(1))] - \mathrm{E}[Y(0,M(0))] = &\left\{\mathrm{E}[Y(1,M(1))] - \mathrm{E}[Y(1,M(0))]\right\} \\ &+ \left\{\mathrm{E}[Y(1,M(0))] - \mathrm{E}[Y(0,M(0))]\right\}. \end{aligned} \quad 26.$$

Here $\mathrm{E}[Y(1,M(1))] - \mathrm{E}[Y(1,M(0))]$ is the indirect effect of the mediator, holding T constant at 1, and $\mathrm{E}[Y(1,M(0))] - \mathrm{E}[Y(0,M(0))]$ is the direct effect of treatment T conditional on $M(0)$. Note that the decomposition is not unique, since we could have instead added and subtracted $\mathrm{E}[Y(0,M(1))]$. The curious feature of the decomposition in Equation 26 is the counterfactual quantity $\mathrm{E}[Y(1,M(0))]$. We can think of this as the mean outcome if the entire population were treated ($T = 1$) but the fraction of persons assigned to mediator ($\mathrm{M} = 1$) was $\Pr(M(0) = 1)$ rather than $\Pr(M(1) = 1)$. Three statistical approaches have emerged to model and estimate this decomposition. All rely on sequentially ignorable mediator assignment given covariates, and each allows statistical interaction between the treatment T and the mediator M.

7.3.1. Elaborated regression approaches. Petersen et al. (2006) and VanderWeele (2015) elaborate the conventional model to allow for interactions between M and Y and for nonlinear associations between M and Y. They also emphasize eliminating observable confounding by including pretreatment covariates, call them X, in their models. A series of regressions and a strategy for combining results across regressions are required to identify the causal effects of interest. We do not describe these methods in detail because our space is limited and these references are admirably clear. We can, however, conclude that this line of work essentially relaxes the assumptions of the conventional model by making the model more complex. Nonadditive and nonlinear structural forms can replace the linear and additive forms but must be explicitly specified. The assumption that the covariance in Equation 25 is null is weakened by virtue of conditioning on covariates X. The price to be paid using this approach is a series of functional form assumptions required to efficiently estimate an increased number of parameters.

7.3.2. Weighting-based approaches. Using Equation 26, Hong (2015) proposed a ratio of inverse probability of treatment weighting, a strategy that reweights the experimental group to have the same distribution as the control group on the mediator. This weighting effectively occurs within levels of the propensity score conditional on X. Like the elaborated regression approaches, weighting approaches assume sequential randomization given X. However, the elaborated regression approach is more ambitious because it seeks to estimate paths between T and M and M and Y while the weighting approach just estimates two quantities: the average indirect and average direct effects. The more ambitious elaborated regression approach requires more functional form assumptions. It is also presumably more efficient under those assumptions. In contrast, the weighting approach uses an essentially nonparametric model for the outcome.

7.3.3. Simulation-based methods. Imai (2010) proposes Monte Carlo methods for estimating the counter-factual quantity $\mathrm{E}[Y(1,M(0))]$. First, sample M given the covariates X and treatment assignment $T = 0$. Next, simulate $\mathrm{E}[Y(1,M(0))]$ from the model for Y given covariates X and $T = 1$ and $M(0)$. These simulations can be obtained under a variety of models, linear and nonlinear, for a variety of discrete and continuous predictors.

7.4. Multilevel Mediation Models

The multilevel setting can increase the complexity of the mediation model. For example, one or more of the coefficients in the extended regression coefficient approach can vary over sites, and direct and indirect effects may vary and covary across sites. However, the multilevel setting offers new opportunities to learn about mediation processes, and we consider three briefly.

7.4.1. Multilevel intervention as a process of sequential randomization. The 4Rs intervention is a school-wide program that aims to encourage instruction that is rigorous, is relevant to student interests, and reduces aggressive behavior (VanderWeele et al. 2013). Schools were matched on demographics and then, within pairs, randomly assigned to treatment or control conditions. Within the experimental condition teachers received training needed to implement the program. The fidelity of implementation depended on the teacher's interest and skill. Sequential randomization seems an appropriate model here. Define $T_k = 1$ if school k is assigned to treatment, $T_k = 0$ if not; next, $M_{jk} = 1$ if teacher j within school k implements the treatment and $M_{jk} = 0$ if not. We assume that children are nested within schools, but assignment to teachers is uncertain. This generates four potential outcomes for each child i: $Y_{ijk}(T_k, M_{jk}(T_k))$ where T_k and M_{jk} are each binary.

Hong & Raudenbush (2013) regard this as an iconic model of how many policies operate. A policy is framed for each of many organizational units, for example, schools, clinics, or police precincts, the success of which depends on heterogeneous agents (teachers, physicians, police officers), with consequences for the intended recipients (students, patients, community residents). VanderWeele et al. (2013) also model spillovers across classrooms. They reasoned that even if the 4Rs intervention was poorly implemented within a particular classroom, students might benefit from effective implementation in other classrooms. For example, if a teacher in a neighboring classroom was effective at promoting well-regulated behavior among her students, those students would be less likely to engage in altercations with other students during lunch and after school. Hence, the conventional SUTVA must be relaxed for two reasons: because the treatment depends for its implementation on heterogeneous teachers, and because of spillovers within schools.

7.4.2. Exploiting site-to-site variation. Moving to Opportunity is a program that aimed to encourage educational, economic, and health outcomes among low-income residents of inner-city housing projects. Local housing authorities in each of five cities randomly assigned the offer of a housing voucher to residents under the condition that the voucher be used to relocate to low-poverty neighborhood. To study the impact of using mediators (e.g., the posttreatment poverty level of the residence), Kling et al. (2007) proposed what is now called the multisite, multimediator model. Reardon & Raudenbush (2013) derived the assumptions needed to identify causal effects within this model, which has been implemented by Duncan et al. (2011) and Nomi & Raudenbush (2016).

To understand how this model works, we can regard the model for noncompliance as a special case where we have adopted the exclusion restriction. Recall that, within a multisite trial with binary treatment and binary compliance measure, we can define $\beta_j = \gamma_j \delta_j$ as the average effect of treatment assignment within site j, where γ_j is the average effect of assignment on participation and δ_j is the average impact of participation on the outcome in that site. We can therefore write a simple random effects regression model for the estimate impact of random assignment on the outcome in site j:

$$\begin{aligned} \hat{\beta}_j &= \gamma_j \delta_j + \hat{\beta}_j - \beta_j \\ &= \gamma_j \delta + \gamma_j (\delta_j - \delta) + \hat{\beta}_j - \beta_j. \end{aligned} \qquad 27.$$

Note that the model has no intercept, which reflects the exclusion restriction in the compliance model: Random assignment can affect the potential outcome only if it affects participation in the program. The model can be identified by substituting the estimate $\hat{\gamma}_j$ for γ_j under (*a*) the exclusion restriction, (*b*) $\text{Var}(\gamma_j) > 0$ and/or $\text{E}(\gamma_j) > 0$, and (*c*) $\text{E}[(\delta_j - \delta)|\gamma_j] = \text{E}[(\delta_j - \delta)] = 0$. Assumption *a* is standard in noncompliance studies while *c* is strong and uncheckable. For that reason, we do not include this approach in our review of noncompliance above. However, the nice feature of Equation 27 is that it can be expanded to include multiple mediators. Ignorability assumptions such as *c* remain strong, but a major advantage here is that we need not make an assumption that the mediator is randomly assigned. We are, in fact, exploiting the association between the quantities $\hat{\beta}_j$ and $\hat{\gamma}_j$, each of which is protected by randomization. Reardon et al. (2014) propose a bias correction when assumption *c* is violated. If the exclusion restriction comes under doubt, we can add an intercept to the model under the assumption that unobserved mediators are not associated with the observed mediators.

7.4.3. Modeling variation in the mediation process. Qin et al. (2019) extend the weighting approach under sequential randomization to study such variation across sites in the mediation process. Recall that the outcome model associated with this method is remarkably simple. This makes the process of extended the model to include random coefficients quite straightforward without adding undue complexity.

8. GENERALIZABILITY

Although the turn toward large-scale randomized experiments has improved the internal validity of causal claims in education, it is well known that many RCTs use convenience samples (Stuart et al. 2017). Two notable exceptions are Head Start and Job Corps, which took probability samples of program centers from full national frames. Of course, the predominance of convenience samples is also a major concern in other fields that rely on randomized experiments, such as medicine (Huebschmann et al. 2019, Kennedy-Martin et al. 2015). The result is that observed samples in RCTs cannot be regarded as representative of actual populations of interest, so estimated ATEs may not closely reflect the actual ATEs of interest. Some authors have begun to argue that in many settings this external validity bias is likely intolerably large by the standards of internal validity (Bell & Stuart 2016). As a result, the past few decades have also seen renewed attention to methods to measure and improve the generalizability of study designs and actually make causal generalizations under explicit assumptions when the RCT constitutes a convenience sample.

8.1. Defining Generalizability

Some social scientists make a provocative argument that the most interesting generalizations draw upon the underlying science behind why a treatment works the way it does (Deaton & Cartwright 2018, Rothman et al. 2013); this poses causal generalization as a scientific problem about the underlying nature of treatments. Shadish et al. (2002) describe an approach in which the investigator's knowledge of program theory leads to hypotheses about characteristics of settings and persons that ought to amplify or muffle the impact of the intervention. A broadly generalizable program has similar positive effects across those characteristics. If the impact of the program depends on favorable moderators, generalizations are hedged accordingly.

From an alternative point of view, the statistics literature frames causal generalization as a problem of sampling—extending inferences from the observed (convenience) sample to some

well-defined target population about which one has considerable information. In this section we focus on this perspective.

In some cases, the sample is a subset of the target population. In others, the sample is not a subset. For example, the RCT may contain schools in one state while the target population in schools is located in another state. In this case the problem of generalization may be recast as a problem of transportability (Pearl & Bareinboim 2014, Westreich et al. 2017), though this usage has not been adopted universally (Hernán & VanderWeele 2011). Mathematically, the assumptions required to solve these problems are nearly identical (Tipton 2013). The informal language we use to discuss target populations seems to lend itself to the classical finite population theory of survey sampling (Cochran 1977), since often the target population is some specific population with a known frame. Classically, only sample inclusion and treatment assignment are random, though some recent work also treats covariates and outcomes as random (Tipton 2014). Instead, we might use a superpopulation model, which may include an embedded finite population of interest (Deming & Stephan 1941, Graubard & Korn 2002, Hartley & Sielken 1975) or not (Ding et al. 2017; Imbens & Rubin 2015, p. 39). Beyond statistics, our reading of the literature is that education research has not widely confronted these alternatives or reached consensus about which point of view best fits most educational applications.

8.2. Identification and Estimation Strategies for Population Average Treatment Effects

Recently, statisticians have proposed a new type of method for generalization when the experimental sample is a convenience sample (Stuart et al. 2011, Tipton 2013). The strategy is to augment the experimental data with data from a large, representative sample survey or even a census from an interesting population. For example, Tipton (2013) considers a cluster-randomized trial of mathematics software in 92 middle schools in Texas and merges these data with publicly available data on 1,713 middle schools in Texas. Crucially, the experimental data and the auxiliary data must have measured the same covariates. Typically the auxiliary data do not contain the outcome of interest, or it is at least assumed that these units have not received the treatment, so all potential outcomes may be missing in the population. The analyst merges the auxiliary data with the data from the experimental trial and attempts to model the connection between the experimental and auxiliary data.

Some have modeled the sample selection process (sampling propensity score methods); others study the outcome as a function of covariates (response surface modeling) in the experimental sample and predict outcomes in the auxiliary portion. A third approach combines selection modeling and outcome modeling (doubly robust methods). We discuss the crucial assumptions underlying these methods in the next subsection.

8.2.1. Sampling propensity score methods. Using the merged data, the analyst fits the conditional probability of being in the experiment given covariates. One way to use the propensity score for causal generalization is weighting (Kern et al. 2016, Stuart et al. 2011). Units in the experimental sample with high propensity scores are weighted down, while units with low propensity scores are weighted up. The reweighted experimental sample then has approximately the same distribution on covariates as does the larger, representative survey or census. This approach closely resembles methods that use weighting to correct for nonresponse in surveys. An alternative strategy classifies members of the experimental sample into strata with closely matching propensity scores and then weights observations by stratum size (O'Muircheartaigh &

Hedges 2014, Tipton 2013). This method may produce more stable weights, yielding a smaller sampling variance but slightly more bias than simply using propensity scores as weights.

8.2.2. Response surface modeling. In this case the analyst regresses the outcomes on treatment, covariates, and treatment-by-covariate interactions using only the RCT data. Using this fitted model, the analyst predicts both potential outcomes for every member of the auxiliary data set to estimate an ATE in the target population. This approach is less popular in other areas of causal inference, perhaps because researchers feel more confident estimating propensity scores than specifying functional forms for the outcome regression; it is not clear a priori that either strategy is more challenging. One notable exception is Kern et al. (2016), who use a semiparametric Bayesian tree-based regression method that allows flexible functional forms but penalizes complexity through clever default priors (Chipman et al. 2010); it has been used successfully for sample ATEs in observational data (Dorie et al. 2019, Hill 2011).

8.2.3. Doubly robust methods. These approaches model both the sampling process and the outcome regression and, somewhat comfortingly, give consistent population average treatment effect estimates when either model is specified correctly; they have met some success in generalizability but are not guaranteed to strike the optimal bias-variance tradeoff (Dahabreh et al. 2019, Kern et al. 2016).

8.3. Assumptions for Generalizability

The key challenge in causal generalization using the methods just described is that we must explicitly model how the experimental sample provides information about the ATE in the target population. Typically researchers have made two main assumptions to this end: unconfounded sampling and common support between the sample and population (Tipton 2014). Together, these assumptions may be called strongly ignorable sampling, borrowing from the language of Rosenbaum & Rubin (1983).

Define the potential outcomes as $Y_i(1)$ if unit i were treated and $Y_i(0)$ if not, and define $S_i = 1$ if unit i is in the experiment and $S_i = 0$ if that unit is in the auxiliary data. The unconfoundedness assumption states that every unit's person-specific treatment effect is independent of membership in the RCT,

$$Y_i(1) - Y_i(0) \perp S_i | X_i, \qquad 28.$$

and is analogous to the missing at random assumption (Little & Rubin 2002) in the context of missing data. The common support assumption, also referred to as positivity, requires that

$$0 < \Pr(S_i = 1 | X_i) < 1. \qquad 29.$$

The first consequence of Equation 29 is that the sampling process cannot systematically exclude subsets of the target population—this is analogous to a coverage error in survey sampling and is discussed at length by Tipton (2013). A second consequence is that the sample used for making generalizations cannot include units not reflected in the target population. In many studies, common support is not met in at least some populations of interest, so researchers must instead target subpopulations for which common support holds in order to avoid heroic extrapolation.

Often the strongly ignorable sampling assumption is stated in terms of all available covariates, but there are two important nuances to covariate selection that must be addressed statistically and scientifically. First, for the purpose of estimating the population average treatment effect,

generalization methods only need to use covariates that affect both the probability of sampling and the treatment effects (i.e., moderators) (Hill & Su 2013, Stuart et al. 2011). Second, as some have pointed out (D'Amour et al. 2017), unconfoundedness and common support assumptions are antagonistic goals—including more covariates makes unconfoundedness more plausible but also makes common support harder to satisfy as the dimension of the covariate space grows.

8.4. Measuring Generalizability

Methods to assess the generalizability of RCT results are important for a variety of reasons even before an RCT is conducted or generalized estimates are produced. First, they may be useful during study design in considering eligibility criteria and sample selection or recruitment. Second, in analysis, some potential populations of interest may be much more amenable to generalizations from a particular study, so investigators may discover that while one target population is intractable, another one is not. For specific diagnostics, Stuart et al. (2011) consider the standardized mean difference of propensity scores in the sample and population data, while Tipton (2014) propose an index between 0 and 1 that depends directly on more nuanced features of these two propensity score distributions.

8.5. Open Problems

Two factors conspire to make population average treatment effects more difficult to estimate accurately than their sample counterparts—basic sampling variability (Tipton et al. 2017) and the strong ignorability of sampling assumption. These factors suggest that the field needs more methods for thorough but approachable sensitivity analysis (Nguyen et al. 2017) since all causal generalizations from RCTs generally rest on assumptions that data cannot definitively test. Relatedly, we need a better understanding of optimal design for RCTs that are more generalizable (Tipton et al. 2014, Tipton & Peck 2017).

9. SPILLOVERS AND HETEROGENEOUS AGENTS

Recall that SUTVA requires that there be only one version of the treatment and that one person's potential outcome is not influenced by the treatment assignment of other people. However, education is a social process. Any educational policy must be implemented by teachers who are heterogeneous in skill. Hence, students may experience different versions of the treatment in different classrooms or schools. Moreover, students typically learn not only from the teacher but from each other. Thus, one student's treatment assignment may spill over to another student. Recognizing this, educational researchers and other social scientists have devised strategies for relaxing SUTVA.

Again, define $T_{ij} = 1$ if student i in cluster j is assigned to a new treatment and $T_{ij} = 0$ if not. Define $\mathbf{T}_j = (T_{1j}, T_{2j}, \ldots, T_{Nj})$ as the vector of randomly varying treatment assignments of all N_j students in site j, and define the outcome that student i will display under the realization $\mathbf{T}_j = \mathbf{t}_j$ as $Y_{ij}(\mathbf{t}_j)$. The causal effect on the outcome of student i of $\mathbf{t}_j$ relative to an alternative assignment vector $\mathbf{t}_j^*$ is therefore $Y_{ij}(\mathbf{t}_j) - Y_{ij}(\mathbf{t}_j^*)$, the difference between these two potential outcomes. The possibility that perturbing the treatment assignment of any subset of students in a site would modify the potential outcome of all students in that site makes causal inference intractable because the number of potential outcomes for any student is 2^{N_j}. SUTVA requires that $Y_{ij}(\mathbf{t}_j) = Y_{ij}(t_{ij})$.

SUTVA makes causal inference tractable, but educational researchers have questioned its applicability in some cases. Hong & Raudenbush (2006) studied the effect of requiring children

who show little progress during kindergarten to repeat kindergarten rather than being promoted to first grade, a practice known as grade retention. The authors reasoned that the impact of grade retention on a given student would depend upon how many of their classroom peers were retained. It also seemed plausible that the retention of slowly progressing peers might affect the potential outcomes of children who would always be promoted. They proposed a relaxation of SUTVA (Hudgens & Halloran 2008, Sobel 2006) such that $Y_{ij}(\mathbf{t}_j) = Y_{ij}(t_{ij}, f(\mathbf{t}_j))$ where $f(\mathbf{t}_j)$ is a scalar function of the entire vector of treatment assignments in site j. For example, $f(\mathbf{t}_j) = \mathbf{t}_j^T \mathbf{t}_j / n_j$, the fraction of students in school j who are retained, might be regarded as capturing the mechanism by which peer treatment assignments affect a student's potential outcome. Hence, $Y_{ij}(0, f(\mathbf{t}_j)) - Y_{ij}(0, f(\mathbf{t}_j^*))$ would represent the effect on a student who would be promoted of having $f(\mathbf{t}_j)$ rather than $f(\mathbf{t}_j^*)$ of her peers retained. Notice that $f(\mathbf{t}_j)$ is a site-level causal variable, suggesting a useful experiment in which clusters are randomly assigned to values of $f(\mathbf{t}_j)$ and eligible students within clusters are assigned at random to be retained (Hudgens & Halloran 2008). Basse et al. (2019) studied an intervention to reduce truancy and used this framework to study the effect of a sibling's treatment assignment on untreated siblings. Key assumptions in this framework are (*a*) intact sites, that is, students do not migrate from one site to the other during the course of the study, and (*b*) no interference between sites. Randomization by cluster may be regarded as a strategy for relaxing SUTVA. Suppose that cluster j is assigned at random to treatments. We can then write the child-specific causal effect as

$$B_{ij} \equiv Y_{ij}(1, 1, \ldots, 1) - Y_{ij}(0, 0, \ldots, 0). \qquad 30.$$

Thus, assignment of the entire cluster affects the treatment assignment of every student in that cluster and we have identification without SUTVA.

10. PROMISING EXPERIMENTAL DESIGNS

In this section we review two promising experimental designs that are leading to success in other fields and are beginning to arouse interest in education. We hope that education research will see another methodological renaissance, this time focused on experimental designs that answer even richer questions about real-world interventions.

10.1. Sequential Multiple Assignment Randomized Trial Designs for Adaptive Interventions

Adaptive interventions, also called dynamic treatment regimes, are multistage interventions that include a sequence of decision rules that use current information about each person to determine individualized treatment at each stage (Chakraborty & Murphy 2014). The SMART is the basic experimental design used to study adaptive interventions, and it sequentially randomizes subjects to the treatments available at each stage according to their previous outcomes (Murphy 2005). Estimating and describing uncertainty about the optimal dynamic treatment regime (i.e., the set of adaptive decision rules that lead to the best average outcome at the end of the intervention) are challenging because the parameters of interest are nonsmooth functions of the model, so most classical statistical theory relying on asymptotic expansions is not relevant (Laber et al. 2014, Luedtke & van der Laan 2016).

These methods have been primarily used in behavioral science and clinical trials (Lei et al. 2012, Wahed & Tsiatis 2004), though adaptive interventions are clearly of great interest in education (Almirall et al. 2018, Raudenbush 2008). For example, Kilbourne et al. (2018) describe

an ongoing SMART on training school staff to provide cognitive behavioral therapy (CBT) to students seeking mental health services. More than 100 high schools in Michigan with over 200 school staff members were cluster-randomized to receive two types of CBT training, and then after three months, within both initial randomization groups, schools were again randomized to different follow-up training. In another recent example, Kim et al. (2019) used a SMART to refine a literacy intervention for kindergarten to grade 2 students by considering a second stage targeted at improving outcomes for children who were not helped by the first-stage treatment. Of course, all teaching can also be considered an adaptive intervention in which teachers continually assess their students' understanding and then tailor instruction accordingly. In one study, children aged 3–5 were assessed every 10 weeks on early math skills, and teachers used these assessments to make updated instructional plans (Raudenbush et al. 2020). One might also imagine studying adaptive instruction at a much more fine-grained scale in educational software that provides near-constant assessment of student knowledge, perhaps borrowing from methods for so-called microrandomized trials in mobile health (Boruvka et al. 2018, Klasnja et al. 2015).

Observational data have also been used to estimate optimal dynamic treatment regimes in educational examples including intensive elementary mathematics instruction (Hong & Raudenbush 2008) and tracking in high schools (Zajonc 2012), though the additional required ignorability assumptions may be very strong. There is great room to apply SMARTs more widely in education. Some methods exist to study adaptive interventions in multilevel settings (Kilbourne et al. 2013, NeCamp et al. 2017) but this is an understudied area, especially in the context of multisite trials where much might be learned about how these complex interventions differ in effectiveness across sites.

10.2. Factorial Designs for Multicomponent Interventions

Long after their storied past in classical experimental design, factorial designs are receiving renewed attention as part of the MOST (multiphase optimization strategy) framework for developing and improving complex interventions (Collins 2018, Collins & Kugler 2018, Dziak et al. 2012). The MOST approach, which has been used predominantly in behavioral medicine (McClure et al. 2012, Pellegrini et al. 2014), is analogous to the multiple phases of clinical trials for drug approval but is focused on interventions that have multiple components that may be manipulated. Factorial experiments here are used as an exploratory tool to identify the form of an intervention with the optimal combination of features. Formal evaluation of this optimized intervention is based on a traditional confirmatory two-arm randomized trial (Wyrick et al. 2014). This research strategy is highly promising in education, where interventions are often complex by necessity and even seemingly straightforward interventions may be implemented in many ways (McLaughlin 1987).

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

We would like to thank Elizabeth Tipton, Luke Miratrix, Tom Louis, Xu Qin, Avi Feller, Jessaca Spybrook, Elizabeth Stuart, Sean Reardon, and Tyler VanderWeele for their helpful comments.

LITERATURE CITED

Allen JP, Pianta RC, Gregory A, Mikami AY, Lun J. 2011. An interaction-based approach to enhancing secondary school instruction and student achievement. *Science* 333(6045):1034–37

Almirall D, Kasari C, McCaffrey DF, Nahum-Shani I. 2018. Developing optimized adaptive interventions in education. *J. Res. Educ. Eff.* 11:27–34

Angrist JD, Cohodes SR, Dynarski SM, Pathak PA, Walters CR. 2016. Stand and deliver: effects of Boston's charter high schools on college preparation, entry, and choice. *J. Labor Econ.* 34(2):275–318

Angrist JD, Imbens GW, Rubin DB. 1996. Identification of causal effects using instrumental variables. *J. Am. Stat. Assoc.* 91(434):444

Angrist JD, Pathak PA, Walters CR. 2013. Explaining charter school effectiveness. *Am. Econ. J. Appl. Econ.* 5(4):1–27

Baron RM, Kenny DA. 1986. The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *J. Personal. Soc. Psychol.* 51(6):1173–82

Basse GW, Feller A, Toulis P. 2019. Randomization tests of causal effects under interference. *Biometrika* 106(2):487–94

Bell SH, Stuart EA. 2016. On the "where" of social experiments: the nature and extent of the generalizability problem. *New Dir. Eval.* 2016(152):47–59

Berger RL, Boos DD. 1994. *P* values maximized over a confidence set for the nuisance parameter. *J. Am. Stat. Assoc.* 89(427):1012–16

Bitler MP, Hoynes HW, Domina T. 2014. *Experimental evidence on distributional effects of Head Start*. NBER Work. Pap. 20434

Bloom HS, Raudenbush SW, Weiss MJ, Porter K. 2017. Using multisite experiments to study cross-site variation in treatment effects: a hybrid approach with fixed intercepts and a random treatment coefficient. *J. Res. Educ. Eff.* 10(4):817–42

Bloom HS, Richburg-Hayes L, Black AR. 2007. Using covariates to improve precision for studies that randomize schools to evaluate educational interventions. *Educ. Eval. Policy Anal.* 29:30–59

Bloom HS, Unterman R. 2014. Can small high schools of choice improve educational prospects for disadvantaged students? *J. Policy Anal. Manag.* 33(2):290–319

Bloom HS, Weiland C. 2015. Quantifying variation in Head Start effects on young children's cognitive and socio-emotional skills using data from the National Head Start Impact Study. *SSRN Electron. J.* **http://dx.doi.org/10.2139/ssrn.2594430**

Borman GD, Dowling NM, Schneck C. 2008. A multisite cluster randomized field trial of open court reading. *Educ. Eval. Policy Anal.* 30(4):389–407

Borman GD, Slavin RE, Cheung ACK, Chamberlain AM, Madden NA, Chambers B. 2007. Final reading outcomes of the national randomized field trial of Success for All. *Am. Educ. Res. J.* 44(3):701–31

Boruvka A, Almirall D, Witkiewitz K, Murphy SA. 2018. Assessing time-varying causal effect moderation in mobile health. *J. Am. Stat. Assoc.* 113(523):1112–21

Bound J, Jaeger DA, Baker RM. 1995. Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *J. Am. Stat. Assoc.* 90(430):443–50

Boyd-Zaharias J. 1999. Project STAR: the story of the Tennessee class-size study. *Am. Educ.* 23(2):30–36

Bryk AS, Weisberg HI. 1976. Value-added analysis: a dynamic approach to the estimation of treatment effects. *J. Educ. Stat.* 1(2):127–55

Bullock JG, Green DP, Ha SE. 2010. Yes, but what's the mechanism? (Don't expect an easy answer). *J. Personal. Soc. Psychol.* 98(4):550–58

Chakraborty B, Murphy SA. 2014. Dynamic treatment regimes. *Annu. Rev. Stat. Appl.* 1:447–64

Chipman HA, George EI, McCulloch RE. 2010. BART: Bayesian additive regression trees. *Ann. Appl. Stat.* 4:266–98

Clark MA, Gleason PM, Tuttle CC, Silverberg MK. 2015. Do charter schools improve student achievement? *Educ. Eval. Policy Anal.* 37(4):419–36

Cochran WG. 1977. *Sampling Techniques*. New York: Wiley. 3rd ed.

Collins LM. 2018. *Optimization of Behavioral, Biobehavioral, and Biomedical Interventions: The Multiphase Optimization Strategy (MOST)*. New York: Springer
Collins LM, Kugler KC, eds. 2018. *Optimization of Behavioral, Biobehavioral, and Biomedical Interventions: Advanced Topics*. New York: Springer
Confrey J. 2006. Comparing and contrasting the National Research Council report on evaluating curricular effectiveness with the What Works Clearinghouse approach. *Educ. Eval. Policy Anal.* 28(3):195–213
Cook TD. 2002. Randomized experiments in educational policy research: a critical examination of the reasons the educational evaluation community has offered for not doing them. *Educ. Eval. Policy Anal.* 24(3):175–99
Dahabreh IJ, Robertson SE, Tchetgen EJ, Stuart EA, Hernán MA. 2019. Generalizing causal inferences from individuals in randomized trials to all trial-eligible individuals. *Biometrics* 75(2):685–94
D'Amour A, Ding P, Feller A, Lei L, Sekhon J. 2017. Overlap in observational studies with high-dimensional covariates. arXiv:1711.02582 [math.ST]
Deaton A, Cartwright N. 2018. Understanding and misunderstanding randomized controlled trials. *Soc. Sci. Med.* 210:2–21
Deming WE, Stephan FF. 1941. On the interpretation of censuses as samples. *J. Am. Stat. Assoc.* 36(213):45–49
Ding P, Feller A, Miratrix L. 2016. Randomization inference for treatment effect variation. *J. R. Stat. Soc. B* 78(3):655–71
Ding P, Feller A, Miratrix L. 2019. Decomposing treatment effect variation. *J. Am. Stat. Assoc.* 114(525):304–17
Ding P, Li X, Miratrix LW. 2017. Bridging finite and super population causal inference. *J. Causal Inference* 5(2):20160027
Dobbie W, Fryer RG. 2013. Getting beneath the veil of effective schools: evidence from New York City. *Am. Econ. J. Appl. Econ.* 5(4):28–60
Donner A, Birkett N, Buck C. 1981. Randomization by cluster: sample size requirements and analysis. *Am. J. Epidemiol.* 114(6):906–14
Dorie V, Hill J, Shalit U, Scott M, Cervone D. 2019. Automated versus do-it-yourself methods for causal inference: lessons learned from a data analysis competition. *Stat. Sci.* 34:43–68
Duncan GJ, Morris PA, Rodrigues C. 2011. Does money really matter? Estimating impacts of family income on young children's achievement with data from random-assignment experiments. *Dev. Psychol.* 47(5):1263–79
Duncan OD. 1966. Path analysis: sociological examples. *Am. J. Sociol.* 72:1–16
Dziak JJ, Nahum-Shani I, Collins LM. 2012. Multilevel factorial experiments for developing behavioral interventions: power, sample size, and resource considerations. *Psychol. Methods* 17(2):153–75
Feller A, Grindal T, Miratrix L, Page LC. 2016. Compared to what? Variation in the impacts of early childhood education by alternative care type. *Ann. Appl. Stat.* 10(3):1245–85
Finn JD, Achilles CM. 1990. Answers and questions about class size: a statewide experiment. *Am. Educ. Res. J.* 27(3):557–77
Fortmann SP, Flora JA, Winkleby MA, Schooler C, Taylor CB, Farquhar JW. 1995. Community intervention trials: reflections on the Stanford Five-City Project experience. *Am. J. Epidemiol.* 142(6):576–86
Frangakis CE, Rubin DB. 2002. Principal stratification in causal inference. *Biometrics* 58:21–29
Graubard BI, Korn EL. 2002. Inference for superpopulation parameters using sample surveys. *Stat. Sci.* 17:73–96
Green DP, Kern HL. 2012. Modeling heterogeneous treatment effects in survey experiments with Bayesian additive regression trees. *Public Opin. Q.* 76(3):491–511
Guo W, Ji Y, Catenacci DVT. 2017. A subgroup cluster-based Bayesian adaptive design for precision medicine: SCUBA. *Biometrics* 73(2):367–77
Haavelmo T. 1943. The statistical implications of a system of simultaneous equations. *Econometrica* 11:1–12
Hartley HO, Sielken RL. 1975. A "super-population viewpoint" for finite population sampling. *Biometrics* 31(2):411–22
Hassrick EM, Raudenbush SW, Rosen LS. 2017. *The Ambitious Elementary School: Its Conception, Design, and Implications for Educational Equality*. Chicago: Univ. Chicago Press

Heckman JJ. 1979. Sample selection bias as a specification error. *Econometrica* 47:153
Heckman JJ, Robb R. 1985. Alternative methods for evaluating the impact of interventions. *J. Econom.* 30(1–2):239–67
Heckman JJ, Schmierer D, Urzua S. 2010. Testing the correlated random coefficient model. *J. Econom.* 158(2):177–203
Heckman JJ, Vytlacil E. 2001. Policy-relevant treatment effects. *Am. Econ. Rev.* 91(2):107–11
Hedges LV, Hedberg EC. 2013. Intraclass correlations and covariate outcome correlations for planning two- and three-level cluster-randomized experiments in education. *Eval. Rev.* 37(6):445–89
Hernán MA, VanderWeele TJ. 2011. Compound treatments and transportability of causal inference. *Epidemiology* 22(3):368–77
Hill JL. 2011. Bayesian nonparametric modeling for causal inference. *J. Comput. Graph. Stat.* 20:217–40
Hill JL, Su Y-S. 2013. Assessing lack of common support in causal inference using Bayesian nonparametrics: implications for evaluating the effect of breastfeeding on children's cognitive outcomes. *Ann. Appl. Stat.* 7(3):1386–1420
Holland PW. 1986. Statistics and causal inference. *J. Am. Stat. Assoc.* 81(396):945–60
Holland PW. 1988. Causal inference, path analysis, and recursive structural equations models. *Sociol. Methodol.* 18:449–84
Hong G. 2015. *Causality in a Social World: Moderation, Mediation and Spill-over*. New York: Wiley
Hong G, Raudenbush SW. 2006. Evaluating kindergarten retention policy: a case study of causal inference for multilevel observational data. *J. Am. Stat. Assoc.* 101(475):901–10
Hong G, Raudenbush SW. 2008. Causal inference for time-varying instructional treatments. *J. Educ. Behav. Stat.* 33(3):333–62
Hong G, Raudenbush SW. 2013. Heterogeneous agents, social interactions, and causal inference. In *Handbook of Causal Analysis for Social Research*, ed. SL Morgan, pp. 331–52. New York: Springer
Hudgens MG, Halloran ME. 2008. Toward causal inference with interference. *J. Am. Stat. Assoc.* 103(482):832–42
Huebschmann AG, Leavitt IM, Glasgow RE. 2019. Making health research matter: a call to increase attention to external validity. *Annu. Rev. Public Health* 40:45–63
Imai K, Keele L, Tingley D. 2010. A general approach to causal mediation analysis. *Psychol. Methods* 15(4):309–31
Imai K, Ratkovic M. 2013. Estimating treatment effect heterogeneity in randomized program evaluation. *Ann. Appl. Stat.* 7:443–70
Imbens GW, Rubin DB. 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge, UK: Cambridge Univ. Press. 1st ed.
Kennedy-Martin T, Curtis S, Faries D, Robinson S, Johnston J. 2015. A literature review on the representativeness of randomized controlled trial samples and implications for the external validity of trial results. *Trials* 16:495
Kern HL, Stuart EA, Hill J, Green DP. 2016. Assessing methods for generalizing experimental impact estimates to target populations. *J. Res. Educ. Eff.* 9:103–27
Kilbourne AM, Abraham KM, Goodrich DE, Bowersox NW, Almirall D, et al. 2013. Cluster randomized adaptive implementation trial comparing a standard versus enhanced implementation intervention to improve uptake of an effective re-engagement program for patients with serious mental illness. *Implement. Sci.* 8:136
Kilbourne AM, Smith SN, Choi SY, Koschmann E, Liebrecht C, et al. 2018. Adaptive school-based implementation of CBT (ASIC): clustered-SMART for building an optimized adaptive implementation intervention to improve uptake of mental health interventions in schools. *Implement. Sci.* 13:119
Kim JS, Asher CA, Burkhauser M, Mesite L, Leyva D. 2019. Using a sequential multiple assignment randomized trial (SMART) to develop an adaptive K–2 literacy intervention with personalized print texts and app-based digital activities. *AERA Open*. **https://doi.org/10.1177/2332858419872701**
Klar N, Donner A. 1997. The merits of matching in community intervention trials: a cautionary tale. *Stat. Med.* 16(15):1753–64

Klasnja P, Hekler EB, Shiffman S, Boruvka A, Almirall D, et al. 2015. Microrandomized trials: an experimental design for developing just-in-time adaptive interventions. *Health Psychol.* 34(Suppl.):1220–28
Kline P, Walters CR. 2019. On Heckits, LATE, and numerical equivalence. *Econometrica* 87(2):677–96
Kling JR, Liebman JB, Katz LF. 2007. Experimental analysis of neighborhood effects. *Econometrica* 75:83–119
Krueger AB, Whitmore DM. 2001. The effect of attending a small class in the early grades on college-test taking and middle school test results: evidence from Project Star. *Econ. J.* 111(468):1–28
Laber EB, Lizotte DJ, Qian M, Pelham WE, Murphy SA. 2014. Dynamic treatment regimes: technical challenges and applications. *Electron. J. Stat.* 8:1225–72
Lei H, Nahum-Shani I, Lynch K, Oslin D, Murphy SA. 2012. A "SMART" design for building individualized treatment sequences. *Annu. Rev. Clin. Psychol.* 8:21–48
Little RJ, Rubin DB. 2002. *Statistical Analysis with Missing Data*. New York: Wiley. 2nd ed.
Louis TA. 1984. Estimating a population of parameter values using Bayes and empirical Bayes methods. *J. Am. Stat. Assoc.* 79(386):393
Luedtke AR, van der Laan MJ. 2016. Statistical inference for the mean outcome under a possibly non-unique optimal treatment strategy. *Ann. Stat.* 44(2):713–42
McClure JB, Derry H, Riggs KR, Westbrook EW, St. John J, et al. 2012. Questions about quitting (Q2): design and methods of a Multiphase Optimization Strategy (MOST) randomized screening experiment for an online, motivational smoking cessation intervention. *Contemp. Clin. Trials* 33(5):1094–102
McLaughlin MW. 1987. Learning from experience: lessons from policy implementation. *Educ. Eval. Policy Anal.* 9(2):171
Mosteller F. 1995. The Tennessee study of class size in the early school grades. *Future Child.* 5(2):113
Mosteller F, Boruch RF, eds. 2002. *Evidence Matters: Randomized Trials in Education Research*. Washington, DC: Brookings Inst.
Murphy SA. 2005. An experimental design for the development of adaptive treatment strategies. *Stat. Med.* 24(10):1455–81
Murray DM. 1995. Design and analysis of community trials: lessons from the Minnesota Heart Health Program. *Am. J. Epidemiol.* 142(6):569–75
NeCamp T, Kilbourne A, Almirall D. 2017. Comparing cluster-level dynamic treatment regimens using sequential, multiple assignment, randomized trials: regression estimation and sample size considerations. *Stat. Methods Med. Res.* 26(4):1572–89
Neyman J. 1935. Statistical problems in agricultural experimentation. *Suppl. J. R. Stat. Soc.* 2(2):107
Nguyen TQ, Ebnesajjad C, Cole SR, Stuart EA. 2017. Sensitivity analysis for an unobserved moderator in RCT-to-target-population generalization of treatment effects. *Ann. Appl. Stat.* 11:225–47
Nomi T, Allensworth E. 2009. "Double-dose" algebra as an alternative strategy to remediation: effects on students' academic outcomes. *J. Res. Educ. Eff.* 2(2):111–48
Nomi T, Raudenbush SW. 2016. Making a success of "Algebra for All": the impact of extended instructional time and classroom peer skill in Chicago. *Educ. Eval. Policy Anal.* 38(2):431–51
O'Muircheartaigh C, Hedges LV. 2014. Generalizing from unrepresentative experiments: a stratified propensity score approach. *J. R. Stat. Soc. C* 63(2):195–210
Paddock SM, Ridgeway G, Lin R, Louis TA. 2006. Flexible distributions for triple-goal estimates in two-stage hierarchical models. *Comput. Stat. Data Anal.* 50(11):3243–62
Pearl J. 2001. Direct and indirect effects. In *Proceedings of the 17th Conference on Uncertainty in Artificial Intelligence*, ed. J Breese, D Koller, pp. 411–20. San Francisco: Morgan Kaufmann
Pearl J, Bareinboim E. 2014. External validity: from do-calculus to transportability across populations. *Stat. Sci.* 29(4):579–95
Pellegrini CA, Hoffman SA, Collins LM, Spring B. 2014. Optimization of remotely delivered intensive lifestyle treatment for obesity using the Multiphase Optimization Strategy: Opt-IN study protocol. *Contemp. Clin. Trials* 38(2):251–59
Petersen ML, Sinisi SE, van der Laan MJ. 2006. Estimation of direct causal effects. *Epidemiology* 17(3):276–84
Puma M, Bell S, Cook R, Heid C, Shapiro G, et al. 2010. *Head Start impact study final report*. Rep., US Dep. Health Hum. Serv., Washington, DC

Qin X, Hong G, Deutsch J, Bein E. 2019. Multisite causal mediation analysis in the presence of complex sample and survey designs and non-random non-response. *J. R. Stat. Soc. A.* **https://doi.org/10.1111/rssa.12446**

Raudenbush SW. 1997. Statistical analysis and optimal design for cluster randomized trials. *Psychol. Methods* 2(2):173–85

Raudenbush SW. 2008. Advancing educational policy by advancing research on instruction. *Am. Educ. Res. J.* 45(1):206–30

Raudenbush SW, Bloom HS. 2015. Learning about and from a distribution of program impacts using multisite trials. *Am. J. Eval.* 36(4):475–99

Raudenbush SW, Hernandez M, Goldin-Meadow S, Carrazza C, Leslie D, et al. 2020. *Longitudinally adaptive instruction increases the numerical skill of pre-school children*. Work. Pap., Dep. Psychol., Univ. Chicago

Raudenbush SW, Martinez A, Spybrook J. 2007. Strategies for improving precision in group-randomized experiments. *Educ. Eval. Policy Anal.* 29:5–29

Raudenbush SW, Reardon SF, Nomi T. 2012. Statistical analysis for multisite trials using instrumental variables with random coefficients. *J. Res. Educ. Eff.* 5(3):303–32

Raudenbush SW, Schwartz D. 2019. *Estimating the average treatment effect in a multisite trial with heterogeneous treatment effects*. Work. Pap., Univ. Chicago

Reardon SF, Raudenbush SW. 2013. Under what assumptions do site-by-treatment instruments identify average causal effects? *Sociol. Methods Res.* 42(2):143–63

Reardon SF, Unlu F, Zhu P, Bloom HS. 2014. Bias and bias correction in multisite instrumental variables analysis of heterogeneous mediator effects. *J. Educ. Behav. Stat.* 39:53–86

Rosenbaum PR, Rubin DB. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70:41–55

Rothman KJ, Gallacher JE, Hatch EE. 2013. Why representativeness should be avoided. *Int. J. Epidemiol.* 42(4):1012–14

Rubin DB. 1978. Bayesian inference for causal effects: the role of randomization. *Ann. Stat.* 6:34–58

Rubin DB. 1981. Estimation in parallel randomized experiments. *J. Educ. Stat.* 6(4):377–401

Rubin DB. 1986. Comment: which ifs have causal answers. *J. Am. Stat. Assoc.* 81(396):961–62

Schochet P. 2015. *Statistical theory for the RCT-YES software: design-based causal inference for RCTs*. Rep., Natl. Cent. Educ. Eval. Reg. Assist., US Dep. Educ.

Shadish WR, Cook TD, Campbell DT. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston: Houghton Mifflin

Shen W, Louis TA. 1998. Triple-goal estimates in two-stage hierarchical models. *J. R. Stat. Soc. B* 60(2):455–71

Sobel ME. 2006. What do randomized studies of housing mobility demonstrate? Causal inference in the face of interference. *J. Am. Stat. Assoc.* 101(476):1398–1407

Spybrook J. 2014. Detecting intervention effects across context: an examination of the precision of cluster randomized trials. *J. Exp. Educ.* 82(3):334–57

Spybrook J, Shi R, Kelcey B. 2016. Progress in the past decade: an examination of the precision of cluster randomized trials funded by the U.S. Institute of Education Sciences. *Int. J. Res. Method Educ.* 39(3):255–67

Stuart EA, Bell SH, Ebnesajjad C, Olsen RB, Orr LL. 2017. Characteristics of school districts that participate in rigorous national educational evaluations. *J. Res. Educ. Eff.* 10:168–206

Stuart EA, Cole SR, Bradshaw CP, Leaf PJ. 2011. The use of propensity scores to assess the generalizability of results from randomized trials: use of propensity scores to assess generalizability. *J. R. Stat. Soc. A* 174(2):369–86

Tipton E. 2013. Improving generalizations from experiments using propensity score subclassification: assumptions, properties, and contexts. *J. Educ. Behav. Stat.* 38(3):239–66

Tipton E. 2014. How generalizable is your experiment? An index for comparing experimental samples and populations. *J. Educ. Behav. Stat.* 39(6):478–501

Tipton E, Hallberg K, Hedges LV, Chan W. 2017. Implications of small samples for generalization: adjustments and rules of thumb. *Eval. Rev.* 41(5):472–505

Tipton E, Hedges L, Vaden-Kiernan M, Borman G, Sullivan K, Caverly S. 2014. Sample selection in randomized experiments: a new method using propensity score stratified sampling. *J. Res. Educ. Eff.* 7:114–35
Tipton E, Peck LR. 2017. A design-based approach to improve external validity in welfare policy evaluations. *Eval. Rev.* 41(4):326–56
VanderWeele TJ. 2015. *Explanation in Causal Inference: Methods for Mediation and Interaction*. Oxford, UK: Oxford Univ. Press
VanderWeele TJ, Hong G, Jones SM, Brown JL. 2013. Mediation and spillover effects in group-randomized trials: a case study of the 4Rs educational intervention. *J. Am. Stat. Assoc.* 108(502):469–82
Wager S, Athey S. 2018. Estimation and inference of heterogeneous treatment effects using random forests. *J. Am. Stat. Assoc.* 113(523):1228–42
Wahed AS, Tsiatis AA. 2004. Optimal estimator for the survival distribution and related quantities for treatment policies in two-stage randomization designs in clinical trials. *Biometrics* 60:124–33
Wald A. 1940. The fitting of straight lines if both variables are subject to error. *Ann. Math. Stat.* 11(3):284–300
Walters CR. 2015. Inputs in the production of early childhood human capital: evidence from head start. *Am. Econ. J. Appl. Econ.* 7(4):76–102
Wang R, Ware JH. 2013. Detecting moderator effects using subgroup analyses. *Prev. Sci.* 14(2):111–20
Weiss MJ, Bloom HS, Verbitsky-Savitz N, Gupta H, Vigil AE, Cullinan DN. 2017. How much do the effects of education and training programs vary across sites? Evidence from past multisite randomized trials. *J. Res. Educ. Eff.* 10(4):843–76
Westreich D, Edwards JK, Lesko CR, Stuart E, Cole SR. 2017. Transportability of trial results using inverse odds of sampling weights. *Am. J. Epidemiol.* 186(8):1010–14
Wright S. 1921. Correlation and causation. *J. Agric. Res.* 20(7):557–85
Wyrick DL, Rulison KL, Fearnow-Kenney M, Milroy JJ, Collins LM. 2014. Moving beyond the treatment package approach to developing behavioral interventions: addressing questions that arose during an application of the Multiphase Optimization Strategy (MOST). *Transl. Behav. Med.* 4(3):252–59
Zajonc T. 2012. Bayesian inference for dynamic treatment regimes: mobility, equity, and efficiency in student tracking. *J. Am. Stat. Assoc.* 107(497):80–92
Zhelonkin M, Genton MG, Ronchetti E. 2016. Robust inference in sample selection models. *J. R. Stat. Soc. B* 78(4):805–27

Annual Review of Statistics and Its Application

Volume 7, 2020

Contents

Errata

An online log of corrections to *Annual Review of Statistics and Its Application* articles may be found at http://www.annualreviews.org/errata/statistics