MDRC Working Papers on Research Methodology

Estimating the Standard Error of the Impact Estimator in Individually Randomized Trials with Clustering

Michael J. Weiss
MDRC

J. R. Lockwood
ETS

Daniel F. McCaffrey
ETS

April 2014

Acknowledgments

This paper was supported by funding from the Judith Gueron Fund for Methodological Innovation in Social Policy Research at MDRC. The authors thank Kristin Porter, Howard Bloom, Alex Mayer, Dan Cullinan, Tim Moses, Yi-Hsuan Lee, and Charles Michalopoulos for their extremely helpful comments. The ideas presented and positions taken in the paper are solely the responsibility of the authors. The findings and conclusions in this paper do not necessarily represent the official positions or policies of the funders.

Dissemination of MDRC publications is supported by the following funders that help finance MDRC's public policy outreach and expanding efforts to communicate the results and implications of our work to policymakers, practitioners, and others: The Annie E. Casey Foundation, The Harry and Jeanette Weinberg Foundation, Inc., The Kresge Foundation, Laura and John Arnold Foundation, Sandler Foundation, and The Starr Foundation.

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

Corresponding Author:

Michael J. Weiss
MDRC
16 East 34th Street (19th Floor)
New York, NY 10016
(212) 340-8651
Michael.weiss@mdrc.org

Abstract

In many experimental evaluations in the social and medical sciences, individuals are randomly assigned to a treatment arm or a control arm of the experiment. After treatment assignment is determined, individuals within one or both experimental arms are frequently grouped together (e.g., within classrooms or schools, through shared case managers, in group therapy sessions, or through shared doctors) to receive services. Consequently, there may be within-group correlations in outcomes resulting from (1) the process that sorts individuals into groups, (2) service provider effects, and/or (3) peer effects. When estimating the standard error of the impact estimate, it may be necessary to account for within-group correlations in outcomes. This article demonstrates that correlations in outcomes arising from nonrandom sorting of individuals into groups leads to bias in the estimated standard error of the impact estimator reported by common estimation approaches.

Keywords: randomized trials, nested designs, clustering, standard error, individually randomized group treatment

Contents

Acknowledgments ii
Abstract iii
List of Exhibits vii

Introduction 1

Section

1 Sources and Lack of Independence 3
The Processes That Sort Individuals into Groups 3
Service Provider Effects 4
Peer Effects 4

2 Random Provider Effects 7

3 Fixed Provider Effects 11

4 Simulation Study 15
The Data-Generating Mechanisms 15

5 Results of the Simulation Study 19

6 Discussion and Conclusions 23
Practical Recommendations for Moving Forward 24

References 27

List of Exhibits

Table

1	Data-Generation Mechanisms (DGM) for the Simulation Study	16
2	Bias of the Standard Error of the Impact Estimator (True Impact = 0)	20

Introduction

Randomized experiments have become an increasingly popular design to evaluate the effectiveness of social policy interventions (Michalopoulos, 2005; Spybrook, 2008). Many of these interventions are delivered to clients (e.g., students or patients) by service providers (e.g., teachers or therapists). Intervention delivery can occur in a group or cluster (e.g., classrooms or schools, or group therapy sessions) or one-on-one, with many individuals receiving the intervention from the same service provider.

One of the more popular experimental designs has been labeled by public health researchers the Individually Randomized Group Treatment (IRGT) trial (Pals et al., 2008). The name reflects the fact that, following the terminology of Bauer, Sterba, and Hallfors (2008), individuals are randomized to experimental conditions, or *arms*, and the treatment is delivered at the *group* level within an arm. This can occur, for example, in a public health intervention where patients are randomly assigned to experimental conditions and the intervention is delivered in a group therapy session; or in a social welfare program where persons or families are randomly assigned to experimental conditions and the intervention is delivered by case managers, each of whom provides services to multiple program participants. The key characteristic of IRGTs is that randomization occurs at the individual level (often referred to as Level 1), and treatment occurs at the group level (often referred to as Level 2).

This research design has been used in many evaluations across many fields of research. For examples in education see: Abdulkadiroglu et al. (2009); Bernstein et al. (2010); Bloom, Thompson, and Unterman (2010); Corrin et al. (2008); Decker, Mayer, and Glazerman (2004); Dynarski and Gleason (2002); Hoxby and Murarka (2009); Kemple, Herlihy, and Smith (2005); Lang et al. (2009); Love et al. (2002); Richburg-Hayes, Visher, and Bloom (2008); Scrivener and Weiss (2009); Weiss, Visher, and Weissman (2012); Wolf et al. (2009). For a list of examples in psychotherapy, see Pals et al. (2008).[1]

Researchers have voiced the concern that in IRGTs, observations within groups are not independent; thus, it has been suggested that analytic adjustments be made in order to obtain an unbiased estimate of the standard error of the impact estimator and to not inflate the likelihood of making type I errors (Bauer, Sterba, and Hallfors, 2008; Crits-Christoph and Mintz, 1991; Roberts and Roberts, 2005). By "impact estimator," we are referring to the function of the observed data used to estimate the average causal effect of the intervention on the population of units. In most cases, although individuals are randomly assigned to an experimental arm, the

[1]Note that some examples of the regression discontinuity design can be thought of as analogous to the IRGT trial (for example see Calcagno and Long, 2008) and are thus subject to the same concerns raised in this paper.

groupings within arms are determined after random assignment through a nonrandom process. In this situation, this paper seeks to clarify if, how, and under what conditions it is necessary to adjust the standard error of the impact estimator to account for the potential lack of independence among observations from individuals treated within the same groups.

Some researchers have suggested that random effects models should be used to address lack of independence among observations from individuals treated in the same group (Crits-Christoph and Mintz, 1991; Crits-Christoph, Tu, and Gallop, 2003; Pals et al., 2008; Roberts and Roberts, 2005; Serlin, Wampold, and Levin, 2003). Others have suggested that fixed effects models should be used (Siemer and Joorman, 2003a, 2003b). Strong arguments have been made in favor of each approach, which are briefly discussed. The paper offers two unique contributions to this topic. First, the three major potential sources of lack of independence are clarified: (1) nonrandom sorting into groups, (2) provider effects, and (3) peer effects. Second, through analytic description and simulations, we demonstrate that the lack of independence caused by the nonrandom sorting of individuals into groups can bias the estimator of the standard error of the impact estimator for both random effects and fixed effects models, likely yielding standard errors that are biased upward in the case of random effects and biased downward in the case of fixed effects.

The remainder of the paper is divided into six sections. Section 1 describes three sources of lack of independence of observations in IRGTs. Section 2 discusses the random effects approach to dealing with lack of independence of observations and shows why nonrandom sorting of individuals into groups can lead to positive bias in the standard error estimator. Section 3 discusses the fixed effects approach to dealing with lack of independence of observations and shows why nonrandom sorting of individuals into groups can lead to negative bias in the standard error estimator. In Section 4, a set of simulations are introduced to demonstrate these biases with simulated data. Section 5 presents the simulation results, and Section 6 offers concluding thoughts.

Section 1

Sources of Lack of Independence

The literature on individually randomized experiments contains considerable discussion of the violation of independence of observations and the need to account for this lack of independence in statistical analyses (Baldwin, Bauer, Stice, and Rohde, 2011; Bauer, Sterba, and Hallfors, 2008; Crits-Christoph, Tu, and Gallop, 1991; Crits-Christoph et al., 2003; Elkin, 1999; Lee and Thompson, 2005; Pals et al., 2008; Roberts and Roberts, 2005; Serlin, Wampold, and Levin, 2003; Siemer and Joorman, 2003b; Walters, 2010). An example of the common rationale for needing to account for violations of independence of observations is provided by Pals et al. (2008), "Regardless of how it develops, any correlation within groups violates one of the major assumptions of statistical methods used in the analysis of randomized clinical trials…An assumption of these methods is that observations are independent within conditions.…" While there is some disagreement over the most appropriate approach for addressing the violation of independence of observations (fixed effects versus random effects, discussed in the next section), there appears to be broad agreement that there may be a correlation among the outcomes of individuals treated in the same group, and that ignoring this correlation could bias standard error estimators.

The sources of this lack of independence of observations have been described in various ways that can be summarized as falling into three main categories: (1) the process that sorts individuals into groups, (2) service provider effects, and (3) peer effects. Each source is discussed in turn.

The Processes That Sort Individuals into Groups

In many individually randomized experiments, after individuals are randomized to experimental arm, they sort into groups for service delivery. The sorting process may be controlled by the experimenter, but more commonly it is controlled by the service delivery organization (as in the case of students being assigned to teachers by a principal) or by the individuals themselves (as in the case of patients self-selecting their doctor or time slot for therapy). To the extent that differences exist between groups in terms of their individuals' characteristics, this may yield correlated future outcomes within groups.

A concrete example from a real-world experiment comes from an evaluation of learning communities in community colleges. In this evaluation, individuals were randomly assigned to the treatment (the opportunity to sign up for learning communities) or the control condition (business-as-usual services at the college) (Richburg-Hayes, Visher, and Bloom, 2008). After

random assignment, students who were offered the treatment self-selected the learning community blocks of courses that fit into their schedules. It is easy to imagine that different types of students would prefer different class time offerings. For example, individuals who work while in school (which is very common in community colleges) may prefer classes later in the day or in the evening, whereas nonworking students may prefer daytime classes. To the extent that working while in school is related to academic outcomes, the sorting process could yield within-group correlations in future outcomes.

Service Provider Effects

In individually randomized experiments, after individuals are randomized to experimental arms and sorted into groups, service providers may be a key influence on their clients' outcomes. Service providers can be educators (e.g., teachers or tutors), health professionals (e.g., doctors, nurses, or therapists), caseworkers, counselors, advisers, or group leaders. To the extent that service providers vary in their effectiveness, the outcomes of the individuals they serve will be correlated (i.e., there will be between-group variation in outcomes).

In addition to the intuitive and anecdotal evidence that service providers affect individuals' outcomes, there is convincing empirical evidence as well. The need to account for therapist effects in IRGT trials has been well documented in the psychotherapy literature (Crits-Christoph and Mintz, 1991; Crits-Christoph, Tu, and Gallop, 2003; Elkin, 1999; Siemer and Joorman, 2003a; Walters, 2010), and results from the education literature show that teachers account for an estimated 10 percent of the total variation in student test scores (Nye, Konstantopoulos, and Hedges, 2004).

Service providers may also be differentially effective at implementing a treatment or otherwise yield differential treatment effects for the individual they treat. This would also yield heterogeneity among the outcomes for individuals treated by different providers.

Peer Effects

A final potential source of lack of independence among observations of sample members who are treated in the same group is peer effects. Essentially, the outcomes of individuals within groups may be correlated due to the interactions that individuals within groups have with one another. In group therapy, for example, a particular patient may influence the outcomes of other patients. In education, peers may influence their classmates. When an individual's outcomes depend on the individuals assigned to the same experimental arm, the outcomes are said to contain partial interference and to violate the stable unit treatment value assumption (Rubin, 1980). When interference among outcomes exists, impact estimates from standard statistical procedures cannot be interpreted as causal effects without additional assumptions (Rubin, 1986;

Tchetgen and VanderWeele, 2012). In order to minimize complexity in our presentation, throughout the remainder of the paper we assume that peer effects do not exist except when otherwise explicitly discussed. The main problems we discuss are driven by individuals sorting into groups and exist even in the absence of peer effects. The problems are not generally going to go away in the more complicated case when there are true peer effects, so focusing on the case of no peer effects, as we do, is a useful simplification to make our main points clear.

In sum, there are at least three main sources of correlations of outcomes among those in the same group: (1) sorting, (2) providers, and (3) peers. In the past, sorting has *not* been viewed as a special nuisance in individually randomized experiments. Rather, it is typically bundled with the other sources of lack of independence and is accounted for through random or fixed effects as advocated by the literature. As will be demonstrated in Sections 2 through 5, this bundling can lead to bias in the estimated standard errors of the impact estimator from common models.

Section 2

Random Provider Effects

Many authors have argued for the use of hierarchical models with random group effects to account for the potential dependence among outcomes of individuals treated in the same group (Baldwin, Bauer, Stice, and Rohde, 2011; Bauer, Sterba, and Hallfors, 2008; Crits-Christoph and Mintz, 1991; Crits-Christoph, Tu, and Gallop, 2003; Lee and Thompson, 2005; Pals et al., 2008; Roberts and Roberts, 2005; Serlin, Wampold, and Levin, 2003). The standard random effects model used in this context is

$$Y_i = \beta_0 + \beta_1 T_i + \theta_{j[i]} + \varepsilon_i \qquad (1)$$

where Y_i is the outcome for individual i, T_i is a treatment indicator set equal to one if individual i is assigned to treatment and zero otherwise, $\theta_{j[i]}$ is the random effect for the group j to which individual i is sorted, and ε_i is an individual error term. The coefficient β_0 equals the control group mean outcome, and β_1 equals the average treatment effect or the impact. The group-level random effects (θ_j) are typically assumed to be normally distributed with mean zero and constant variance, τ^2, and the residual errors (ε_i) are also assumed to be normally distributed with mean zero and constant variance, ω^2. The group-level random effects are assumed to be independent across groups and independent of the individual-level error terms, which are assumed to be independent across individuals. The group-level random effects provide information about heterogeneity in the mean outcomes among groups. The proportion of the total variance in the outcome that is between groups, $\tau^2/(\tau^2 + \omega^2)$, is commonly referred to as the intraclass correlation, or ICC (Kish, 1965).

The papers advocating the use of random effects models have shown that such models can provide consistent estimates of the standard error of the maximum likelihood estimates of the impact of the intervention. Of course, the consistency of the standard errors is contingent on the assumptions of the hierarchical linear model being correct. As has been noted in the literature, one of those assumptions is that the provider effects are a random sample from a population. In many experiments, the providers are not a random sample from a well-defined identifiable population, and some authors argue that the random effects model does not hold (Siemer and Joorman, 2003a, 2003b). Other authors counter that the random effects model can be justified, even if the providers are not truly a random sample, by random assignment of providers to either treatment or control or by treating the providers as a random sample from a hypothetical super-population (Crits-Christoph, Tu, and Gallop, 2003; Serlin, Wampold, and Levin, 2003). For now, we will assume that treating the providers as a random sample is justifiable so that we can focus on standard error estimates.

When deriving results about standard errors, the literature advocating the use of random effects models does not differentiate among the three sources of the clustering in outcomes (sorting of individuals to providers, provider effects, and peer effects). The random effects model in Equation 1 contains only a single random effect for each group with no specification of its source. Consequently, the random effects model is incompletely specified and treats all sources of between-group heterogeneity as a common random variable associated with each provider. This would be appropriate under two distinct settings. The first is if individuals are randomly assigned to arms, and then randomly assigned to providers within each arm. This would make the individual-level error terms mutually independent and independent of any provider effects so that the assumptions of the random affects model are not violated. The second setting where the random effects model would be appropriate is in a cluster-randomized design in which individual unit assignments are determined by the provider assignments. However, IRGT studies typically do not fall into either of these two settings. In most instances, there would be multiple potential providers for each individual conditional on that individual's assigned arm, but which provider each individual was assigned to might not be random because it might depend on attributes of the provider, the individual, and possibly the other individuals assigned to the same arm. Consequently, the random effects model is generally misspecified for IRGT designs.

A simplified example demonstrates how the two-stage sample differs from a clustered sample and how the random effects model can overestimate the standard error of an impact estimate. For this example, we assume no treatment effects and no provider effects (in other words, all providers have the same effect, which is zero). Thus, for every individual, there is a single outcome, which we will call Y_i, which is identical regardless of whether the individual is assigned to treatment or control and regardless of the group into which he or she sorts. We further assume that a sample of *n* individuals is randomly selected from a very large population with constant variance (σ^2) for the treatment group, and a separate sample of the same size n is selected from the same population for control. For the sample in each arm, an equal number c of individuals are assigned to each provider *on the basis of their outcome values*, so that the means of the Y_is of the individuals assigned to different groups are heterogeneous.

The standard estimator for the coefficients from the random effects model, $\beta = (\beta_0, \beta_1)'$, is $\hat{\beta}_{RE} = (\mathrm{X}'\mathrm{V}^{-1}\mathrm{X})^{-1}\mathrm{X}'\mathrm{V}^{-1}\mathrm{Y}$, where Y is the vector individual outcomes, X is the design matrix containing a column of ones for the intercept and a column for the treatment indicators, and V is the model-based estimate of the variance-covariance matrix for Y. Assuming that the data are sorted by group, V is block diagonal, where the off-diagonal elements within the blocks equal the intraclass correlation (ICC) of individuals sharing a group times the within-arm variance in the outcomes. Maximum likelihood or restricted maximum likelihood is used to estimate the parameters of V. Straightforward derivations yield that for equally sized groups of size c, $\mathrm{X}'\mathrm{V}^{-1}\mathrm{X} = \mathrm{X}'\mathrm{X}a$ and $\mathrm{X}'\mathrm{V}^{-1}\mathrm{Y} = \mathrm{X}'\mathrm{Y}a$ where $a = v^2\left[1 + (c-1)\widehat{ICC}\right]^{-1}$,

and v^2 equals the pooled within-arm variance. In large samples, v^2 will converge to σ^2, and the estimated ICC will converge to the true ICC so that the difference between the true and estimated V can be ignored for the purposes of our discussion.

The simple Ordinary Least Squares (OLS) model ignores the clustering of individuals in groups, resulting in

$$Y_i = \beta_0 + \beta_1 T_i + \varepsilon_i \quad (2)$$

where the ε_i are assumed to be independent across individuals. The resulting estimator is $\hat{\beta}_{OLS} = (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{Y}$ (Searle, 1971). In this simple case, the estimate of β_1 equals the difference between the treatment and control arm means.

Regardless of the sample size, when the groups are equally sized, $\hat{\beta}_{RE} = (\mathrm{X}'\mathrm{V}^{-1}\mathrm{X})^{-1}\mathrm{X}'\mathrm{V}^{-1}\mathrm{Y} = [(\mathrm{X}'\mathrm{X})^{-1}/a]\mathrm{X}'\mathrm{Y}a = (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{Y}$, the OLS estimator. The estimated treatment effect from fitting the random effects model will simply be the difference between the treatment sample mean and the control sample mean. This is true regardless of how individuals sort to groups. In this example, the assignment of individuals to groups has no bearing on their outcomes, because providers have no effects. No matter how the individuals are assigned to providers, the arm-level means remain constant. The mean for each arm is simply the mean of the simple random sample of size n. Hence, across repeated experiments the variance in the mean of the treatment arm equals the variance, σ^2, of the Y_i divided by n, and the same is true for the control arm. The samples in each arm are independent and, therefore, the impact estimator from the random effects model, which we label $\hat{\beta}_{1RE}$, has true variance $2\sigma^2/n$.

However, under the random effects model, the estimated variance of $\hat{\beta}_{1RE}$ equals the corresponding diagonal element of $(\mathrm{X}'\mathrm{V}^{-1}\mathrm{X})^{-1}$, which from above equals $(\mathrm{X}'\mathrm{X})^{-1}/a$. This is the OLS variance-covariance matrix times $1 + (c-1)ICC$. That is, in large samples the variance of $\hat{\beta}_{1RE}$ reported from the random effect model is equal to $(2\sigma^2/n)(1 + (c-1)ICC)$, even though the true variance of $\hat{\beta}_{1RE}$ is only $2\sigma^2/n$. Therefore, as long as the ICC is positive, the standard error estimator from the random effects model will be positively biased, that is, $SE(\hat{\beta}_{1RE}) < E[\widehat{SE}(\hat{\beta}_{1RE})]$.

The intuition of this simple example should apply to more complex applications with treatment, provider, and peer effects. In those contexts, the potential outcomes for each individual can be modeled as the sum of the individual-specific component that depends on the individual but not on treatment or control assignment, a provider-specific term that applies to all individuals grouped with that provider, a peer effect term, and interactions among these terms. Heterogeneity across providers (within treatment arm) because of nonrandom assignment to groups does not contribute to the variance of the impact estimator. However, the random effects

model cannot distinguish this source of between-group heterogeneity from other sources of correlation in outcomes among individuals assigned to the same provider. So, generally, it will not recover an estimate of the standard error that corresponds to the actual variability of the impact estimator across repeated realizations of the assignment processes.

If individuals are randomly assigned to providers, then there is not heterogeneity from the second-stage assignment and the standard error estimator from the random effects model will be consistent. Hence, by ignoring the particular sources of heterogeneity among groups, the literature on the use of random effects for IRGT studies implicitly assumes that individuals are randomly assigned to providers.

Generalized estimating equations with cluster-adjusted standard errors are sometimes used as an alternative to random effects models for IRGT experiments (for example, see Visher et al., 2012). However, these methods also do not distinguish among the sources of heterogeneity among providers and will yield inflated standard errors for the same reasons that random effects models do.

Section 3

Fixed Provider Effects

As noted in the previous section, the random effects model assumes that providers are randomly sampled from a larger population. This often is not the case, and some authors have argued that a more appropriate model would treat providers as fixed (Siemer and Joorman, 2003a, 2003b). The standard model used in this case is

$$Y_i = \beta_0 + \beta_1 T_i + \varphi_{j[i]} + \varepsilon_i \qquad (3)$$

where variable and coefficients have the same definitions as in Equation 1, except that $\varphi_{j[i]}$ is the fixed effect for the group where individual i received treatment. Models with fixed provider effects and treatment effects are not identified, so the fixed effects models must assume that the fixed provider effects sum to zero within each experimental arm (Siemer and Joorman, 2003b).

As was the case in the papers advocating the use of random effects, the papers promoting fixed effects do not consider the source of heterogeneity among outcomes from groups of individuals served by different providers. Analogous to the random effects case, the assignment of individuals to providers does not contribute to the true standard error of the fixed effects impact estimate. As discussed in the previous section, the individual-specific components of the outcomes are fixed after assignment to experimental arm. This is true whether we consider the providers as fixed or random.

Again, the simple example of equal sample sizes per provider and no provider, no treatment, and no peer effects will make this clear. The fixed effects estimator of model coefficients, the mean, impact, and group effects, $\hat{\beta}_{FE} = (\mathrm{M}'\mathrm{M})^{-1}\mathrm{M}'\mathrm{Y}$, where M**,** the design matrix, equals the OLS design matrix, X, with columns appended for the group fixed effects, $\mathrm{M} = [\mathrm{XB}]$. The matrix B contains $n/c - 1$ columns corresponding to the groups in each arm. Each column includes 1s in the elements corresponding to the individuals in the group, -1s in the elements corresponding to individuals in the holdout group for the arm, and zeros elsewhere. By construction, $\mathrm{X}'\mathrm{B} = 0$ so $\mathrm{M}'\mathrm{M}$ is block diagonal with blocks $\mathrm{X}'\mathrm{X}$ and $\mathrm{B}'\mathrm{B}$ so that $\hat{\beta}_{FE} = [\{(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{Y}\}', \{(\mathrm{B}'\mathrm{B})^{-1}\mathrm{B}'\mathrm{Y}\}']'$. The impact estimator from the fixed effects model again equals the OLS estimator, which in this simple case equals the difference of the treatment and control arm means, and the square of its standard error equals $2\,\sigma^2/n$.

The estimated standard error for the intercept and impact estimators will be $(\mathrm{X}'\mathrm{X})^{-1}MSE^2$, where MSE^2 equals the sum of squared residuals from the fixed effect model divided by $2n - 2n/c$. The sum of squared residuals from this model equals the within sum of squares from a one-way ANOVA for groups, so MSE^2 equals the mean squared error within.

When n/c and c are large, it can be shown that $1 - \widehat{ICC}$ is approximately equal to the mean squared error within, divided by the OLS mean squared error (i.e., the sum of squared OLS residuals divided by $2n - 2$).[2] Therefore, $(\mathrm{X}'\mathrm{X})^{-1}MSE^2$ is approximately equal to the OLS variance-covariance matrix times $(1 - ICC)$, and the square of the standard error from the fixed effects model underestimates the variance in the estimator by a factor of $(1 - ICC)$. If there are groups only in the treatment arm, the variance is underestimated by a factor of $(1 - ICC)/2$. If individuals are nonrandomly assigned to providers so that individual outcomes are correlated within provider, the expected value of the $ICC > 0$ and $SE(\hat{\beta}_{1FE}) > E\left[\widehat{SE}(\hat{\beta}_{1FE})\right]$. The square of the reported standard error of the fixed effects estimator will underestimate the true variability of the impact estimator. Even with true provider effects, the sorting of individuals into groups does not affect their contribution to the variance of the impact estimator, and its standard error from the fixed effects model will be biased low.

Even if the providers are fixed, any existing peer effects will be random, and peer effects inflate the variance of the impact estimator relative to a model with independent errors within providers. However, the fixed effects estimator does not discriminate among sources of variance at the provider level and removes them all from the standard error estimator. Consequently, the presence of peer effects also results in the fixed effects estimator underestimating the standard error of the impact estimator.

If we consider individuals as fixed so that the only random variable is treatment assignment (Rubin, 1990), the problem remains: The grouping of individuals does not add to the variability of the impact estimator other than through provider and peer effects, but the fixed effects standard error estimator also removes heterogeneity due to the sorting of individuals into groups.

We have been assuming that the assignment of individuals to providers within arm would vary across the potential assignments of the samples to experimental arms. That is, individuals might be assigned to any provider in their experimental arm. An alternative assumption is that each individual in the super-population or a finite sample would be associated with one treatment provider. He or she would be assigned to only that provider if assigned to treatment, and to only one control provider if assigned to control. Different individuals would be assigned to different providers, but each individual would be assigned to the same treatment (control) provider for every realization of the random assignments in which he or she was assigned to treatment (control). Under this assumption, the only probabilistic assignment would be to experimental arm; once assigned to experimental arm, an individual's assignment to provider would be predetermined and fixed. Effectively, individuals would be stratified by the provider they

[2] An ICC is not typically calculated in a fixed effects model but can be approximated by the adjusted r^2 or can be calculated from mean squares values from the fitted model (Snedecor and Cochran, 1989).

were assigned to under treatment or control. In this case, the fixed effects standard error would be correct. The variation in the means of the individual assigned to providers is fixed, so it cannot contribute to variation in the impact estimator.

Given that only one assignment will be used in a study, we cannot determine whether individuals are affixed with providers or whether they have the potential to be assigned to any provider, depending on the other individuals assigned to each experimental arm. However, information on how individuals are chosen for the provider groups may help determine whether one model is more plausible than the other. For example, if slots available at providers are fixed, and there are decisions about how best to place the individuals, given the entire sample, then individual assignments are probably not fixed with providers. Alternatively, if provider assignments are based solely on the characteristics of each individual, with no consideration of the other individuals in the experimental arm, then assuming that assignments are fixed may be plausible. This could be a reasonable assumption, for example, in cases where providers are spread out geographically and units are likely to go to the nearest provider.

Section 4

Simulation Study

The Data-Generating Mechanisms

We conducted a simulation study to demonstrate that under data-generating mechanisms (DGM) that correspond to IRGT designs, the sampling distribution of the impact estimator is not modeled accurately by the OLS, random effects, or fixed effects models under many realistic DGM. A consequence of the model misspecification is bias in the standard error estimates from each of the models under some of the DGM.

We consider six DGM, summarized in Table 1. In all DGM, individuals are randomly assigned to experimental arm, and there is no treatment effect. In DGM (1), the level-one units are randomly assigned to equally sized groups within arm, and there are no provider effects and no peer effects. DGM (2) is identical to DGM (1), except that in DGM (2), level-one units are placed into groups based on their potential outcomes, such that around 10 percent of the overall variation in outcomes is between groups. That is, the ICC is $.10$. Individuals are still randomly assigned to experimental arm. DGM (3) is the same as DGM (1), except that (a) providers vary in their effectiveness,[3] (b) treatment and control group providers are randomly sampled from a population of providers, and (c) the desired causal inference is assumed to be to the super-population of providers. We set the provider effect variance so that when individuals were randomly assigned to providers, the provider ICC was .10, roughly corresponding to the apparent magnitude of teacher effects on student achievement (Nye, Konstantopoulos, and Hedges, 2004).

DGM (4) is like DGM (3), except that the sample of providers is fixed across all the simulated datasets. Initially, one set of providers was drawn from the same population of providers as in DGM (3); this unique set of providers was retained across all data sets, and each provider always remained in the same experimental arm. The provider effects were constrained to sum to zero within each arm, consistent with the necessary identifying assumptions of the fixed provider effects model. In this case, the desired causal inference is with respect to the particular sample of providers observed, not to a super-population of providers from which the sample was drawn.

DGM (5) and (6) mirror DGM (3) and (4), respectively, except that they allow level-one units to sort into groups nonrandomly, such that prior to experiencing provider effects, individuals' potential outcomes are related to group assignment. The assignment of providers to groups of individuals was independent of the group means of the potential outcomes. We discuss the implications of relaxing this assumption in the Discussion section.

[3]Each individual provider is assumed to have homogenous effects on level-one units.

Table 1. Data-Generating Mechanisms (DGM) for the Simulation Study

DGM	Sorting into Experimental Arm	True Treatment Effect	Sorting into Groups	Provider Effects	Provider Sampling
1	Random	Zero	Random	Zero	-
2	Random	Zero	Nonrandom	Zero	-
3	Random	Zero	Random	Vary	Random
4	Random	Zero	Random	Vary	Fixed
5	Random	Zero	Nonrandom	Vary	Random
6	Random	Zero	Nonrandom	Vary	Fixed

NOTES: All DGMs draw a random sample of 10,000 individuals from a population. Half are randomly assigned to the treatment arm and the other half to the control arm. In all DGM there is absolutely no treatment effect. All individuals are assigned to groups of size 25 within their experimental arm. Group assignment is either randomly or through a process that is correlated with individual's average potential outcome. In scenarios where provider effects exist, they are added on after individuals sort into groups. When provider sampling is random, a unique set of providers is randomly drawn from a population of providers for each simulated dataset. When provider sampling is fixed, a fixed set of treatment arm providers and a fixed set of control arm providers are randomly assigned to groups within their respective experimental arm. The fixed set of providers (within each arm) is the same for each simulated dataset. The fixed set of providers used in the fixed sampling case is initially randomly drawn from the same population of providers as is used in the random provider sampling case.

We generated 10,000 datasets for each DGM. For each dataset, we first drew a random sample of $2n = 10{,}000$ individuals from a population. In all scenarios, there is no treatment effect. Because we assume no true treatment effects, the treatment and control potential outcomes for each unit are equal, and these values were randomly generated from a standard normal distribution. Next, we randomly assigned individuals to either a treatment arm or a control arm. We then assigned samples of $c = 25$ individuals to 200 groups within each arm, according to the specification of the DGM. Group sizes of 25 are consistent with typical classroom sizes, and the total number of units is roughly consistent with what might be available in a cohort from a large urban school district. For DGM (3) and (5), we generated random samples of providers for each arm and added those effects to the individual data. Similarly for each dataset from DGM (4) and (6), we added the fixed provider effects to the individual data for each arm.

We estimated the impact and its standard error for each dataset using an OLS regression (Equation 2), a random effects model (Equation 1), and a fixed effects model (Equation 3), even though the data from most of the DGM do not meet the assumptions of these estimation models. We also used generalized estimating equations with a working independence covariance matrix

and cluster-adjusted standard errors. Because the results from this model were nearly identical to those from the random effects model, they are not reported. For each model, the parameter of interest is the standard error of β_1 and the test of the null hypothesis that $\beta_1 = 0$.

For each DGM and model, we estimated the expected value of the impact estimate and its reported standard error by their respective means across the 10,000 samples. We estimated the true standard error by the sample standard deviation of the impact estimate across the 10,000 samples. We compared the estimated expected value of the standard error with our estimate of the true standard error to calculate the bias in the standard error estimator. SAS and R code to replicate the simulation studies are available from the authors.

Section 5

Results of the Simulation Study

The simulation results confirm what is expected theoretically. Table 2 presents results from the simulations. Table 2, the first row of panel 1, labeled the "truth estimate," presents the estimates of the true standard error of the impact estimator from each model. In this case, each impact estimator is the difference in the mean outcomes from the two experimental arms. Thus the true standard error (the estimand) is the same across models and equals .0202 (i.e., $SE(\hat{\beta}_{1OLS}) = SE(\hat{\beta}_{1FE}) = SE(\hat{\beta}_{1RE}) \approx .0202$). The second row in panel 1, labeled "mean of 10,000 SE," is the simulation-based estimate of the expected value of the standard error from the OLS model, the fixed effects model, and the random effects model. If the standard error of the impact estimate from a model accurately reflects the uncertainty in the impact estimate, the truth estimate and the mean of 10,000 SEs should be very close. The next two rows, the estimated bias and the estimated percent bias, compare the first two rows. The estimated bias is the difference between the truth estimate and the mean standard error, and the estimated percent bias is the estimated bias divided by the truth estimate. Finally, the last row presents the percent of times that the null hypothesis of no treatment effect was rejected. Since all DGMs have no treatment effect, the target value for the bottom row is 5 percent. Under the simple DGM (1), all impact models result in negligible bias and rejection rates that are acceptably close to 5 percent. Note that in this case the only between-group variation in mean outcomes is a result of random sampling variability, so it is unsurprising that including fixed or random effects in our model is not consequential.

Panel 2 in Table 2 presents the results from DGM (2). Notice that the first row in panel 2 is the same (in practical terms) as the first row in panel 1 — the standard deviation of 10,000 impact estimates is .0200 (again, $SE(\hat{\beta}_{1OLS}) = SE(\hat{\beta}_{1FE}) = SE(\hat{\beta}_{1RE}) = .0200$). This occurs because there is no additional source of uncertainty in the impact estimate that results from nonrandom sorting into groups. Consequently, the OLS standard error, which ignores grouping, is estimated without bias (that is, $SE(\hat{\beta}_{1OLS}) = \widehat{SE}(\hat{\beta}_{1OLS})$).

In comparison, the average standard error from the fixed effect regression is .0190, which is a 5.1 percent underestimate, or as the theoretical results predicted, about $\sqrt{1 - ICC}$ times the true value (that is, $SE(\hat{\beta}_{1FE}) \approx \widehat{SE}(\hat{\beta}_{1FE})\sqrt{1 - \widehat{ICC}}$). In contrast, the average standard error from the random effect regression is .0369, which is 84.7 percent too large or $\sqrt{1 + ICC \times 24}$ times the truth as predicted by the theoretical results, since $c = 25$ (that is, $SE(\hat{\beta}_{1RE}) \approx \widehat{SE}(\hat{\beta}_{1RE})\sqrt{1 + \widehat{ICC} \times 24}$). Consequently, even though there is absolutely no treatment effect, the null hypothesis is rejected at the .05 level 6.3 percent of the time using the fixed effects model and 0.0 percent of the time using the random effects model. This simplified

Table 2. Bias of the Standard Error of the Impact Estimator (True Impact = 0)

Data-Generating Mechanism	OLS	Fixed Effects	Random Effects
(1) Random sorting, no provider effect (ICC = 0.00)			
Truth estimate (sd of 10,000 impact estimates)	0.0202	0.0202	0.0202
Mean of 10,000 SE	0.0200	0.0200	0.0203
Estimated bias	-0.0002	-0.0002	0.0001
Estimated percent bias	-0.9%	-0.9%	0.4%
P(reject null)	5.4%	5.4%	5.0%
(2) Nonrandom sorting, no provider effect (ICC = 0.10)			
Truth estimate (sd of 10,000 impact estimates)	0.0200	0.0200	0.0200
Mean of 10,000 SE	0.0200	0.0190	0.0369
Estimated bias	0.0000	-0.0010	0.0169
Estimated percent bias	0.0%	-5.1%	84.7%
P(reject null)	4.9%	6.3%	0.0%
(3) Random sorting, random provider effect (ICC = 0.10)			
Truth estimate (sd of 10,000 impact estimates)	0.0388	0.0388	0.0388
Mean of 10,000 SE	0.0211	0.0200	0.0388
Estimated bias	-0.0178	-0.0188	0.0000
Estimated percent bias	-45.7%	-48.5%	0.0%
P(reject null)	28.6%	31.1%	4.9%
(4) Random sorting, fixed provider effect (ICC = 0.10)			
Truth estimate (sd of 10,000 impact estimates)	0.0201	0.0201	0.0201
Mean of 10,000 SE	0.0211	0.0200	0.0391
Estimated bias	0.0010	-0.0001	0.0190
Estimated percent bias	5.0%	-0.5%	94.6%
P(reject null)	3.9%	5.0%	0.0%
(5) Nonrandom sorting, random provider effect (ICC = 0.19)			
Truth estimate (sd of 10,000 impact estimates)	0.0388	0.0388	0.0388
Mean of 10,000 SE	0.0211	0.0190	0.0497
Estimated bias	-0.0177	-0.0198	0.0109
Estimated percent bias	-45.6%	-51.1%	28.2%
P(reject null)	28.4%	33.6%	1.3%
(6) Nonrandom sorting, fixed provider effect (ICC = 0.19)			
Truth estimate (sd of 10,000 impact estimates)	0.0199	0.0199	0.0199
Mean of 10,000 SE	0.0211	0.0190	0.0499
Estimated bias	0.0012	-0.0009	0.0300
Estimated percent bias	5.9%	-4.7%	150.7%
P(reject null)	3.7%	6.1%	0.0%

NOTES: The ICC in DGM (1) is zero because there is no sorting and no provider effects. The ICCs in DGM (2-4) are 0.10 because each contains either nonrandom sorting or true provider effects, but not both. The ICCs in DGM (5-6) of 0.19 reflect the combination of both sorting and provider effects.

case demonstrates how the nonrandom sorting of level-one units into groups can lead to bias in the standard error of the impact estimator for both the fixed and random effects models, because neither model accurately captures the IRGT design in which the sample of randomly assigned individuals is then sorted into groups.

The inclusion of provider effects in DGM (3) inflates the true standard error of the impact estimator to 0.0388 (panel 3, row 1). Using OLS, which ignores provider effects, the average standard error (.0211) is greatly underestimated (by an amount approximately equal to the reciprocal of $\sqrt{1 + ICC \times 24}$, as it fails to account for the additional uncertainty caused by the sampling variability associated with providers. Unsurprisingly, the average standard error of the fixed effects estimator (.0200) is also underestimated — providers are not fixed. In this case, this is a result of the desired inference being misaligned with the estimand. The average standard error of the random effects estimator (.0388) is unbiased, because in this case the between-group heterogeneity in outcomes depends on the providers, so that the random effects model approximates the DGM.

When provider effects are fixed as in DGM (4), they do not contribute to the variability in the impact estimator. The true standard error is 0.0201 (panel 4, row 1). As expected, the standard error from the fixed effect estimator (0.0200) is now unbiased, as it removes the provider effects from the analyses through the constrained fixed effects. The OLS standard error estimator is biased upward by a factor of $1/\sqrt{1 - ICC}$. The standard error of the random effect estimator (0.0388) is biased upward also because the desired causal inference is misaligned with the estimand. DGM (3) and (4) are useful demonstrations that if units are randomly assigned to groups of equal sizes, the random and fixed effects models produce standard errors of the impact estimator that are unbiased when they align with the desired inference.

Because DGM (5) and (6) combine provider effects and nonrandom assignment to groups, none of the models yields unbiased standard errors, and the biases are all in the expected directions and are roughly the size predicted by the theoretical results. In these most realistic cases, even with moderate amounts of clustering and modest provider heterogeneity, the bias in any of the estimators can be quite large in magnitude.

Section 6

Discussion and Conclusions

In individually randomized trials where units sort nonrandomly into treatment delivery groups, potentially causing correlation in outcomes for individuals sharing groups, both random and fixed effects modeling approaches for addressing this correlation lead to biased standard error estimators. The random effects model overestimates the standard error of the impact estimator, because it assumes that variability among groups arising from the nonrandom sorting is instead arising from an additive, group-level source of heterogeneity such as provider effects. The fixed effects model, on the other hand, underestimates the standard error of the impact estimator because it uses within-group variability, dampened by the nonrandom sorting, as the basis of the standard error. In a simple case where the only source of correlation in outcomes for individuals sharing groups is due to sorting, OLS provides a correct standard error, regardless of the sorting mechanism. If, on the other hand, sorting to groups was random, but there were true provider effects, both the random effects and fixed effects would provide correct standard errors for their respective inferences. However, when both nonrandom sorting and true provider effects contribute to correlation among outcomes from individuals sharing groups, none of the modeling approaches generally provides correct standard error estimates. IRGT designs that do not experimentally control for group assignment are at risk for both sources of heterogeneity and biased standard errors.

The situation does not improve if the data-generating model is more complicated than the simple case with constant group sizes and no relationship between provider effects and the assignment of individuals to groups. For example, groups could be unequally sized, possibly with the sizes of the groups related to attributes of the individuals (e.g., healthy patients or higher-achieving students are assigned to big groups). Alternatively, unequal group sizes might be related to the provider effects (e.g., more individuals are assigned to more effective providers), or provider effects may be related to the attributes of the individuals assigned to them (e.g., struggling students are assigned to more effective teachers). In general, the behavior of an impact estimator and its associated standard error estimators depends on the joint distribution of the provider effects, the group sizes, and the group means of the individual average potential outcomes, where the individual average potential outcome equals the average of an individual's separate potential outcome for every provider under treatment and control. Data-generating models that allow correlations in the joint distribution of the group sizes, group means of individual average potential outcomes, and provider effects tend to violate one or more assumptions made by the OLS, random effects, and fixed effects models. These violations of the model assumptions affect both the magnitude and direction of biases in the reported standard errors.

In this paper, we focus only on bias in the standard error estimator, but violations of the assumptions of the model because of unequal sample sizes and nonrandom assignment of individuals to groups can also result in bias in the impact estimator (Weiss, Visher, and Weissman, 2012). Users of IRGT designs should be mindful of potential threats to inferences about the impacts of their interventions.

Practical Recommendations for Moving Forward

To the best of the authors' knowledge, some of the issues raised in this article are intractable. In particular, the standard error estimators go wrong in predictable ways under even grossly simplified data-generating models and will not, in general, improve under more realistic scenarios. Nonetheless, many evaluations involve random assignment of individuals to two or more treatment arms, and the individuals in one or more of those arms end up in groups or clusters. Since this design has the benefit of removing individual-level selection bias on impact estimators, it is important to consider the options available to researchers so that the properties of the standard error of the impact estimator are understood and their limitations appreciated. We offer the following practical suggestions.

Randomly Assign Individuals to Groups

The bias in the standard errors results from the nonrandom assignment of individuals to groups. As demonstrated in our simulation, if individuals were randomly assigned to groups, the common statistical models would yield consistent standard errors for their impact estimators. Random assignment of individuals to groups also has benefits for the impact estimator. Consequently, random assignment of individuals should be explored as part of the study design. In some instances, random assignment might be impossible because geography or other constraints restrict each individual to being associated with a single provider in each arm. In these cases, the standard error for the fixed effects estimator would be unbiased if that impact estimator is appropriate. In other cases, researchers will not want to, or be able to control assignments, but they are not fixed. In these cases, another method to reduce bias will be necessary.

Include Covariates

The potential benefits of including covariates in an impact model are well understood for the purpose of improving the precision of the impact estimator (Bloom, Richburg-Hayes, and Black, 2007). Our findings suggest that the inclusion of covariates may serve another important role — accounting for the process that sorts individuals into groups. Under the assumption of strongly ignorable *group* assignment (e.g., DGM (3) and (4)), the standard error of the impact estimator is unbiased (when the model aligns with the inference). To the extent that group membership is ignorable after controlling for covariates, the nuisance that sorting causes

can be eliminated. Even if the effects of sorting cannot be eliminated entirely, reducing the variance in group means due to sorting tends to be beneficial. It makes more of the residual heterogeneity in group means result from sources of variance that are treated approximately correctly by the standard models. For example, if after adjusting for covariates, the vast majority of between-group variance is due to provider effects, the standard random or fixed effects estimators will tend to have reported standard errors that are only mildly inconsistent for their respective targets. In particular, for the random effects model, if nonrandom sorting causes between-group variability to exceed the true providers variance τ^2, including covariates may reduce the residual variability between groups to something closer to τ^2, which will, in turn, reduce the bias in the estimated standard errors.

The most obvious choice of covariates is a baseline measure of the target outcome. Group means of baseline measures are another source of potential covariates that might remove clustering in the observed outcomes.

Another important role of covariates (particularly a baseline measure of the target outcome) in this setting is that a variance decomposition of a pretreatment covariate into between- and within-group components will be informative about the degree of sorting. Between-group variance on a pretreatment assignment (and therefore pregrouping) covariate is attributable to sorting. Thus it provides useful information about the potential bias in reported standard errors from different approaches, when not adjusting for the covariate. In the event that units do not appear to be sorted to groups on the basis of the baseline measure, analysts can be more confident that the standard approaches are likely to yield reasonable standard error estimates. On the other hand, a nonnegligible between-group variance component of the covariate can motivate the need to adjust for the covariate to mitigate the problems discussed here.

References

Abdulkadiroglu, A., Angrist, J., Dynarski, S., Kane, T. J., and Pathak, P. (2009). Accountability and flexibility in public schools: Evidence from Boston's charters and pilots. NBER Working Paper No. 15549. Cambridge, MA: National Bureau of Economic Research. Web site: http://www.nber.org.

Baldwin, S. A., Bauer, D. J., Stice, E., and Rohde, P. (2011). Evaluating models for partially clustered designs. *Psychological Methods, 16*(2), 149-165.

Bauer, D. J., Sterba, S. K., and Hallfors, D. D. (2008). Evaluating group-based interventions when control participants are ungrouped. *Multivariate Behavioral Research, 43*(2), 210-236.

Bernstein, L., Yamaguchi, R., Unlu, F., Edmunds, J., Glennie, E., Willse, J., . . . Dallas, A. (2010). Early findings from the implementation and impact study of early college high school. Paper presented at the Society for Research on Educational Excellence, Washington, DC.

Bloom, H. S., Richburg-Hayes, L., and Black, A. R. (2007). Using covariates to improve precision for studies that randomize schools to evaluate educational interventions. *Educational Evaluation and Policy Analysis, 29*(1), 30-59.

Bloom, H. S., Thompson, S. L., and Unterman, R. (2010). *Transforming the high school experience: How New York City's new small schools are boosting student achievement and graduation rates*. New York: MDRC.

Calcagno, J. C., & Long, B. T. (2008). *The impact of postsecondary remediation using a regression discontinuity approach: Addressing endogenous sorting and noncompliance.* An NCPR Working Paper. New York: National Center for Postsecondary Research, Teachers College, Columbia University.

Corrin, W., Somers, M.-A., Kemple, J. J., Nelson, E., and Sepanik, S. (2008). *The Enhanced Reading Opportunities study: Findings from the second year of complementation.* Executive summary. NCEE 2009-4037. Washington, DC: National Center for Education Evaluation and Regional Assistance. Institute of Education Sciences, U.S. Department of Education. Web site: http://ies.ed.gov/ncee/pubs/.

Crits-Christoph, P., and Mintz, J. (1991). Implications of therapist effects for the design and analysis of comparative studies of psychotherapies. *Journal of Consulting and Clinical Psychology, 59*(1), 20-26.

Crits-Christoph, P., Tu, X., and Gallop, R. (2003). Therapists as fixed versus random effects-some statistical and conceptual issues: A comment on Siemer and Joormann (2003). *Psychological Methods, 8*(4), 518-523.

Decker, P. T., Mayer, D. P., and Glazerman, S. (2004). The effects of Teach For America on students: Findings from a national evaluation. Discussion paper no. 1285-04. Madison, WI: Institute for Research on Poverty.

Dynarski, M., and Gleason, P. (2002). How can we help? What we have learned from recent federal dropout prevention evaluations. *Journal of Education for Students Placed at Risk (JESPAR), 7*(1), 43-69.

Elkin, I. (1999). A major dilemma in psychotherapy outcome research: Disentangling therapists from therapies. *Clinical Psychology: Science and Practice, 6*(1), 10-32. doi: 10.1093/clipsy.6.1.10.

Hoxby, C. M., and Murarka, S. (2009). Charter schools in New York City: Who enrolls and how they affect their students' achievement. NBER Working Paper No. 14852. Cambridge, MA: National Bureau of Economic Research. Web site: http://www.nber.org/cgi-bin /get_bars.pl?bar=pub.

Kemple, J. J., Herlihy, C. M., and Smith, T. J. (2005). *Making progress toward graduation: Evidence from the Talent Development high school model.* New York: MDRC.

Kish, L. (1965). *Survey sampling*. New York: John Wiley.

Lang, L., Torgesen, J., Vogel, W., Chanter, C., Lefsky, E., and Petscher, Y. (2009). Exploring the Relative Effectiveness of Reading Interventions for High School Students. *Journal of Research on Educational Effectiveness, 2*(2), 149 - 175.

Lee, K. J., and Thompson, S. G. (2005). The use of random effects models to allow for clustering in indvidually randomized trials. *Clinical Trials, 2*, 163-173.

Love, J. M., Kisker, E. E., Ross, C. M., Schochet, P. Z., Brooks-Gunn, J., Paulsell, D., . . . Brady-Smith, C. (2002*). Making a difference in the lives of infants and toddlers and their families: The impacts of early Head Start.* Volumes I-III*: Final technical report* [and] *appendixes* [and] *local contributions to understanding the programs and their impacts.* Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services. Web site: http://www.acf.dhhs.gov/ programs/core/ongoing_research/ehs/ehs_intro.html.

Michalopoulos, C. (2005). Precedents and prospects for randomized experiments. In H. Bloom (Ed.), *Learning more from social experiments: Evolving analytic approaches*. New York: Russell Sage Foundation.

Nye, B., Konstantopoulos, S., and Hedges, L. V. (2004). How large are teacher effects? *Educational Evaluation and Policy Analysis, 26*(3), 237-257.

Pals, S. L., Murray, D. M., Alfano, C. M., Shadish, W. R., Hannan, P. J., and Baker, W. L. (2008). Individually randomized group treatment trials: A critical appraisal of frequently used design and analytic approaches. *American Journal of Public Health, 98*(8), 1418-1424. doi: 10.2105/ajph.2007.127027

Richburg-Hayes, L., Visher, M. G., and Bloom, D. (2008). Do learning communities effect academic outcomes? Evidence from an experiment in a community college. *Journal of Research on Educational Effectiveness, 1*(1), 33-65.

Roberts, C., and Roberts, S. A. (2005). Design and analysis of clinical trials with clustering effects due to treatment. *Clinical Trials, 2*(2), 152-162. doi: 10.1191/1740774505cn076oa.

Rubin, D. B. (1980). Discussion of "randomization analysis of experimental data in Fisher randomization test," by Basu. *Journal of the American Statistical Association 75*, 591-593.

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

Rubin, D. B. (1990). Comment: Neyman (1923) and causal inference in experiments and observational studies. *Statistical Science, 5*(4), 472-480.

Scrivener, S., and Weiss, M. J. (2009). *More guidance, better results? Three-year effects of an enhanced student services program at two community colleges.* New York: MDRC.

Searle, S. R. (1971). *Linear models*. New York: Wiley.

Serlin, R. C., Wampold, B. E., and Levin, J. R. (2003). Should providers of treatment be regarded as a random factor? If it ain't broke, don't "fix"it: A comment on Siemer and Joormann (2003). *Psychological Methods, 8*(4), 524-534.

Siemer, M., and Joorman, J. (2003a). Assumptions and consequences of treating providers in therapy studies as fixed versus random effects: reply to Crits-Christoph, Tu, and Gallop (2003) and Serlin, Wampold, and Levin (2003). *Psychological Methods, 8*(4), 535-544.

Siemer, M., and Joorman, J. (2003b). Power and measures of effect size in analysis of variance with fixed versus random nested factors. *Psychological Methods, 8*(4), 497-517.

Snedecor, G. W., and Cochran, W. G. (1989). *Statistical methods* (8th ed.). Ames, Iowa: Blackwell Publishing Professional.

Spybrook, J. (2008). Are power analyses reported with adequate detail? Evidence from the first wave of group randomized trials funded by the Institute of Education Sciences. *Journal of Research on Educational Effectiveness, 1*(3), 215-235.

Tchetgen, E. J. T., and VanderWeele, T. J. (2012). On causal inference in the presence of interference. *Statistical Methods in Medical Research, 21*(1), 55-75.

Visher, M. G., Weiss, M. J., Weissman, E., Rudd, T., and Wathington, H. D. (2012). *The effects of learning communities for students in developmental education: A synthesis of findings from six community colleges.* New York: National Center for Postsecondary Research, Teachers College, Columbia University.

Walters, S. J. (2010). Therapist effects in randomised controlled trials: What to do about them. *Journal of Clinical Nursing, 19*, 1102-1112.

Weiss, M. J., Visher, M., and Weissman, E. (2012). *Learning communities for developmental education students: A synthesis of findings from randomized experiments at six community colleges.* Evanston, IL: Society for Research on Educational Effectiveness. Web site: http://www.sree.org.

Wolf, P., Gutmann, B., Puma, M., Kisida, B., Rizzo, L., and Eissa, N. (2009). *Evaluation of the DC Opportunity Scholarship Program: Impacts after three years.* NCEE 2009-4050. Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. Web site: http://ies.ed.gov/ncee/pubs/.

About MDRC

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

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

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

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

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