MDRC Working Papers on Research Methodology

The Implications of Teacher Selection and Teacher Effects in Some Education Experiments

Michael J. Weiss

April 2010

mdrc
BUILDING KNOWLEDGE
TO IMPROVE SOCIAL POLICY

Acknowledgments

This paper was supported by funding from the Carnegie Foundation. The author thanks Thomas Brock, Howard Bloom, Lashawn Richburg-Hayes, Marie-Andree Somers, Fred Doolittle, Colleen Sommo, Reshma Patel, and Stephen Raudenbush for their extremely helpful comments. The ideas presented and positions taken in the paper are solely the responsibility of the author, however, and do not necessarily reflect views 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 Ambrose Monell Foundation, The Annie E. Casey Foundation, Carnegie Corporation of New York, The Kresge Foundation, Sandler Foundation, and The Starr Foundation.

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

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

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

Copyright © 2010 by MDRC.® All rights reserved.

Contents

List of Tables and Figures

Table

Figure

Abstract

In some experimental evaluations of classroom or school-level interventions it is not practically feasible to randomly assign teachers or schools to experimental conditions. Given such restrictions, researchers may randomly assign students to the program or control group and consider the teacher or school to be a part of the intervention. However, in an individually randomized evaluation of a classroom or school-level intervention, unless teachers or schools are randomized to experimental conditions, it will not be clear whether measured differences between program and control group students are a result of the core components of the intervention or a result of the teachers (that is, teacher effects). This paper clarifies the interpretation of typically calculated "program impacts" in this situation. In addition, using the magnitude of estimated teacher effects from past research, this paper demonstrates that, if teachers or schools are not randomly assigned to experimental conditions, it is significantly more difficult to establish whether the program works or whether the types of teachers selected (or volunteering) to teach in program classrooms are simply more or less effective than their control group counterparts. The significant implications of the correct causal inference to be made are discussed.

Introduction

Randomized experiments have become an increasingly popular design to evaluate the effectiveness of education interventions (Michalopoulos, 2005; Spybrook, 2008). Many of the interventions evaluated in education are delivered to groups of students, rather than to individuals. Experiments designed to evaluate programs delivered at the group level often randomize intact groups, such as classes or schools, to treatment[1] or control conditions in order to obtain unbiased estimates of program effects (Bloom, 2006; Bloom et al., 2008). A growing body of research has discussed the analytics of such group-randomized trials (GRTs), or cluster-randomized trials, as well as the appropriate interpretation of impact estimates (Bloom, 2006; Bloom, Richburg-Hayes, and Black, 2007; Bloom et al., 2008; Hedges, 2007a, 2007b; Hedges and Hedberg, 2007; Konstantopoulos, 2008a, 2008b; Raudenbush, 1997; Spybrook, 2008).

However, in some experimental evaluations of classroom- or school-level interventions, it is not practically feasible to randomly assign teachers or schools to the program or control condition. Instead, such experiments may randomly assign students to the program or control group and deliver the intervention at the classroom or school level. In public health, where this design is common, it has been labeled the Individually Randomized Group Treatment (IRGT) trial (Pals et al., 2008), reflecting the fact that individuals are randomized to experimental conditions, and the treatment is delivered at the group level. 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. The key characteristic of IRGTs is that randomization occurs at the individual level (often referred to as Level 1), and treatment occurs in groups (often referred to as Level 2).

While a great deal of attention has been given to GRTs in education, much less attention has been paid to IRGTs, even though this design is common in many education evaluations (for examples see Decker, Mayer, and Glazerman, 2004; Kemple, 2004; Lang et al., 2009; Love et al., 2002; Richburg-Hayes, Visher, and Bloom, 2008; Scrivener and Au, 2007; Scrivener et al., 2008; Scrivener and Pih, 2007; Visher, Wathington, Richburg-Hayes, and Schneider, 2008).[2] As in GRTs, in IRGTs observations within groups often are not independent; thus, appropriate analytic adjustments must be made in order to obtain accurate standard errors and

[1] Throughout this paper, the terms “treatment group” and “program group” are used interchangeably. Similarly, the terms “treatment,” “program,” and “intervention” are also used interchangeably.

[2] 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. In addition, many “natural experiments” that occur as a result of lotteries held at the individual level fall into this category of experiment.

not inflate the likelihood of making type I errors. The need to account for the lack of independence of observations in IRGTs has been documented, along with many examples where correct adjustments have not been made (Bauer, Sterba, and Hallfors, 2008; Crits-Christoph and Mintz, 1991; Hedges, 2007a; Pals et al., 2008; Roberts and Roberts, 2005). However, the implications of using the IRGT design, with respect to the correct causal interpretation of impact estimates, have rarely been well documented.[3]

This research describes how, in evaluations of classroom-level interventions that randomize students (and not teachers) to experimental conditions, it will be unclear whether estimates of the impact of the program reflect the effect of the core components of the program or the types of teachers delivering the program (that is, teacher effects). This potential confounding of program effects and teacher effects can be a major concern if teachers are sorted into experimental conditions in such a way that they differ at the outset of the study. This paper attempts to make clear the correct causal interpretations of typically calculated "program impacts" in this situation.

In addition, using the magnitude of estimated teacher effects from prior research, this paper demonstrates that if teachers are not randomly assigned to experimental conditions, it is significantly more difficult to establish whether the intervention "works" or whether the types of teachers selected to teach in intervention classrooms are simply more or less effective than their control group counterparts. The implication is that the usefulness of such studies' findings may be severely limited.

This paper is divided into four main sections. The first section provides some background motivating this research. Section Two describes a concrete example of a real IRGT evaluation in education, clarifying the main issue regarding the use of this evaluation design — that program impacts can be confounded with teacher effects. This is followed by Section Three, where the robustness of the desired causal inference in the IRGT example is explored. Finally, Section Four discusses the implications, solutions, and conclusions of this research.

Section One: Background

This section begins with a brief review of some of the reasons why researchers randomize students to experimental conditions in simple experiments (without clustering). Recalling some of these reasons is intended to help draw clear and direct parallels to the implications of not randomizing teachers to experimental conditions in an IRGT trial.

[3] Recent progress on this topic has been made in the field of psychotherapy (for examples see Krause and Lutz, 2009; Lambert and Baldwin, 2009; Stiles, 2009). An additional notable exception is an article by Stephen Raudenbush, where he clearly describes this phenomenon using a potential outcomes framework (Raudenbush, 2008).

Why Randomize Students to Experimental Conditions?

In most observational evaluations in education, students are *not r*andomly assigned to treatment or comparison conditions. In such evaluations, where an unknown mechanism is used to sort students into the program or comparison group, it is difficult to determine the cause of any observed differences in outcomes between program and comparison group members. Differences may be a result of the treatment or intervention, but they also may reflect pre-treatment differences between program and control group members — differences that are a result of the selection process that led students to be in either the program or the comparison group. For example, more motivated students might seek out a new and innovative intervention, leading to a treatment group containing highly motivated individuals and a comparison group containing less motivated individuals. Subsequent differences in outcomes between the program and comparison groups could simply reflect differences in the initial motivation levels of the students in each group.

While researchers can take great care to statistically control for differences in the observable characteristics of students (such as gender or race), there is still uncertainty regarding whether students in the program and comparison conditions differ on unobservable characteristics (such as motivation or ability), which may be related to the outcome of interest. Notably, if the selection process (that is, the mechanism used to sort students into treatment and comparison conditions), is understood completely (for example, in a regression discontinuity design), then randomization may not be needed in order to make strong causal claims. Similarly, if the selection process could be statistically modeled well enough, randomization might not be needed to make causal claims. (See Rosenbaum and Rubin, 1983, for a precise explanation of the assumptions one has to make.) The problem is that it is extremely difficult (and often impossible) to know when the selection process has been modeled "well enough;" consequently, randomization is used to ensure that program and control groups are approximately equivalent at the outset of a study, thus avoiding the selection concern. With the creation of approximately equivalent groups through the deliberate random assignment selection process, researchers can be reasonably assured that average differences in outcomes between experimental groups are not a result of differences in the types of people in the experimental groups; rather, differences in outcomes are *causally* attributable to systematic differential treatment of group members after random assignment.

When and Why Randomize Teachers to Experimental Conditions?

In the situation where a program or intervention is offered at the classroom level, it may be important to randomize teachers to experimental conditions for reasons that are very similar to the reasons why researchers randomize students to experimental conditions. In a study of a classroom-level intervention *without* randomization of teachers to the program or control group, it may not be clear whether measured differences between program and comparison group students are a result of the treatment/intervention or the types of teachers that ended up in the program and comparison conditions. This may be a major concern if teachers are sorted into the program and comparison conditions in such a way that they differ on variable(s) that are related to their effectiveness, their toughness in grading (if the outcomes of interest are not standardized across classrooms), or other variables related to their implementation of the intervention. By not randomizing teachers, the valid inferences one can make with respect to program effectiveness change significantly. This point will be elaborated on in the second section of this paper, which provides a concrete example from a real education experiment.

Group-randomized trials that randomly assign teachers (or schools) to experimental conditions ensure that the treatment and control group (both teachers and students) are approximately equivalent at the outset of the study, avoiding both the student and teacher selection concern. With the creation of approximately equivalent groups through a deliberate random assignment process, researchers can be reasonably assured that average differences on outcomes between experimental groups are a result of systematic differential treatment of group members after random assignment and not a result of the types of teachers (or students) in the groups.

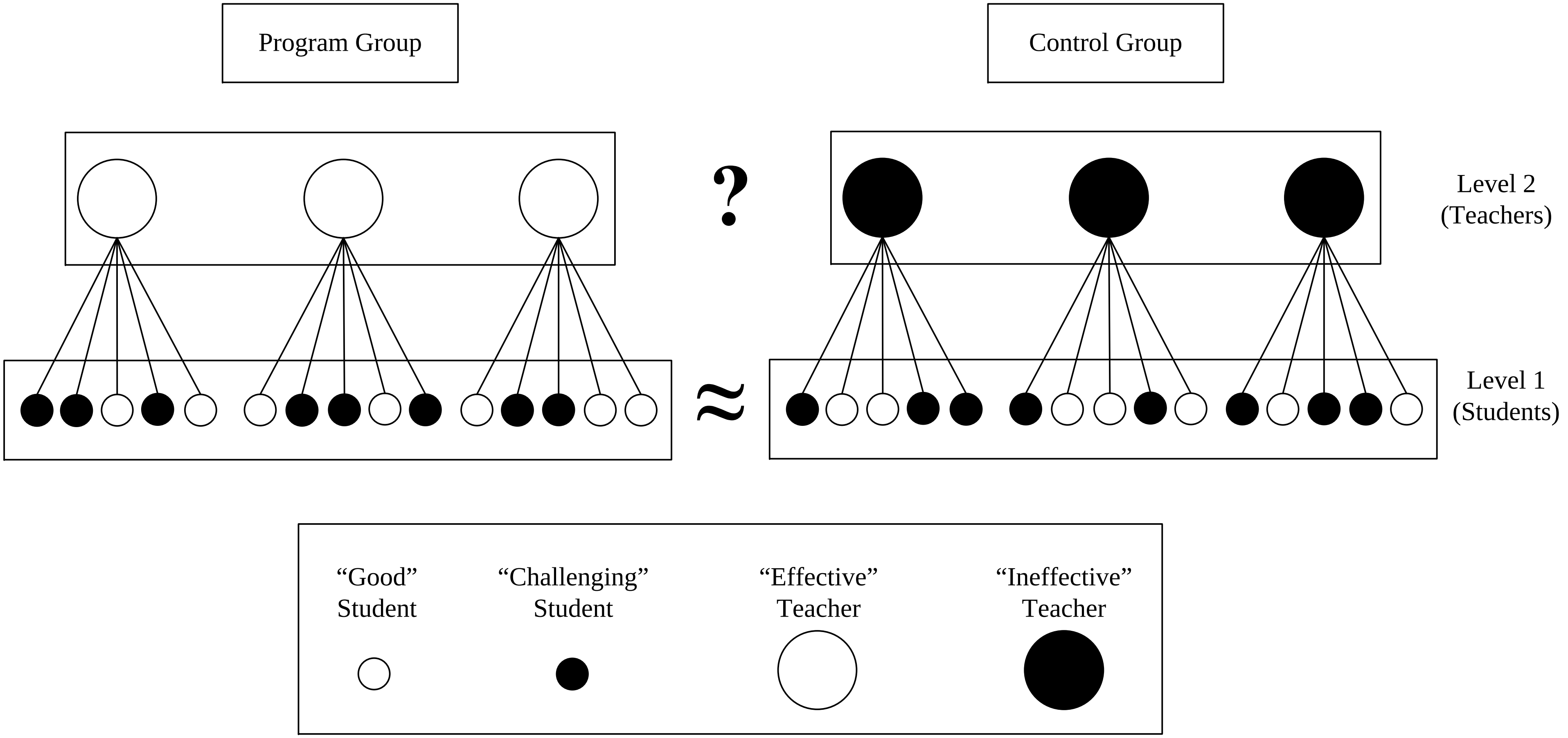
Individually Randomized Group Treatment (IRGT) Trials

The IRGT trial is a study design that shares some of the characteristics of both simple experiments, where individuals are randomized to the treatment group or control group, and group-randomized trials, where intact groups are randomly assigned to the treatment or control group. Figure 1 provides a visual depiction of the IRGT design where students are randomly assigned to experimental condition, the program/intervention is delivered at the classroom level, and teachers are *not* randomly assigned to experimental condition. To simplify reality, in Figure 1 there are two types of students, “Good Students” and “Challenging Students,” and there are two types of teachers, “Effective Teachers” and “Ineffective Teachers.”[4] Since randomization occurs at Level 1, or the student level, there is reasonable assurance that the good students and the challenging students will be split about equally between the program and the control groups

[4] It is worth pointing out that these “types” of students and teachers refer to their pre-experimental statuses. An intervention may be designed to turn ineffective teachers into effective teachers. In such a case, if the intervention is successful, one should expect to see differences in the effectiveness of program group teachers and control group teachers after the study begins. The issues raised here arise when there are pre-experimental differences between program group teachers and control group teachers.

Figure 1

Teacher Selection in Individually Randomized Group Treatment Trials



(hence the "≈" at Level 1 in Figure 1). As a result, observed differences between program and control group students can be causally attributed to systematic differential treatment of program and control group students after random assignment. Such observed differences are often referred to as "program impacts," and they reflect the relative benefit or harm of the experiences of the program group students compared with the experiences of the control group students after random assignment. What is critical when considering the IRGT design is that the estimated program impact may be composed of at least two key factors:

1. The core components of the intervention being tested (for example, a new math curriculum or being part of a learning community)
2. The types and effectiveness of teachers in the program and control group

Figure 1 shows Factor 2 visually, using larger circles at Level 2, which represent teachers. Since teachers are not randomly assigned in the IRGT trial, there is no assurance that the effective and ineffective teachers will be split approximately equally between the program and the control group (hence the "?" at Level 2). In the example shown in Figure 1, the program group teachers are the effective teachers, and the control group teachers are the ineffective teachers. This could occur because of the way teachers are recruited to participate in the program; for example, perhaps a principal selects her best teachers to participate in a new program, or perhaps the most dedicated teachers are drawn to the intervention on their own. As a result, estimated program impacts reflect the "teacher effect" and/or the effect of the core components of the program. In essence, the teachers become a part of the program. Consequently, estimated program impacts must be interpreted as the combined effect of the core components of the program plus the types of teachers that were selected into the program group.

Conditions Required for Teacher Selection/Effects to Be a Concern

The issue described in the previous paragraphs rests on two assumptions. First, that there is a teacher effect; that is, there must be variation in teachers' effectiveness levels. If all teachers were equally effective (and a standardized outcome measure was used to assess student success), IRGT trials could obtain unbiased estimates of the effects of the core components of the program being evaluated.[5] Second, teacher selection must be a concern; that is, the process by which teachers are sorted into the program and control group must potentially produce selection bias. If bias in sorting does not occur, then the IRGT should produce unbiased estimates of the effects of the core components of the program being evaluated.

[5] In addition to this assumption, it also must be assumed that program and control teachers are equally effective at implementing the intervention being tested.

The Teacher Effect

Some of the strongest evidence to date of the existence and magnitude of the teacher effect comes from Nye, Konstantopoulos, and Hedges's (2004) paper titled "How Large Are Teacher Effects." Their analyses using Tennessee STAR's experimental data demonstrate that "teacher effects are real and…there are substantial differences among teachers in the ability to produce achievement gains in their students" (Nye, Konstantopoulos, and Hedges, 2004, p. 253). Depending upon the model, grade, and subject area, the magnitude of their estimates of the teacher effect varies, but in general they find that teachers account for around 10 percent of the variation in student test score gains.[6] The analyses conducted in the third section of this paper will assume that teachers account for 10 percent of the variation in student academic outcomes.[7]

Teacher Selection

Evidence of teacher effects alone does not suggest the need to be concerned about confounding teacher effects with the effects of the core components of the intervention in an IRGT. If teachers sort into the program and control conditions in such a way that there are not pre-experimental differences between them, then impact estimates can correctly be interpreted as the effect of the core components of the intervention being studied. The problem is that, without randomization, researchers cannot be certain whether such group equivalence holds. Accordingly, there is a theoretical basis for being concerned about the confounding of program impacts and teacher effects.

Section One of this paper describes the situation in an experimental evaluation where the impact of the core components of a program cannot be disentangled from teacher effects. In order to demonstrate the implications and significance of this type of situation, Sections Two and Three use a real experiment as an example to make the issues more concrete. The specific example was chosen out of convenience, but it is intended to exemplify the issues more generally.

[6] This finding confirms findings from their review of past nonexperimental research, where Nye et al. found evidence that suggests that "from 7% to 21% of the variance in achievement gains is associated with variation in teacher effectiveness" (Nye, Konstantopoulos, and Hedges, 2004, p. 240).

[7] The work of Nye et al. assessed the effectiveness of teachers from kindergarten through third grade, using data from the 1980's, in a single state. While generalizing from this work alone might not be prudent, their review of the literature suggests that many past nonexperimental studies find evidence of a teacher effect of similar magnitude. In addition, more recent research points in the same direction, suggesting that it is safe to assume that teachers do, in fact, matter (Clotfelter, Ladd, and Vigdor, 2007; Kane and Staiger, 2008; Kukla-Acevedo, 2009; Rockoff, 2003; Rockoff, Jacob, Kane, and Staiger, 2008).

Section Two: An Example of an IRGT Trial in Education — An Evaluation of Learning Communities

In 2002 the Opening Doors Demonstration was launched, a first of its kind to conduct several rigorously designed randomized experiments in the community college setting. The demonstration took place at six community colleges throughout the United States and evaluated four different programs designed to improve students' likelihood of achieving academic success (Brock, LeBlanc, and MacGregor, 2005).

One of these four interventions, small learning communities, was studied at a single college, Kingsborough Community College in Brooklyn, NY. (Bloom and Sommo, 2005; Brock, LeBlanc, and MacGregor, 2005; Scrivener et al., 2008)

In this evaluation, the core components of the learning communities program were:

- **Paired- or clustered-course model:** Students were divided into groups of up to 25. These groups formed "learning communities" where students within each learning community took three courses together.

- **Teacher Collaboration:** The teachers teaching the learning communities courses were expected to collaborate, coordinating their syllabi before the semester began and meeting regularly during the semester to discuss student progress.

The Opening Doors learning communities program was a bundled program that included several other components (enhanced counseling and support, enhanced tutoring, and book vouchers), but for the purposes of this example it is simpler to consider the program's two primary components listed above.

The findings from the Opening Doors evaluation of learning communities were generally positive (Richburg-Hayes, Visher, and Bloom, 2008; Scrivener et al., 2008), and they, in part, led to a multisite evaluation of learning communities known as the Learning Communities Demonstration (Visher et al., 2008). Largely due to practical restrictions, both the Opening Doors evaluation and the Learning Communities Demonstration utilize an IRGT study design. In these studies, students were assigned, at random, to be part of the program group, which was eligible for the learning community, or the control group, which received the college's standard services (Scrivener et al., 2008; Visher, Wathington, Richburg-Hayes, and Schneider, 2008). However, teachers were *not* randomly assigned to teach learning community classes. Consequently, it is possible that, at the outset of the study, the learning community teachers were more (or less) effective teachers than their control group counterparts.

The Implications of Not Randomly Assigning Teachers to Experimental Conditions

As noted earlier, since teachers were not randomly assigned to experimental conditions in the learning communities studies, it is uncertain whether observed impacts are a result of (1) the core components of the program (the paired-course model and teacher collaboration) and/or (2) differences between the learning community teachers and their control group teacher counterparts at the outset of the experiment. For example, it is possible that teachers volunteered to teach in the learning communities, drawing the most motivated teachers to the treatment condition. If these more motivated teachers were also more effective, it is possible that students in the program group would have outperformed students in the control group even in the absence of the learning communities program. If this was the case, the core components of the program may have had no impact at all, and the college may have simply reshuffled its more effective teachers.

Alternatively, it is possible that most teachers resisted teaching in the learning communities and that learning community classes were taught by coerced "volunteers." If these coerced volunteers were, on average, less effective than their control group teacher counterparts, then it is possible that the impacts of the core components of the program were underestimated. If this was the case, the program may have even larger impacts than were estimated, but the program first had to overcome its less effective teachers.

In fact, qualitative work from the Opening Doors learning communities experiment suggests that some teachers enthusiastically taught in learning communities, some were recruited by department chairs, and later in the study some instructors came forward voluntarily (Scrivener et al., 2008). So it is quite possible that the selection mechanism that led teachers to teach in the learning communities influenced the study's estimated impacts.

Teacher Selection: Just One More Component of the Program Package?

One of the understood limitations of many education experiments is the inability to disentangle the effects of each of the components of the studied program. For example, in a recent classroom-level random assignment evaluation of an effective pre-kindergarten mathematics intervention, the "program package" included three components: the classroom components of *Pre-K Mathematics*, the home components of *Pre-K Mathematics*, and access to *DLM Express* math software for classroom use (for more details see: Klein et al., 2008). Because the program was "bundled," it is not possible to disentangle the effects of each component of the intervention. Nonetheless, if a pre-kindergarten program similar to the ones participating in this study decided to implement the entire program package, it might reasonably expect positive overall results, regardless of the component(s) of the program package that mattered in the original study.[8]

Similarly, in the learning communities study, at first glance one might think of the teachers as just another component of the program package. However, under careful scrutiny, the inferences one can make as a result of teacher selection are far more limited. What one might like to say is that the learning communities' program package includes the learning communities and the types of teachers that were selected to teach in them. However, there are serious implications to this assertion.

First, it is unlikely that the selection process that placed teachers into learning communities can be explained completely,[9] so replicating this process at another college may prove difficult. This is unlike the pre-kindergarten mathematics intervention described above, which can be fairly clearly defined and therefore may be replicable in another setting.

Second, the implications of nonrandom teacher selection are different than the implications of the packaging of other components. Consider a school that is not implementing learning communities that decides to implement the program package based on the success of the Opening Doors learning communities study. Such a school could expect overall positive academic outcomes only if the core components of the learning communities program (not teacher selection) were the causal mechanism that led to positive impacts in the learning communities study. If the positive impacts observed in the study were a result of pre-experimental differences in teacher effectiveness (and not a result of the core components of learning communities), a college implementing learning communities should *not* expect overall positive academic outcomes. This is true even if the college had its teachers select into the program through a mechanism identical to the one used in the study.

Rather, if teacher selection was the underlying causal mechanism leading to the learning community study's positive findings, in order for a college to experience overall improvements it would need to use the teacher selection mechanism from the learning communities study for the purpose of hiring and firing teachers. Moreover, since it is not possible to discern which part(s) of the program package mattered (the learning community components or the teachers), a school that wants to ensure improvement has to both (1) implement learning

[8] For the purpose of making a point, please ignore the reasons that this statement might not be completely true, such as the generalizability of results from a study conducted in 40 classrooms.

[9] Here, to "explain completely" does mean "to describe in words." For example, knowing that program teachers were "volunteers" does not suffice. Rather, a precise measure of all teachers' (program and control) propensities to volunteer is needed, as well as an understanding of the relationship between teachers' propensities to volunteer and the outcome measures of interest. However, it is precisely because of our lack of confidence in the ability to model such selection processes that experiments are conducted in the first place.

communities and (2) make hiring and/or firing decisions based on the selection mechanism used to determine which teachers taught in learning communities in the study.

In order to understand the different implications of the ways program "impacts" could be observed when teacher selection is a concern, it is useful to consider a simplified example.

Table 1 depicts one school under six hypothetical scenarios of results of a hypothetical IRGT experiment. Two variables, located in columns A and B, determine the values in the remaining columns. Column A describes whether the core components of the program "work" or not.[10] When the program works, program group students get a 10 percentage point bump in their pass rates. Column B describes whether teacher selection was (a) not an issue, (b) was an issue favoring the program group (that is, the program group got more effective teachers), or (c) was an issue favoring the control group (that is, the control group got more effective teachers). In general, both control group students and program group students get a 10 percentage point increase in their pass rates when they have an effective teacher compared with when they have an ineffective teacher.

Columns C through F provide the course pass rates for control group students and program group students, which depend upon both whether the program works and the effectiveness of their teachers. Column G shows the program's estimated impact, and column H shows the overall pass rate among all students in the hypothetical study (program and control). For simplicity, assume that there are an equal number of program and control group students. As a result, the impact estimate is the average of the control groups' pass rates (columns C and E) subtracted from the average of the program groups' pass rates (columns D and F), and the overall pass rate is the average pass rate for the row (columns C through F).

Row 1 can be thought of as a baseline hypothetical example. As indicated by column A, the core components of the program are neither beneficial nor harmful to students in this scenario. In addition, in row 1, teacher selection is not a problem; that is, there is an equal split of more and less effective teachers between the program and control group. This would be expected to occur if teachers are randomly assigned to their experimental group; however, even without teacher random assignment it could happen. In scenario 1, the pass rate for program and control group students with more effective teachers is 55 percent, whereas the pass rates for program and control group students with less effective teachers is 45 percent.[11] Since the program has no effect, program and control group students pass at the same rate, *given that they have teachers of equal effectiveness*. As a point of reference, in this baseline scenario the estimated program impact is zero and the overall pass rate is 50 percent.

[10] For simplicity, the scenario where the program is harmful is not included.

[11] These are fairly arbitrary numbers chosen for this example.

Table 1

Course Pass Rates Under Six Hypothetical Scenarios

	(A)	(B)	(C)	(D)	(E)	(F)	(G)	(H)
	Does the Program Work?	Is Teacher Selection an Issue?	More Effective Teachers		Less Effective Teachers		Impact Estimate	Overall Pass Rate
			Control	Program	Control	Program		
(1)	No	No	55%	55%	45%	45%	0%	50%
(2)	No	Yes (Favoring Program Group)	--	55%	45%	--	10%	50%
(3)	No	Yes (Favoring Control Group)	55%	--	--	45%	-10%	50%
(4)	Yes	No	55%	65%	45%	55%	10%	55%
(5)	Yes	Yes (Favoring Program Group)	--	65%	45%	--	20%	55%
(6)	Yes	Yes (Favoring Control Group)	55%	--	--	55%	0%	55%

Rows 2 and 3 present results from two hypothetical scenarios where the program is again neither beneficial nor harmful, but teacher selection occurred in such a way that the more effective teachers ended up in the program group (scenario 2) or the control group (scenario 3). Critically, because the core components of the program have no effect in these two scenarios, the school's overall pass rate remains at 50 percent. However, because of teacher selection, both scenarios result in nonzero impact estimates. Under scenario 2, a researcher might erroneously conclude that the core components of the program are working, and she might expand the program with little success. Under scenario 3, a researcher might erroneously conclude that the core components of the program are actually harmful, when in fact they are having no impact on course pass rates at all. In scenarios 2 and 3, all that has happened is that teachers have been reorganized in such a way that one group of students is taught by more effective teachers than the other, but *overall,* students are no better or worse off than they would have been in the absence of the program.

Rows 4 through 6 present results from three hypothetical scenarios where the core components of the program are beneficial, providing a 10 percent increase in students' likelihood of passing their courses. In all three scenarios, the overall pass rate at the school has increased as a result of half of the school's students receiving a program that is more effective than business as usual. However, in scenarios 5 and 6, the estimated impact of the core components of the program is confounded with teacher effects that resulted from teacher selection. In scenario 5, the program's impact is overestimated, and in scenario 6 the program is estimated to have no impact at all.

Teacher selection is a concern in the real world because scenarios 2 and 4 are indistinguishable, as are scenarios 1 and 6. Consequently, it is unknown whether the core components of the program work (or are harmful) or if there is a teacher effect and teacher selection is influencing impact estimates. Notably, it is possible to make use of the impact estimates by considering the program as a package. However, doing so requires acknowledging that the impacts are potentially a result of the core components of the program and/or the types of teachers selected into the program. The implication is that a school that wants to experience the benefits of the program package would need to both implement the program and make hiring/firing decisions based on the teacher selection mechanism that was used in the research study. Thus, the usefulness of the program impacts may be drastically reduced when teacher selection is a concern.

Section Three: Robustness of the Desired Causal Inference

Section Two of this paper described the Opening Door learning communities study, explaining why concerns of teacher selection led to the potential confounding of program impacts that are a result of the core components of the learning communities and teacher effects. Here, the observed impacts from that study are reexamined to test how sensitive they are to the teacher selection concern.

Table 2 provides a summary of the main positive impacts from the Opening Doors learning communities study (Richburg-Hayes et al., 2008; Scrivener et al., 2008). During the program semester,[12] highly statistically significant positive program impacts were observed on students' likelihood of passing all courses, the number of courses they passed, and the number of credits they earned. In addition, there was somewhat weaker evidence that the program boosted registration rates in the third postprogram semester and improved students' likelihood of passing both English tests by the end of the second postprogram semester. In order to demonstrate the sensitivity of these results to teacher selection, the program's impact on credits earned is focused on here.[13] On average, program group students earned 11.5 credits and control group students earned 10.4 credits, resulting in an estimated program impact of 1.2 credits earned. The control group's standard deviation on this outcome was 7.2, a value that proves useful when assessing the sensitivity of these findings.

To assess the sensitivity of this impact estimate to the teacher effect/selection, an assumption must be made regarding the magnitude of the teacher effect (that is, the proportion of

[12] The learning communities program was a one-semester intervention. The "program semester" refers to the semester during which program group students were taking the learning community's linked classes. "Postprogram" semesters refer to those semesters after the program semester.

[13] Credits earned was selected because it fell in the middle of the impacts in terms of its effect size, where effect size was calculated as the impact estimate divided by the control group standard deviation.

Table 2

Main Impacts from the Learning Communities Study

	Program Group	Control Group	Difference (Impact)	Standard Error	Control Group S.D.
Passed all courses (%)	43.1	33	10.1 ***	2.5	47
Courses passed (%)	3.8	3.2	0.6 ***	0.1	2.2
Credits earned	11.5	10.4	1.2 ***	0.4	7.2
Registered for any course (%) (3rd postprogram semester)	52.9	47.8	5.1 *	2.7	50
Passed both English tests (%) (by end of 2nd postprogram semester)	65.2	60	5.2 *	2.7	49

SOURCE: Scrivener (2008) and Richburg-Hayes (2008). The control group standard deviations were not reported in the cited articles, but were provided to the author by MDRC.

NOTES: A two-tailed t-test was applied to differences between research groups. Statistical significance levels are indicated as: *** = 1 percent; ** = 5 percent; * = 10 percent.

variation in student outcomes that is explained by teachers). Based on previous research, it is assumed that teachers account for 10 percent of the variation in credits earned; therefore, the standard deviation of the teacher effect is equal to 2.3 credits earned (See Appendix A for derivation). Assuming the teacher effectiveness distribution is normal, a student's expected credits earned can be estimated given the average effectiveness of her teachers. Figure 2 displays the teacher effectiveness distribution, where the x-axis represents the expected number of credits earned for students. The middle vertical line in Figure 2 shows that a student in the Opening Doors learning communities study with teachers of average effectiveness (50th percentile) can be expected to have earned around 11.0 credits during the first program semester (this is equal to the overall mean number of credits earned for all students). Likewise, a student with 30th-percentile teachers[14] is expected to have earned 9.8 credits,[15] and a student with 70th-percentile teachers is expected to have earned 12.1 credits.[16] Thus, the difference in mean expected credits earned for students with 30th- and 70th-percentile teachers is 2.4 credits earned. Comparing 2.4 with the observed impact of 1.2 on credits earned provides an indication of the sensitivity of the observed impacts to selection bias resulting from teacher effects.

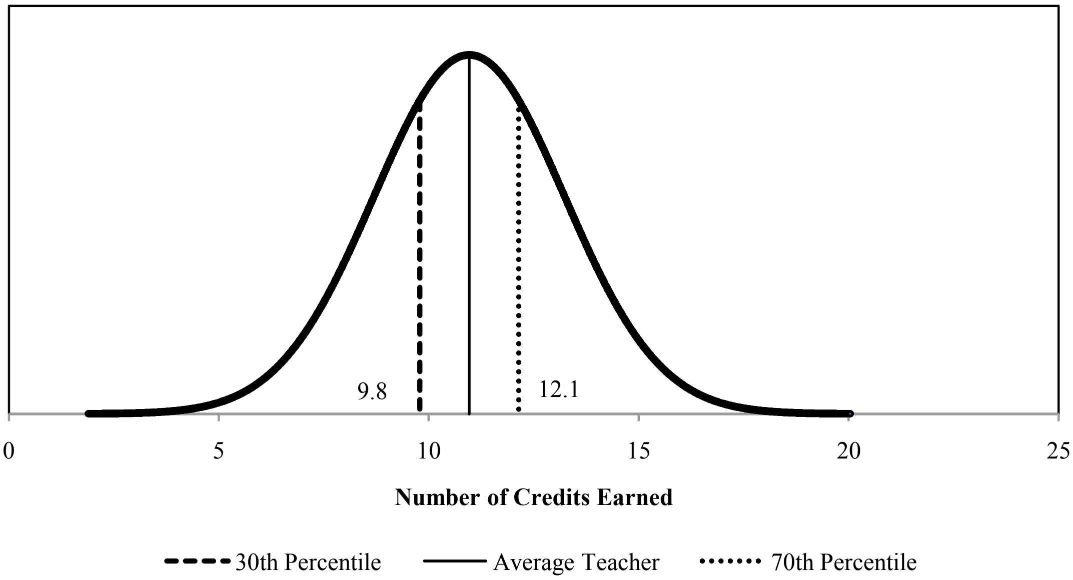
[14] 30th-percentile teachers are defined as teachers 0.52 standard deviations below the mean, since the probability of an observation being at least 0.52 standard deviations below the mean on a standard normal distribution is 30 percent.

[15] This is calculated as 11.0 - 0.52*2.3, where 0.52 is obtained as described in the previous footnote, and 2.3 is equal to the standard deviation of the teacher effect.

[16] Numbers may appear off due to rounding.

Figure 2

Teacher Effectiveness Distribution

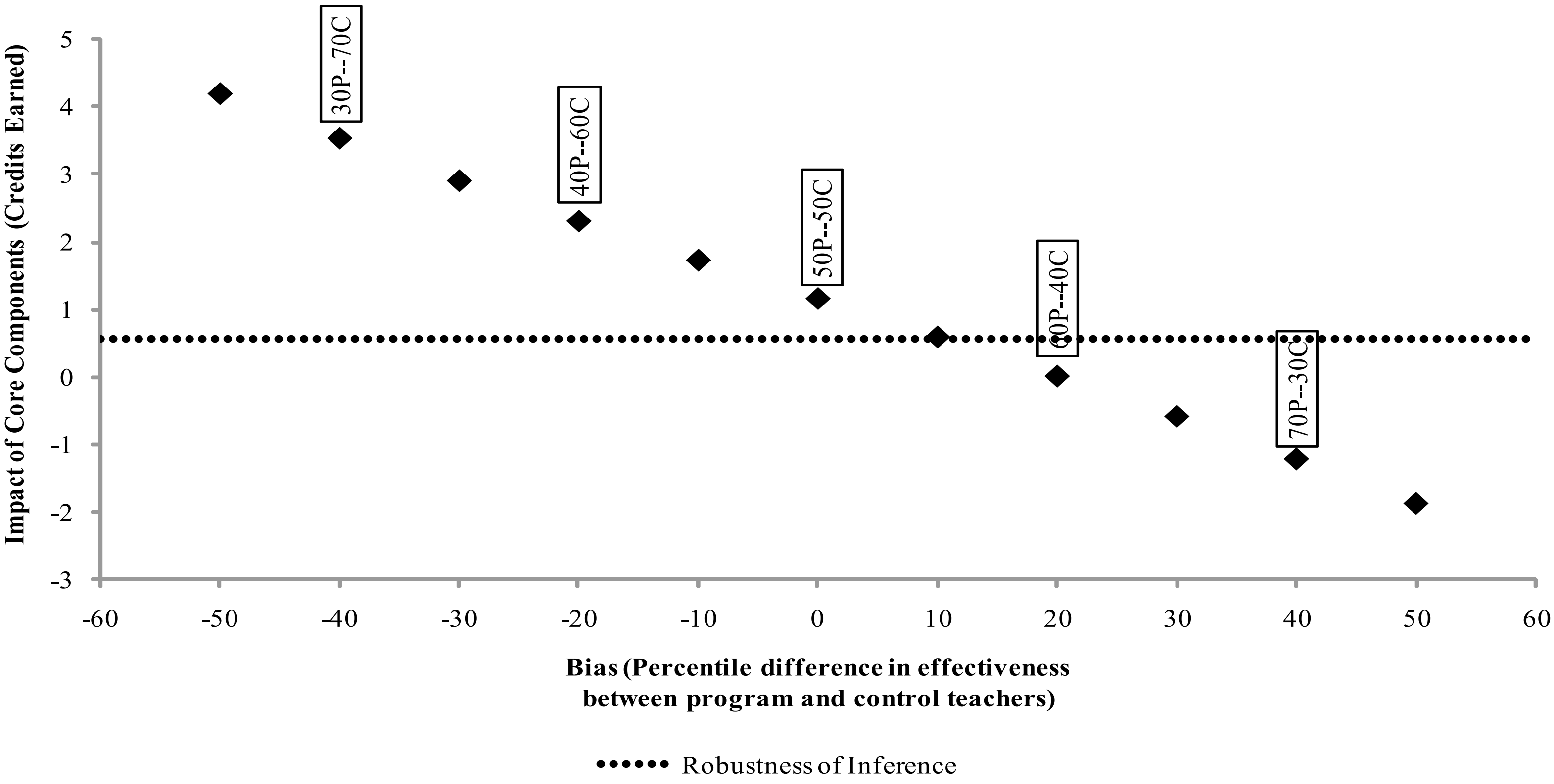


For example, if program group teachers were in the 70th percentile of the teacher effectiveness distribution and control group teachers were in the 30th percentile of the teacher effectiveness distribution, then the estimated impact of the core components of the program would actually be -1.2 credits earned. In other words, the program could actually have been harmful. On the other hand, if the reverse occurred, and the program group teachers were in the 30th percentile of the teacher effectiveness distribution and the control group teachers were in the 70th percentile of the teacher effectiveness distribution, then the impact of the core components of the program could actually have been 3.4 credits earned. In other words, the program may have been much more effective than was reported.

Figure 3 provides a visual depiction of the influence of the potential confounding of teacher effects with the impacts of the core components of the learning communities program. In this graph, the y-axis represents the estimated impact of the core components of the learning communities program. The x-axis represents different amounts of selection bias that could have resulted from the nonrandom assignment of teachers to the program and control group. Here, bias is used to refer to the percentile difference in average effectiveness of the program group teachers compared with the control group teachers. For example, in the middle of the x-axis is 0 percent bias. The data point above 0 is labeled "50P--50C," meaning that program and control group teachers were, on average, in the 50th percentile of the teacher effectiveness distribution. In this situation there is 0 bias, and the estimated impact of the core components of learning

Figure 3

Sensitivity of Estimated Impacts to Teacher Selection



communities is the same as the estimated impact from the original analyses (1.2 credits earned). Sliding right to the value of 40 on the x-axis leads to the data point described in the previous paragraph, where program group faculty were in the 70th percentile of the teacher effectiveness distribution and control group faculty were in the 30th percentile of the teacher effectiveness distribution (hence the data point label "70P--30C"). As noted earlier, in this scenario the estimated impact of the core components of the program is actually negative 1.2 credits earned. What is most striking about this graph is that, if teacher selection is at all significant, it can swamp the observed impact estimates. In other words, if the assumed magnitude of the teacher effect is accurate, then teacher selection should be a serious concern because the teacher effect is quite large compared with the program's impacts.

Another way to consider Figure 3 is to ask the question "how serious would teacher selection have to have been to alter the inference with regard to the program's effectiveness?"[17] In Figure 3 the dotted horizontal line labeled "Robustness of Causal Inference" represents the magnitude of the program's estimated impact that had to be exceeded in order for the impact to be deemed statistically significant (.577 credits earned).[18] This is equivalent to the Minimum Detectable Effect (MDE) $\alpha = .10$ and $\beta = .50$. What the figure shows is that the positive, statistically significant finding would no longer have been deemed significant if teacher selection were such that program group teachers were in the 56th percentile of the teacher effectiveness distribution and control group teachers were in the 44th percentile of the teacher effectiveness distribution. In other words, if the desired causal inference of the learning communities study is about the effect of the core components of the program (and not teacher effectiveness), then the program's estimated impact on credits earned is quite sensitive to the possibility of teacher effects due to teacher selection.

This section has explored the robustness of the desired causal inference in spite of the potential confounding of that inference with teacher effects. As demonstrated, under fairly reasonable assumptions, the teacher effect is a serious concern in IRGT trials in education.

Section Four: Solutions, Bigger Picture, Conclusions

The final section of this paper discusses solutions to the problem described in Sections One through Three, considers bigger picture implications of the general idea described in this work, and offers some final conclusions.

[17] This general way of considering sensitivity analyses has been referred to as the "impact threshold for a confounding variable (ITCV)." In this case it reflects the amount of the teacher selection necessary to make positive and statistically significant observed program impact become positive and *just* statistically significant (Frank, 2000).

[18] The statistical significance threshold reported by MDRC (using a 2-tailed test with $\alpha = .10$) is used for this analysis (Scrivener et al., 2008).

Solutions

While the concerns outlined in this paper may be serious, there are potential solutions or improvements that can be made in education experiments to mitigate these concerns. Outlined here are four potential solutions and improvements. Three solutions involve modifying the experimental design, and one is a nondesign-related improvement.

The ideal solution to the problem described in this paper is to conduct random assignment at the level at which the program or intervention is being delivered (that is, the classroom or school level). Doing so circumvents the teacher effect problem because it helps ensure group equivalence at the level of random assignment; that is, it can ensure that the program and control group teachers are of similar levels of effectiveness. Famously, this was achieved in the Tennessee STAR study, where *both* students and teachers were randomly assigned to different class sizes in order to measure the impact of class size on student learning. Notably, while it may be preferable to randomly assign *both* students and teachers, in many program evaluations it is sufficient to simply randomly assign the higher unit of analysis to program or control conditions (that is, the teacher or the school). Doing so is internally valid *if assignment of the lower-level units (students) to the higher-level units (teachers or schools) is done prior to, or is unaffected by, the random assignment of the higher-level units (teachers or schools) to experimental conditions.* This design is becoming increasingly popular in education (Spybrook, 2008). Such group-randomized trials can eliminate the concern that program impacts are confounded with teacher effects. While more ideal than the IRGT, this design is not always practically feasible; in fact, in the example of the learning communities it is rather difficult to imagine any reasonable way this design could have been used.[19]

A second solution to this problem is to evaluate interventions that are uniformly delivered at the student level. For example, performance-based scholarships, online tutoring programs, or free laptops for economically disadvantaged students can be evaluated with little concern that program impacts will be contaminated by teacher effects. In such studies, program and control group students may be mixed together in classes such that teacher effects influence the two experimental groups approximately equally.[20] While such studies can maintain a high level of internal validity, this solution is rather unsatisfying since it severely limits the types of interventions that can be evaluated.

When the teacher effect/selection is a concern, a third solution may be to have study teachers teach in both the program and the control group. In this way the teacher effect can be

[19] There are several reasons for this assertion. First, control group students were free to take whatever courses they wanted. In order to randomly assign faculty, every faculty member at the college would have to agree to participate in the study. In addition, insurmountable scheduling difficulties would arise using a design that involved random assignment at a higher level (both from a faculty and a student perspective).

[20] This can become complicated if the program or intervention influences course selection.

balanced across experimental groups, and impact estimates may better reflect the core components of the program.[21] While it is useful for researchers to have this design in their toolkit, it may lose some of its merit if “contamination” is a concern. To clarify what is meant by contamination, imagine a study seeking to determine the effectiveness of a new curriculum. If study teachers teach in both the program group (delivering the new curriculum) and the control group (delivering the old curriculum), contamination may occur when the teacher uses some of the new program’s practices in her control classroom. Such contamination may reduce the program versus control contrast, diluting estimated treatment effects.

Finally, if a study is designed such that teacher selection and the teacher effect may influence the impact estimates, researchers can use nonexperimental methods to attempt to control for teacher effects. One way to do this would be to control for teacher effects by incorporating teachers’ value-added scores (from the year prior to the experiment) into the impact model. When value-added scores are not available, researchers could control for factors that are known (or thought) to be related to teacher effectiveness (for example, years of experience, certification, or content knowledge). Any such controls are unlikely to eliminate completely the influence of teacher effects on impact estimates, but it is possible that they may help. At a minimum, obtaining baseline information on program and control group teachers will allow researchers to explore the ways that teacher selection may be a concern.

Bigger Picture

This paper focuses on teacher selection and the teacher effect in IRGT trials in education; however, the concerns raised here have implications beyond this particular experimental design applied to education research. First, the IRGT trial study design may have similar limitations in noneducation fields where nested data structures are common, such as health care (patients nested within therapists/doctors) or welfare (families nested within case managers). If there is variation in therapists’ effectiveness (Crits-Christoph and Mintz, 1991; Krause and Lutz, 2009; Lambert and Baldwin, 2009; Stiles, 2009), or variation in the effectiveness of case managers (Brock and Harknett, 1998), for example, then concerns that are analogous to those described in this work should exist in some experiments conducted in these fields as well.

Moreover, even if the data structure is not nested, the concerns raised in this paper may arise under any circumstance where the agent(s) or deliverer(s) of an intervention are not randomly assigned. For example, imagine an evaluation seeking to compare one-on-one tutoring program X with one-on-one tutoring program Y. Program X requires tutors to support

[21] To some extent, this occurred in the learning communities study described in this paper, since some program group teachers also taught non-learning communities classes. However, since control group students were free to take whatever classes they liked, it is unclear how frequently control group students took classes with instructors who taught in the program group.

their tutee academically *and* socially, whereas program Y has a solely academic focus. If the evaluation randomly assigns each *tutee* to either program X or program Y, but the *tutors* are not randomly assigned, the exact same concerns described in this paper will arise (even if there is one tutor per tutee). Because the tutor effect and the program effect cannot be disentangled, such a study might not be able to answer a question such as, "Does one-on-one tutoring that includes academic and social support outperform one-on-one tutoring that includes academic support alone?" Although the general problem described in this paper may be more common when the treatment is delivered in a group environment, the problem can still exist even when the treatment is delivered at the individual level.

In addition, the issue described in this paper may stretch further, from understanding the correct causal inference that can be made from an experiment to understanding the implications of that causal inference for scale-up. Consider, for example, Mathematica Policy Research Inc.'s Early Head Start evaluation (Love et al., 2002). In this study, families were randomly assigned either to the program group, which was eligible to participate in Early Head Start programs, or to the control group, which was free to use non-Early Head Start services. Early Head Start grantees (the providers of the services) were allowed to select from among several program options (center-based, home-based, or a mixed approach). The two key goals of the study were to:

1. Understand "the extent to which the Early Head Start intervention can be effective for infants and toddlers and their low-income families" (Love et al., 2002, p. 16).
2. Understand "what kinds of programs and services can be effective for children and families with different characteristics living in varying circumstances and served by programs with varying approaches" (Love et al., 2002, p. 16-17).

When considering *the cause* of the impact estimates from the Mathematica study, it cannot accurately be described as the core components of, for example, *center-based* Early Head Start services. These core components might be loosely defined as providing all services to families through center-based child care and education, parent education, and a minimum of two home visits per year to each family (Love et al., 2002, p. xxiv). However, the true *cause* is a combination of these core components *plus* the delivery of those services by the types of people who chose and/or were selected to work at those Early Head Start centers (that is, the service deliverers). In addition, any attempts to assess whether "different program approaches have different program impacts" (Love et al., 2002, p. 71) may be influenced by the types of people working at center-based, home-based, or mixed approach Early Head Start centers.

The fact that the impacts of the core components of the different types of Early Head Start programs may be confounded with the effect of the deliverers of the services limits the conclusions one can draw from this research. For example, it could be the case that center-based programs have much more highly skilled staff than home-based programs. If so, it would be wrong to conclude that a better way to deliver services is through center-based programs (assuming that impact estimates were larger for center-based programs), as the results might be related to the type of staff delivering these programs.

Still, it can be argued that the types of staff delivering the program are, in fact, a part of the program, in which case the impact estimates do have meaning. Perhaps center-based programs tend to have a better recruitment process, and this is critical to their success. The challenge then becomes figuring out what policymakers and program administrators can do with the study's findings. If it is known that the core components of a program lead to impacts, spreading the practice is probably a good idea. However, if it is only the type of staff delivering services that leads to impacts, the challenge becomes finding out ways to recruit and retain staff that look like the program staff, a challenge that can be quite complex in a large scale-up, especially if market forces come into play.[22]

Conclusions

What is essential to the entire discussion in this paper is the need for a more complete understanding of the correct causal inference one can make from each experimental study. While random assignment enables a researcher to feel confident that differences in average outcomes between the program and control group are a result of systematic differential treatment of the two groups after random assignment, it is critical that the components of this differential treatment be understood. Once the components are understood (including the delivery system and agents), the correct causal inference to be made can be clearer. Finally, it is necessary to consider the implications and/or value of the claims that can be made once the correct causal inference is understood.

In the learning communities study example, there were three main components of the program: paired-course taking, teacher collaboration, and the teachers who ended up teaching the learning community classes. Ideally, researchers, administrators, and teachers would like to know the causal effects of the first two components of the program. However, the study design does not allow for the isolation of these effects from the teacher effect. In this situation, evaluators might want to claim that the learning community program is in fact a combination of the three components, and that the types of teachers that volunteer to teach in the program are essential to the program's success. While the impact estimate will provide an unbiased estimate

[22] For a nice example of when the selection mechanism is intended to be part of the program and thus is not necessarily a "problem" as described in this paper, see Mathematica Policy Research Inc.'s evaluation of Teach for America (Decker, Mayer, and Glazerman, 2004). In this evaluation, part of the goal of the program, Teach for America, is to recruit teachers *who otherwise would not enter the profession*. As a result, it is reasonable to consider teacher selection to be a part of the program.

of this bundled program package, it is unclear what to do with such a result. If it turns out that the program's positive impact are fully a result of the types of teachers who volunteered, then all that occurred was a rearrangement of teachers, with no real overall improvement at the school.

In general, researchers need to be cautious when designing experiments to pay careful attention to the unit of randomization as well as the level and mechanism through which the treatment is delivered. Simply randomizing a unit to experimental groups does not ensure that the causal effect researchers are attempting to measure is the one that really matters.

Appendix A

Given the work of Nye et al. and the previous estimates described in their work, in this analysis it is assumed that teachers explain 10 percent of the variation in student achievement outcomes. The proportion of variation between teachers can be described by the intraclass correlation (ICC),[23] and can be expressed as:

$$ICC = \frac{\tau^2}{\tau 2 \quad +\sigma^2} \tag{1}$$

In equation (1), τ^2 represents the amount of variation in student outcomes between classes, σ^2 represents the amount of variation in student outcomes within classes, and $\tau^2 + \sigma^2$ represents the overall variance. The assumption that 10 percent of the total variation in credits earned is explained by teachers suggests that the ICC is 0.10. The denominator in equation (1) is the overall variance in credits earned, which for the control group in the learning communities study was $(7.2)^2$ for total credits earned, as shown in Table 2.[24] Through substitution:

$$0.10 = \frac{\tau^2}{7.2^2} \tag{2}$$

and therefore:

$$\tau = 2.3 \tag{3}$$

The standard deviation of the teacher effect, τ, is therefore equal to 2.3.

[23] Typically, the ICC is a result of both teacher effects and student selection into classes. However, used here are Nye et al.'s experimental results (which are unaffected by teacher/student selection), which can be used to get an estimate of the standard deviation of the teacher effect.

[24] The program group's standard deviation was 6.9 and the pooled standard deviation was 7.04, so the choice of standard deviation has little effect on the sensitivity analysis.

References

Bauer, D. J., Sterba, S. K., & Hallfors, D. D. (2008). Evaluating Group-Based Interventions When Control Participants Are Ungrouped. *Multivariate Behavioral Research*, *43*(2), 210-236.

Bloom, D., & Sommo, C. (2005). *Building Learning Communities: Early Results from the Opening Doors Demonstration at Kingsborough Community College*: MDRC. 16 East 34th Street 19th Floor, New York, NY 10016-4326. Tel: 212-532-3200; Fax: 212-684-0832; e-mail: publications@mdrc.org; Web site: http://www.mdrc.org.

Bloom, H. S. (2006). *The Core Analytics of Randomized Experiments for Social Research. MDRC Working Papers on Research Methodology*: MDRC. 16 East 34th Street, 19th Floor, New York, NY 10016-4326. Tel: 212-532-3200; Fax: 212-684-0832; e-mail: publications@mdrc.org; Web site: http://www.mdrc.org.

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

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

Brock, T., & Harknett, K. (1998). A Comparison of Two Welfare-to-Work Case Management Models. *Social Service Review*, *72*(4), 493-520.

Brock, T., LeBlanc, A., & MacGregor, C. (2005). *Promoting Student Success in Community College and Beyond. The Opening Doors Demonstration*: MDRC. 16 East 34th Street 19th Floor, New York, NY 10016-4326. Tel: 212-532-3200; Fax: 212-684-0832; e-mail: publications@mdrc.org; Web site: http://www.mdrc.org.

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*: National Center for Postsecondary Research. Teachers College, Columbia University, Box 174, 525 West 120th Street, New York, NY 10027. Tel: 212-678-3091; Fax: 212-678-3699; e-mail: ncpr@columbia.edu; Web site: http://www.tc.columbia.edu/centers/ncpr/.

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2007). Teacher Credentials and Student Achievement: Longitudinal Analysis with Student Fixed Effects. *Economics of Education Review*, *26*(6), 673-682.

Crits-Christoph, P., & 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.

Decker, P. T., Mayer, D. P., & Glazerman, S. (2004). *The Effects of Teach For America on Students: Findings from a National Evaluation. Discussion Paper no. 1285-04*: Publications Department, Institute for Research on Poverty, 1180 Observatory Drive, Madison, WI 53706-1393. Tel: 608-262-6358; Fax: 608-265-3119; e-mail: irppubs@ssc.wisc.edu.

Frank, K. A. (2000). Impact of a Confounding Variable on a Regression Coefficient. *Sociological Methods Research*, *29*(2), 147-194.

Hedges, L. V. (2007a). Correcting a Significance Test for Clustering. *Journal of Educational and Behavioral Statistics*, *32*(2), 151-179.

Hedges, L. V. (2007b). Effect Sizes in Cluster-Randomized Designs. *Journal of Educational and Behavioral Statistics*, *32*(4), 341-370.

Hedges, L. V., & Hedberg, E. C. (2007). Intraclass Correlation Values for Planning Group-Randomized Trials in Education. *Educational Evaluation and Policy Analysis*, 2*9*(1), 60-87.

Kane, T. J., & Staiger, D. O. (2008). *Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation. NBER Working Paper No. 14607*: National Bureau of Economic Research. 1050 Massachusetts Avenue, Cambridge, MA 02138-5398. Tel: 617-588-0343; Web site: http://www.nber.org/cgi-bin/get_bars.pl?bar=pub.

Kemple, J. J. (2004). *Career Academies: Impacts on Labor Market Outcomes and Educational Attainment*. MDRC. 16 East 34th Street 19th Floor, New York, NY 10016-4326. Tel: 212-532-3200; Fax: 212-684-0832; e-mail: publications@mdrc.org; Web site: http://www.mdrc.org.

Klein, A., Starkey, P., Clements, D., Sarama, J., & Iyer, R. (2008). Effects of a Pre-Kindergarten Mathematics Intervention: A Randomized Experiment. *Journal of Research on Educational Effectiveness*, *1*(3), 155-178.

Konstantopoulos, S. (2008a). The Power of the Test for Treatment Effects in Three-Level Block Randomized Designs. *Journal of Research on Educational Effectiveness*, *1*(4), 265-288.

Konstantopoulos, S. (2008b). The Power of the Test for Treatment Effects in Three-Level Cluster Randomized Designs. *Journal of Research on Educational Effectiveness*, *1*(1), 66-88.

Krause, M. S., & Lutz, W. (2009). Process Transforms Inputs to Determine Outcomes: Therapists Are Responsible for Managing Process. *Clinical Psychology: Science and Practice*, *16,* 73-81.

Kukla-Acevedo, S. (2009). Do Teacher Characteristics Matter? New Results on the Effects of Teacher Preparation on Student Achievement. *Economics of Education Review, 28*(1), 49-57.

Lambert, M. J., & Baldwin, S. A. (2009). Some Observations on Studying Therapists Instead of Treatment Packages. *Clinical Psychology: Science and Practice, 16*, 82-85.

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

Love, J. M., Kisker, E. E., Ross, C. M., Schochet, P. Z., Brooks-Gunn, J., Paulsell, D., et al. (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*: For full text: 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., & Hedges, L. V. (2004). How Large Are Teacher Effects? E*ducational Evaluation and Policy Analysis, 26*(3), 237-257.

Pals, S. L., Murray, D. M., Alfano, C. M., Shadish, W. R., Hannan, P. J., & Baker, W. L. (2008). Individually Randomized Group Treatment Trials: A Critical Appraisal of Frequently Used Design and Analytic Approaches. *Am J Public Health, 98*(8), 1418-1424.

Raudenbush, S. W. (1997). Statistical Analysis and Optimal Design for Cluster Randomized Trials. *Psychol. Methods, 2*(2), 173-185.

Raudenbush, S. W. (2008). Advancing Educational Policy by Advancing Research on Instruction. *American Educational Research Journal, 45*(1), 206-230.

Richburg-Hayes, L., Visher, M. G., & 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., & Roberts, S. A. (2005). Design and Analysis of Clinical Trials with Clustering Effects Due to Treatment. *Clinical Trials, 2*(2), 152-162.

Rockoff, J. E. (2003). *The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data*:
For full text: http://econwpa.wustl.edu:8089/eps/pe/papers/0304/0304002.pdf.

Rockoff, J. E., Jacob, B. A., Kane, T. J., & Staiger, D. O. (2008). *Can You Recognize an Effective Teacher When You Recruit One? NBER Working Paper No. 14485*: National Bureau of Economic Research. 1050 Massachusetts Avenue, Cambridge, MA 02138-5398. Tel: 617-588-0343; Web site: http://www.nber.org/cgi-bin/get_bars.pl?bar=pub.

Rosenbaum, P. R., & Rubin, D. B. (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika, 70*(1), 41-55.

Scrivener, S., & Au, J. (2007). *Enhancing Student Services at Lorain County Community College: Early Results from the Opening Doors Demonstration in Ohio*: MDRC. 16 East 34th Street 19th Floor, New York, NY 10016-4326. Tel: 212-532-3200; Fax: 212-684-0832; e-mail: publications@mdrc.org; Web site: http://www.mdrc.org.

Scrivener, S., Bloom, D., LeBlanc, A., Paxson, C., Rouse, C. E., & Sommo, C. (2008). *A Good Start: Two-Year Effects of a Freshmen Learning Community Program at Kingsborough Community College*: MDRC. 16 East 34th Street 19th Floor, New York, NY 10016-4326. Tel: 212-532-3200; Fax: 212-684-0832; e-mail: publications@mdrc.org; Web site: http://www.mdrc.org.

Scrivener, S., & Pih, M. (2007). *Enhancing Student Services at Owens Community College: Early Results from the Opening Doors Demonstration in Ohio*: MDRC. 16 East 34th Street 19th Floor, New York, NY 10016-4326. Tel: 212-532-3200; Fax: 212-684-0832; e-mail: publications@mdrc.org; Web site: http://www.mdrc.org.

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.

Stiles, W. B. (2009). Responsiveness as an Obstacle for Psychotherapy Outcome Research: It's Worse Than You Think. *Clinical Psychology: Science and Practice, 16,* 86-91.

Visher, M. G., Wathington, H., Richburg-Hayes, L., & Schneider, E. (2008). *The Learning Communities Demonstration: Rationale, Sites, and Research Design. An NCPR Working Paper*: National Center for Postsecondary Research. Teachers College, Columbia University, Box 174, 525 West 120th Street, New York, NY 10027. Tel: 212-678-3091; Fax: 212-678-3699; e-mail: ncpr@columbia.edu; Web site: http://www.tc.columbia.edu/centers/ncpr/.

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