Abstract Title Page

Title: The Implications of Teacher Selection and Teacher Effects in Individually Randomized Group Treatment Trials

Author(s): Michael J. Weiss, MDRC

Abstract Body

Background/context:

Randomized experiments have become an increasingly popular design to evaluate the effectiveness of interventions in education (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 and control conditions in order to obtain unbiased estimates of intervention effects (H. S. Bloom, 2006; H. S. 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 (H. S. Bloom, 2006; H. S. Bloom, Richburg-Hayes, & Black, 2007; H. S. Bloom, et al., 2008; Hedges, 2007; Hedges & Hedberg, 2007; Konstantopoulos, 2008a, 2008b; Spybrook, 2008).

However, in some experimentally designed evaluations of classroom-level interventions, it is not practically feasible to randomly assign teachers 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 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 used in some education evaluations (for examples see: Love, et al., 2002; Richburg-Hayes, Visher, & Bloom, 2008; Scrivener, et al., 2008; Scrivener & Weiss, 2009; Visher, Wathington, Richburg-Hayes, & Schneider, 2008).[2] Like GRTs, in IRGTs observations within groups oftentimes are not independent; thus, appropriate analytic adjustments must be made in order to obtain accurate standard errors and 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, & Hallfors, 2008; Crits-Christoph & Mintz, 1991; Pals, et al., 2008; Roberts & Roberts, 2005). However, the implication of using the IRGT design with respect to the correct causal interpretation of impact estimates has not been well-documented – that is the focus of this research.

Purpose / objective / research question / focus of study:

This work describes how, in studies of classroom-level interventions that do not randomize teachers to experimental conditions, it will be unclear whether measured differences between program and control group students are a result of the core components of the intervention or the teachers (i.e., teacher effects). This potential confounding is a major concern if teachers are sorted into experimental conditions in such a way that they differ on variable(s)

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

that are related to their effectiveness. This work attempts to make clear the correct interpretation of typically calculated "program impacts" in this situation. In addition, using the magnitude of estimated teacher effects from prior research, this work demonstrates that if teachers are not randomly assigned to experimental conditions, then it is significantly more difficult to establish whether the intervention "works" or if the types of teachers selected to teach in intervention classrooms are simply more/less effective than their control group counterparts. The implications may be quite serious in terms of the usefulness of such studies' findings.

Setting:

This work is largely methodological / theoretical. In order to make the theoretical point more clear, and to examine the practical significance of the issue raised in this paper, a concrete example using extant data is used. The setting of the example is Kingsborough Community College (KCC), in Brooklyn, New York.

Population / Participants / Subjects:

The study participants include 1,534 community college students. For a full description of the research sample see (Scrivener, et al., 2008).

Intervention / Program / Practice:

This research is about the implications of randomly assigning students to experimental conditions when the treatment is delivered at a higher level (e.g., classrooms). The study used to exemplify this research design is the program evaluation of a one semester learning community intervention at KCC. In this study, the learning community intervention included two core components:

- Paired- or clustered-course model: Students in learning communities 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.
  (D. Bloom & Sommo, 2005; Brock, LeBlanc, & MacGregor, 2005; Scrivener, et al., 2008)

For details on the administration, duration, and implementation of the intervention, see (Scrivener, et al., 2008).

Research Design:

This work is about the research design where random assignment occurs at the student level, but the deliverers of the intervention (typically teachers) are not randomly assigned. The result of such a research design is that the intervention's deliverers become a part of the intervention. Consequently, the correct causal interpretation of the intervention's impacts is that they represent the effect of both the core components of the intervention and the intervention's deliverers. This is the design that was used in the learning community's evaluation, a case studyu of this design.

Data Collection and Analysis:

Transcript data were collected from KCC. The analysis described here focus of the intervention's impact on students' credits earned. A description of the original strategies used to analyze the data can be found at (Richburg-Hayes, et al., 2008; Scrivener, et al., 2008). The reanalysis conducted here are sensitivity analyses that describe the potential implications of teacher selection on the original impact estimates.

Findings / Results:

As shown in Table 1, original analyses of the learning community's data suggest that the program had a highly statistically significant positive impact on students' number of credits earned. 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.

The original analyses of the program's effectiveness reflect the impact of both the core components of the intervention (paired-course model and teacher collaboration), as well as the types of instructors that ended up teaching in the learning communities classes (instructors were not randomly assigned). Below, analyses are conducted to test how sensitive the findings are to the possibility that more/less effective instructors ended up teaching in the program/control group.

To assess the sensitivity of these results, an assumption must me made regarding the proportion of variation in student outcomes that are explained by teachers (i.e., the magnitude of the teacher effect). Past research suggests that teachers likely explain around 10 percent of the variation in student achievement outcomes (Nye, Konstantopoulos, & Hedges, 2004). The proportion of variation between teachers can be described by the intraclass correlation[3], and can be expressed as:

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

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 assumed ICC is 0.10. The denominator in Equation 1 is the overall variance in credits earned, which for the control group in this study was $(7.2)^2$ as shown in Table 1.[4] Through substitution we see that:

$$0.10 = \frac{\tau^2}{7.2^2} \quad (2)$$

and therefore:

$$\tau = 2.3 \quad (3)$$

The standard deviation of the teacher effect, τ, is therefore equal to 2.3. If we assume teacher effectiveness is normally distributed, then we can estimate a student's expected credits earned given the average effectiveness of her teachers. Figure 1 displays the teacher effectiveness distribution (which is assumed to be normal),

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

[4] 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.

where the x-axis represents the expected number of credits earned for students. The middle vertical line in Figure 1 shows that a student in the learning communities study with average teachers (50th percentile) can be expected to have earned around 11.0 credits during the first program semester (this is simply the overall mean number of credits earned for all students). Likewise, a student with 30th percentile teachers[5] has an expected number of credits earned of 9.8,[6] and a student with 70th percentile teachers has an expected number of credits earned of 12.1.[7] Thus, the difference in mean expected credits earned for students with 30th and 70th percentile teachers is 2.4 credits earned. Comparing 2.4 to 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 selection. 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 True impact of the core components of the program would actually be negative 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 teachers were in the 30th percentile of the teacher effectiveness distribution and the control 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 2 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 non-random 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 to 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 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 label "70P--30C"). As noted earlier, in this scenario the True 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 there is a significant teacher selection problem, it can swamp the observed impact estimates. In other words, if the assumed ICC is accurate, then teacher selection should be a serious concern because the teacher effect is quite large compared to the program's impacts.

Another way to consider

[5] 30th percentile teachers are defined as teachers 0.52 standard deviations below the mean since the probability of an observation being at least .52 standard deviations below the mean on a standard normal distribution is 30 percent.

[6] 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.

[7] Numbers may appear off due to rounding.

Figure 2 is to ask the question "how bad would teacher selection have to have been to alter the inference with regard to the program's effectiveness?"[8] In

Figure 2 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).[9] 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 the sorting of teachers), then the program's impact on credits earned is quite sensitive to the possibility of teacher effects due to teacher selection**.**

Conclusions:

What is essential to the discussion in this paper is the need for a more complete understanding of the correct causal inference one can make from certain experimental studies. While student 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 all the components of this differential treatment are understood. Once the components are understood (including the delivery system), then the correct causal inference to be made should 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 all 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 of this bundled program package, it is unclear what to do with this impact from a policy or administration perspective. **If the impact was 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.** If other schools were to implement learning communities based on these positive impacts, it would be wrong of them to expect any positive findings.

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 received. Simply randomizing a unit to experimental groups does not ensure that the causal effect we are attempting to measure is the one we are actually interested in.

---

[8] 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).

[9] Here I've chosen to use the same statistical significance threshold as reported by MDRC, using a 2-tailed test with $\alpha = .10$ (Scrivener, et al., 2008).

Appendices

Appendix A. 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., 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/.

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.

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

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

Hedges, L. V., & Hedberg, E. C. (2007). Intraclass Correlation Values for Planning Group-Randomized Trials in Education. *Educational Evaluation and Policy Analysis, 29*(1), 60-87.

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.

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.

Nye, B., Konstantopoulos, S., & 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., & 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.

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.

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., & Weiss, M. J. (2009). *More Guidance, Better Results? Three-Year Effects of an Enhanced Student Services Program at Two Community Colleges*: 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.

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/.

Appendix B. Tables and Figures

Table 1. Program Impacts from the Original Evaluation of Learning Communities

	Program Group	Control Group	Difference (Impact)		Standard Error	Control Group S.D.
# of Credits Earned	11.5	10.4	1.2	***	0.4	7.2

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 differenences between research groups. Statistical significance levels are indicated as: *** = 1 percent; ** = 5 percent; * = 10 percent.

Figure 1. The Teacher Effect

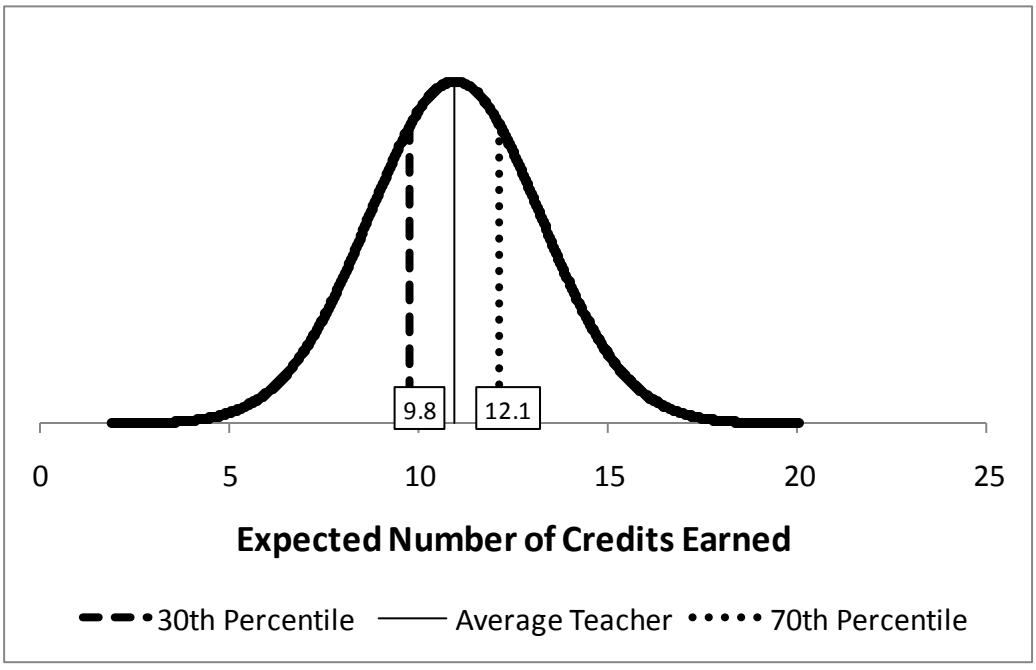

Figure 2. Sensitivity of Estimated Impacts to Teacher Selection