Abstract Title Page

Title: Using Observed Characteristics to Resolve Differences in Impact Estimates in an RCT Replication Study

Authors and Affiliations:
Jared Eno (American Institutes for Research)
Ryan Williams (University of Memphis)
Kelly Hallberg (American Institutes for Research)
Arie van der Ploeg (American Institutes for Research)
Andrew Swanlund (American Institutes for Research)
Shazia R. Miller, (American Institutes for Research)
Spyros Konstantopoulos (Michigan State University)
Abstract Body

Background / Context

The Indiana Diagnostic Assessment Tools study was a pair of randomized controlled trials of interim assessment tools that included in the design a replication of the experiment. Two separate cohorts of schools were randomly assigned to treatment and control conditions for the same intervention. Analysis of the two experiments yielded results that differed in policy-relevant ways. In the first cohort of schools, positive and significant impacts were found in mathematics. In the second cohort of schools, the effect in math was insignificant and smaller in magnitude. While the difference in estimates is not statistically significant, the two sets of results present varying differing degrees of evidence for the efficacy of the intervention that could lead policymakers to different decisions.

The mechanisms through which schools were selected for the study may have led to substantive differences between the two cohorts of schools. Each cohort was selected from among the pool of schools volunteering for the intervention in separate academic years (2009-10 and 2010-11). Thus, the schools in each RCT can be viewed as a sample of the schools that applied to implement the assessment system in each year, but not necessarily as representative of the other year or the state as a whole. To the extent that the type of schools interested in implementing the benchmark assessment system change between the two years, we would expect the samples in the two studies to differ.

Differences in the characteristics of the samples may explain the differences in estimated impacts if the treatment effect is heterogeneous. If the impact varies across subpopulations, then the representation of these subpopulations in each study sample will drive the estimated average treatment effect. Primary research focusing on variation in treatment effects across subpopulations is limited. However, some recent research has focused on issues of external validity and how to use observable characteristics to generalize impact estimates to broader or different populations (e.g., Stuart et al., 2012). In particular, new methods of weighting estimates based on propensity scores have been proposed for generalization of results from samples to populations based on observable characteristics (Tipton, 2012a).

Purpose / Objective / Research Question / Focus of Study

The purpose of this paper is to explore the degree to which differences in observable characteristics between the participants in the two separate experiments under the Indiana Diagnostic Tools study explain the differences in estimated impacts. In doing so, the paper contributes to the literature on external validity by providing an empirical examination of the extent to which treatment effect heterogeneity can be accounted for by observable characteristics commonly measured in education research.

Setting

This work is part of the Indiana Diagnostic Assessment Tools randomized control trial (RCT). The study occurred in Indiana in 2009-2010 and 2010-11.
Population / Participants / Subjects

The Indiana Diagnostic Assessment Tools study included a total of 120 schools, 57 from 2009-10 and 70 from 2010-11. These two cohorts of schools were randomized into treatment and control conditions separately (designated RCT1 and RCT2, respectively). Study schools were drawn from state lists of applicants for the intervention. The intervention became available in the 2008-09 school year, and the study took place in the 2009-10 and 2010-11 school years; as such, study schools applied for the intervention in the second and third years in which it was available. Therefore, the schools are not the earliest adopters, but are RCT1 schools are earlier adopters than RCT2 schools.

The sampling approach used to select schools into the study also differed by study year. Schools could apply for interim assessment systems at different grade levels, K-2 and grades 3 to 8. In RCT1, a stratified random sample (by urbanicity) was selected from the applicant pool of schools applying for the intervention in grades K-8. In RCT2, the pool of such schools was smaller, with far fewer schools applying for the K-2 component. Therefore, all schools applying for the intervention in grades K-8 or K-2 were included in the study, as well as a sample of schools applying for the intervention in grades 3-8. These differences could also contribute to differences in the composition of schools participating in RCT1 and RCT2.

Intervention / Program / Practice

The intervention included two interim assessment systems: Wireless Generation’s mCLASS: Reading 3D and mCLASS: Math was used in grades K–2, and CTB/McGrawHill’s Acuity was used in grades 3–8. This paper focuses on the Acuity program, as student outcomes on the state assessment are available for all students in these grades.

Research Design

Each RCT was a two-level cluster randomized design with students nested in schools. Randomization occurred at the school level—that is, schools were randomly assigned to treatment and control conditions—a conceptual match between the design and practice because the intervention was designed as a whole-school intervention.

Impact estimates in grades 3-8 were positive and statistically significant in math for RCT1. The impacts appeared to be driven largely by effects in grades 5 and 6, where there were positive, significant impacts on the order of 0.2 to 0.3 standard deviations. However, in RCT2, the impact in math was smaller and insignificant, both for grades 3-8 and grades 5 and 6 individually. We used the subclassification estimator proposed by Tipton (2012a) in an attempt to align the results from the two RCTs using observed school characteristics. Using this approach, we first estimated propensity scores for inclusion in RCT1 for study schools in both years. We then divided the RCT1 sample into equally-weighted strata according to these propensity scores, and used the same cut points to arrange RCT2 schools into the same strata. Finally, we estimated the average treatment effect for RCT1 by averaging the within-strata RCT2 impacts, weighting each within-strata impact by the representation of the strata in RCT1. Comparison of the direction and magnitude of estimates using the convention and subclassification estimators suggest the
degree to which heterogeneity of treatment effects based on observed characteristics are responsible for the differences in estimated impacts.

**Data Collection and Analysis**

Indiana Department of Education school-level data on student, teacher, and school characteristics were collected for study schools for the years of the study and prior years and used to compare the two sets of schools. Student-level data on ISTEP+ assessment scores and student demographics were also collected to estimate treatment impacts in mathematics and ELA. The analysis first examined the observed characteristics of study participants by study year. Descriptive statistics for each set of schools were examined to identify dimensions on which study schools differed.

Subclassification estimation followed the procedure outlined in Tipton (2012a). To estimate propensity scores, an indicator for participation in RCT1 was regressed on selected school characteristics using logistic regression. The characteristics included were those on which the study samples were found to differ, including school level previous achievement, racial and ethnic composition, size, grade level, and urbanicity, as well as district size. Predicted probabilities of inclusion in RCT1 were calculated using the estimated coefficients. Propensity scores for RCT1 schools were divided into three equally-weighted strata. The cut points for these strata were applied to the propensity scores for RCT2 schools, resulting in three unequally-weighted strata. Within-strata treatment effects were then estimated for RCT2 and combined according the RCT1 strata weights to calculate a predicted average treatment effect for RCT1. The variance of this estimator was calculated as the sum of the products of the variance of each within-strata estimate and the square of its strata weight. To account for the nesting structure of students within schools, two-level hierarchical linear models were used to analyze the data, controlling for student demographics.

In addition to overall impact estimation including grades 3 to 8, a grade-level analysis was used because the original results suggested that treatment effects varied by grade. This analysis focused on grade 5, as the effect in this grade also showed differences in both magnitude and significance (results for other grades generally also differed, but not always in significance).

**Findings / Results**

Descriptive analyses found substantive differences between schools in RCT1 and RCT2. On average, schools in RCT2 had lower ISTEP+ scores, higher percentages of minorities (both African Americans and Hispanics), and more students; came from larger districts; were more urban and less rural; and were more likely to have middle school grades. Differences in school composition based on free or reduced price lunch, English language learner, and special education status were smaller.

These differences were reflected in the estimated propensity scores, which showed limited overlap between the two sets of schools. Figure 1 shows a kernel density plot of the distribution of school propensity scores for inclusion in RCT1, by study year. As a result, most RCT2 schools fell into one strata, with only five or six schools in each additional stratum. However,
estimation using subclassification was still possible because every stratum contained both treatment and control schools.

Preliminary analyses did not suggest that these observed characteristics were responsible for the difference between the RCT1 and RCT2 impact estimates. In both school-level and grade 5 analyses, the differences between the estimates of the RCT1 impact using a subclassification estimator on RCT1 and RCT2 data were larger in magnitude than the differences between estimates for RCT1 and RCT2 using a conventional estimator (simple HLM). That is, weighting the RCT2 sample to more closely match the RCT1 sample on observed characteristics did not appear to reduce the difference in impact estimates. Within-strata estimates for RCT2 were consistently lower than those for RCT1, suggesting that the two school populations differed in unobserved ways that affected the treatment outcomes.

Conclusions

Preliminary analyses found that while the sample of schools in RCT1 and RCT2 showed substantial differences on some observed characteristics, accounting for these observed characteristics did not appear to explain the difference in impact estimates between study years. The results suggest that commonly-available observable characteristics such as school demographics, prior achievement, size, and grade level may not always be sufficient to explain treatment effect heterogeneity.

These findings are subject to important limitations. Comparisons between estimates from the two RCTs have less power to the degree to which schools are excluded or down-weighted in the subclassification estimator. Therefore, random variation may play a larger role, which could contribute to differences between the two RCTs’ results and limit the likelihood of aligning the estimates. In addition, this analysis examined a limited set of school characteristics and used a simple propensity score specification (e.g., no interactions or higher-order terms were included). A richer or more complex specification may capture more of the treatment effect heterogeneity. However, this simple analysis may still be relevant because the characteristics used represent variables commonly included in extant education datasets.

As focus on generalization of causal estimates grows, researchers will need to carefully consider how important reference population characteristics are included in data collection (e.g., Tipton, 2012b). Identifying potentially important variables of this type will likely be an important new area of investigation.
Appendix A. References


Appendix B. Tables and Figures

Figure 1. Kernel density plot of propensity for inclusion in RCT1, by study year