Abstract Title Page

Title: What Do We Know about the Control Condition?

Authors and Affiliations:

Shazia R. Miller    American Institutes for Research
Coby Meyers        American Institutes for Research
Arie van der Ploeg  American Institutes for Research
Spyros Konstantopoulos Michigan State University
Abstract Body

Background / Context

Randomized controlled trials (RCTs) are the best way of determining causality in social experiments (IES, 2003; Bloom, 2005; Imbens & Woolridge, 2009). Yet as strong as they are, large scale RCTs are not perfect, especially in education, and concerns have been raised about the utility of RCTs in actual policy environments (Schochet & Burghardt, 2008). This study focuses on the control condition within each of two RCTs designed to examine the impact of Indiana’s system of diagnostic assessment tools on student achievement.

Purpose / Objective / Research Question / Focus of Study

This paper is based on a larger study of the impact of Indiana’s system of diagnostic assessment tools which consisted of two, independent RCTs conducted in 2009-2010 and 2010-2011. This paper uses the results from surveys with all teachers in treatment and control schools, interviews with teachers in grades two and five, and checklists, completed every two weeks by grade 2 and 5 teachers, which report on the instructional practices used with 8 randomly selected students. This set of data allow for close examination and comparison of treatment and control schools on issues related to the implementation of the policy initiative. The paper provides insight into the question of what it means for control schools to conduct “business as usual.”

Setting

The two RCTs took place in Indiana in 2009-2010 and 2010-2011 and included K-8 public schools that had volunteered to implement diagnostic assessments in the Spring of 2009 and 2010 respectively.

Population / Participants / Subjects

Indiana schools have annually volunteered to participate in the roll out of the state’s diagnostic assessment tools intervention since spring 2008. For the full study, we randomly selected schools from the spring 2009 and spring 2010 queues for participation. Selection was limited to schools with broad grade spans and that had no prior use or familiarity with the vendor products that comprised the treatment or similar products (such as NWEA MAP or DIBELS). All schools in the queue were informed that they might be selected to the study and, if assigned to control, would be delayed for one year in implementing the diagnostic assessment tools. They were also informed of study incentives and requirements. Schools were then assigned randomly to treatment or control.

Intervention / Program / Practice:

IDOE identified two commercial products through a standard public agency request for information, request for proposals, and negotiated bidding process. Indiana selected Wireless Generation’s mCLASS:Reading 3D and mCLASS:Math as the K–2 solution and CTB/McGraw-
Hill’s *Acuity* for Grades 3–8. From IDOE’s perspective, this is a single intervention, a system of periodic diagnostic assessment tools that is *consistent*, because students throughout Indiana take the same assessments; *periodic*, because students are tested at the same time windows during the school year statewide; and *diagnostic*, because the assessments identify and report on individual learning needs.

**Research Design**

The RCT was a two-level cluster randomized design (see Boruch, Weisburd, & Berk, 2010), with students nested within schools. The school was the unit of randomization because Indiana treats the intervention as a whole-school implementation. We implemented a single-stage stratified random sampling procedure to establish a sampling pool of 70 schools for RCT 1 and an additional 70 schools for RCT 2. Power analyses determined that a minimum of 50 schools would be necessary.

**Data Collection and Analysis**

A rich set of data was used to measure program implementation and impact. These included Indiana ISTEP+ test results for students in grades three and up and Terra Nova results for K-2 students. Teachers in treatment and control conditions completed surveys about instructional practices and school conditions. In addition, administrators and teachers in Grades 2 and 5 of treatment and control schools were interviewed each spring of the study. Grade 2 and 5 treatment and control teachers also completed approximately bi-weekly online logs of instructional practices used with eight randomly selected students in their classrooms. While the results of the two studies are presented elsewhere, this paper uses the surveys, interviews, and instructional logs to explore inner workings of the RCT itself.

**Findings / Results**

First, the study was designed to compare schools using Indiana’s diagnostic assessment tools (treatment) to those who were not (control). For control schools, we used a typical business-asusual approach, allowing schools to do what they would ordinarily do with the exception of using the new Wireless and Acuity systems. We took the additional precaution of removing from our pre-randomization sample those schools who reported using similar systems (such as NWEA MAP or DIBELS) in the past. While these seemed like adequate precautions at the time, our surveys, interviews and informal fact checking later revealed that study schools, both treatment and control, were using numerous assessment systems before the study began. Published reports suggest that nationally, almost one third of education’s K–12 assessment expenditures now funds formative assessment: “everyone is doing it” (Cech, 2008, p.1), and our study suggests that Indiana was no exception (See Figure 1 for an illustration from RCT2). While a program intervention can check explicitly for use of the particular program, a policy intervention has a much broader set of antecedents, and use of an assessment system was not a new intervention for most of our study schools.

Second, the emphasis on data driven decision-making was present in both treatment and control schools during the study. While control schools generally did not report using the mCLASS or
Acuity tools, they did use a wide range of interim assessments. For instance, Renaissance Learning’s STAR Reading and STAR Math were present and used in about half the control and treatment schools (see Figure 2 for an illustration of RCT2). Moreover, most schools reported using multiple, different assessment systems during the period of the study, with often using more than 2. (See Figure 3) As such, the study requires careful clarification of what was examined, of the business-as-usual condition. In this case, results should be understood as the effect of adding the Indiana system of diagnostic assessment tools to supplement whatever was already in place in schools.

Third, our theory of action assumed that teachers would use these diagnostic assessment tools to more closely examine students’ skills and deficits, and given that information, they would differentiate instruction so that each student was more precisely taught what was most needed. However, a focus on differentiating instruction was not new to Indiana during the time of our study. Supports for differentiating instruction have become a major part of the educational landscape, common in in-school and out-of-school professional development efforts. We had assumed that given Indiana’s system, treatment teachers would differentiate more. However, we actually saw less differentiation among treatment teachers than control teachers. Exploring further, we found that 81 percent of control school teachers reported receiving training on differentiated instruction during the study year, with fully half reporting a moderate or large extent of such training. Similarly, 80 percent of control schools reported training in using assessments to guide instruction, and 73 percent reported training on ability grouping. As such, our specific year one results on improved student achievement must be re-interpreted. More broadly, we are left with a reminder to the research community that educational trends can affect moderating variables as well as final outcomes, and need to be explicitly measured and considered.

**Conclusions**

Well conducted RCTs remain researchers’ best option for providing causal evidence about whether an intervention works. However, as this study demonstrates, even the best run studies are conducted in real world settings where not all things can be controlled. This is perhaps particularly true for experiments targeting policy innovations. This paper points to the critical importance of examining business as usual, of mapping existing conditions prior to the intervention, one-off contamination during the study, and other factors that might lead to changes in intermediate and final outcomes.
Appendix A. References


Appendix B. Tables and Figures

Figure 1. Reported assessment system usage in the year prior to joining the study: 2009-10, RCT2

![Bar chart showing assessment system usage](chart1.png)

Sample size = 743.

Figure 2. Reported assessment system usage during the study: 2010-11, RCT2

![Bar chart showing assessment system usage](chart2.png)

Sample size = 743.
Figure 4. Number of assessment systems used during the study, 2010-11, RCT2

Sample size = 743.