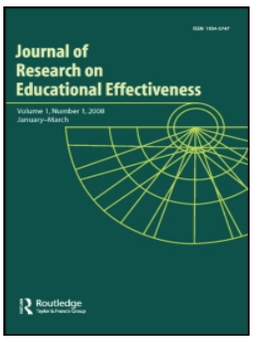

Routledge
Taylor & Francis Group

Journal of Research on Educational Effectiveness

ISSN: 1934-5747 (Print) 1934-5739 (Online) Journal homepage: https://www.tandfonline.com/loi/uree20

Effects of Interim Assessments on Student Achievement: Evidence From a Large-Scale Experiment

Spyros Konstantopoulos, Shazia R. Miller, Arie van der Ploeg & Wei Li

To cite this article: Spyros Konstantopoulos, Shazia R. Miller, Arie van der Ploeg & Wei Li (2016) Effects of Interim Assessments on Student Achievement: Evidence From a Large-Scale Experiment, Journal of Research on Educational Effectiveness, 9:sup1, 188-208, DOI: 10.1080/19345747.2015.1116031

To link to this article: https://doi.org/10.1080/19345747.2015.1116031

Accepted author version posted online: 30 Jan 2016.
Published online: 20 Jun 2016.

JOURNAL OF RESEARCH ON EDUCATIONAL EFFECTIVENESS
2016, VOL. 9, NO. S1, 188–208
http://dx.doi.org/10.1080/19345747.2015.1116031

Effects of Interim Assessments on Student Achievement: Evidence From a Large-Scale Experiment

Spyros Konstantopoulos[a], Shazia R. Miller[b], Arie van der Ploeg[b], and Wei Li[a]

ABSTRACT

We use data from a large-scale, school-level randomized experiment conducted in 2010–2011 in public schools in Indiana. Our sample includes more than 30,000 students in 70 schools. We examine the impact of two interim assessment programs (i.e., mCLASS in Grades K–2 and Acuity in Grades 3–8) on mathematics and reading achievement. Two-level models were used to capture the nesting in the data. Results indicate that the treatment effect is insignificant in Grades 3–8, and thus students in treatment schools perform as well as students in control schools. In contrast, the treatment effect is negative and significant in Grades K–2 (i.e., kindergarten and second grade), indicating that students in control schools perform higher than students in treatment schools.

KEYWORDS

field experiment
interim assessments
student achievement
IV estimation
multilevel models

Accountability mandates that were introduced nearly a decade ago have resulted in accountability systems that measure and report school and student performance annually across all states. As a result, a plethora of assessment-based school interventions targeted at improving student performance have been designed and implemented over the last 10 years (Bracey, 2005; Sawchuk, 2009). Among the assessment-based solutions offered are periodic assessments variously known as benchmark, diagnostic, or interim assessments (Perie, Marion, Gong, & Wurtzel, 2007). Typically, these assessments are administered three or four times during the school year and provide information about students' understanding of instruction. Interim assessments typically include training and instructional resources to help teachers use assessment-based evidence to make better instructional decisions that meet students' learning needs.

Interim assessments are viewed as promising levers for increasing student achievement because they provide teachers with limited objective data about student performance on a regular basis (Davidson & Frohbieter, 2011); that is, teachers can use recurrent evidence of student performance to better diagnose and monitor student learning and adjust instructional practices accordingly to maximize student learning. Interim assessments are expected to improve ongoing classroom instruction by providing useful feedback to students, teachers, and administrators. This line of reasoning has received widespread support in the literature (e.g., Datnow, Park, & Wohlstetter, 2007; Luce & Thompson, 2005; Michael & Susan Dell Foundation, 2009).

CONTACT Spyros Konstantopoulos spyros@msu.edu Michigan State University, Measurement and Quantitative Methods, 620 Farm Lane, 461 Erickson Hall, East Lansing, MI 48824, USA.
[a]Michigan State University, East Lansing, Michigan, USA
[b]American Institutes for Research, Chicago, Illinois, USA

A hypothesized mechanism is that interim assessments should lead to constructive feedback and differentiation of instruction (Tomlinson, 2000), key mediators between teaching and learning. Thus, interim assessments should help teachers identify areas of immediate instructional need for each student by providing more detailed insight on students' strengths and weaknesses. Given data from an iterative series of assessments, teachers are expected to modify instruction in ways that will improve the rate and amount of student learning. It is presumed that the closer that instruction matches students' needs and capabilities, the greater the probability of improved learning. Ongoing evidence about student performance should guide teachers' choices about which instruction is more effective. Teachers should be able to diagnose weaknesses and strive to promote positive changes in student learning by adjusting their instructional practices to what the student data at hand suggest. A nuanced explication of how teachers may be enabled to do this has been reported by Connor and colleagues (Connor, Piasta, et al., 2009; Connor et al., 2011). For example, Connor, Piasta, et al. (2009) found that first graders who received recommended amounts of individualized instruction had a stronger literacy skill growth.

There is a need to test this theory of action by rigorously evaluating the effects of schools' use of interim assessments at scale, and our study of Indiana's system of interim assessments allows for that. Specifically, we designed and conducted a large-scale cluster randomized experiment to investigate whether schools in which teachers have regular and repeated access to objective data on student progress during an academic year will produce students who perform higher on state assessments than other schools. In particular, the purpose of this study is to examine the impact of interim assessments on mathematics and reading achievement in Grades K–8 in public schools in Indiana. We used data from 70 public schools that were randomly assigned to a treatment or a control condition during the 2010–2011 school year. We evaluated the effectiveness of two interim assessment programs: mCLASS in Grades K–2 and Acuity in Grades 3–8. Two-level regression models, which capture the nesting of students within schools, were used to analyze the data.

The internal validity of our estimates should be high and should justify causal inferences about the intervention effects because we used data from a well-designed and well-executed randomized experiment. Because our sample included 70 schools from different areas of Indiana, our estimates should have higher external validity than those obtained from localized or convenience samples. Given the interest in interim assessments as levers to accelerate school improvement, the results of this study should be informative to education researchers, practitioners, and policymakers.

The article is structured as follows. First, we describe the intervention and set it in the Indiana policy context that produced it. In the second section, we situate Indiana's initiative and schools' possible responses within an overview of relevant literature. In the third section, we describe the sample and the data we used, and then we outline the analytic procedures we employed. Section four presents the results, and section five discusses our findings and their implications, as well as the limitations of our analysis.

Interim Assessment System in Indiana

In 2006, the Indiana Legislature charged the Indiana State Board of Education and the Indiana Department of Education (IDOE) with developing a long-term plan for a new assessment system that would measure student growth from year to year, and provide diagnostic information to teachers to use to improve ongoing instruction and thereby student

learning. The intention was to "encourage the advanced and gifted child, drive progress in the student who is ready, and accelerate progress for the student whose learning reflects gaps in preparation and readiness" (Indiana State Board of Education, 2006, pp. 11–12).

Indiana expected that its system of interim assessments would change teachers' instructional practices. Specifically, Indiana anticipated that interim assessments would provide teachers with more detailed knowledge about their students' academic progress and that they would use this knowledge to differentiate their instruction appropriately for individuals or groups of students. The main hypothesis was that changes in teaching practice would translate to improvements in student achievement.

IDOE began the rollout of the Diagnostic Assessment Tools in the fall of 2008. Schools and school districts volunteered to participate by signing up in early spring 2008. In summer 2008 teachers from more than 500 schools enrolling some 220,000 K–8 students began training. The state and its vendors used train-the-trainer models under which one to four teachers from each volunteering school received two to three days of training in late summer. Teachers trained in the summer received a supply of materials to train colleagues and were expected to conduct two to three training sessions at their schools during the first six months of the program. This process was repeated in each of the ensuing years until essentially all Indiana public schools participated.

Vendor and IDOE staff closely monitored training delivery to each school and provided additional support as needed. In order to retain the tools, paid for by the state, mCLASS and Acuity schools were required to provide evidence that 95% of their students participated in the interim assessments. According to IDOE, the schools in the study met this condition. Schools in the control condition did not use mCLASS or Acuity. The research team had access to year-long transaction data compiled by the web servers that both products maintained. These were used to confirm that schools and staff used the assessments and interfaces at the appropriate times to inspect results, draw down detailed reports, and view analytics for their classes and students.

However, the concept of interim assessments was not unknown to schools and teachers, which is not surprising in the era of data-driven decision making. According to a survey the researchers administered to treatment and control school teachers, prior to the study most schools were using some sort of periodic assessments to inform of student progress during the school year. These ranged from homegrown efforts to multiple commercial products. The STAR series from Renaissance Learning and NWEA's MAP were often found. Informal DIBELS-based efforts were quite common in the early grades. Schools that had histories of using technology-supported products similar to mCLASS and Acuity were excluded from the lists from which the schools of this study were drawn. Control schools continued their use of whatever products and procedures were in place. Treatment schools, with time, came to depend less on whatever preexisting methods they had in place.

Indiana chose two commercial interim assessment products for its Diagnostic Assessment Tools policy. Indiana required both vendors to align their products closely to state academic content standards. In Grades K–2, the mCLASS assessment was selected, where diagnostic probes are conducted face-to-face, and students and teachers work together. For reading and English language arts (ELA), the student performs tasks while the teacher records characteristics of the work using a personal digital assistant or tablet. Teachers are guided through the assessment process by the interface and they can immediately view results and compare them to prior performance. At any point, teachers are able to monitor individual student

progress in the classroom using short one-on-one, one-minute probes and then see those results linked to previous results graphically on the screen. The mCLASS mathematics assessments are conducted using paper and pencil, with results entered into a computer database by the teacher. This database synchronizes with district and state servers. Short delays, typically one to two days, are normal in obtaining results. Additionally, detailed individual and group reports as well as ad hoc queries are available to the classroom teacher and other authorized personnel via a web-based interface for all mCLASS assessments.

In Grades 3–8, Indiana selected CTB/McGraw-Hill's Acuity. Its assessments are 30- to 35-item multiple-choice online tests that can be completed within a class period, usually in group settings. The Acuity assessments are of two types, diagnostic or predictive, with most schools selecting the predictive tests. The former focuses on identifying specific student needs and then appropriately targeting and personalizing instruction, and the latter forecasts student performance on the Indiana state test ISTEP$^+$. Both types of Acuity permit teachers to construct practice or progress monitoring assessments from banks of aligned items. Instructional resources—packaged student exercises to practice skills or explore others—are available and may be assigned directly from Acuity's computerized report displays. Teacher access to online reports and queries is immediate.

Review of the Literature

The effects of interim assessments on student achievement have begun to be documented in the literature in recent years. Data provided by leaders, principals and teachers who have implemented interim assessments suggest that by and large the value of such assessments seems to be in establishing a link between district policy and instructional practices. Teacher focus group and survey data have indicated that teachers believe that interim assessments can be useful in redesigning lessons, modifying instruction, and preparing students for standardized testing (Clune & White, 2008). Of course to achieve meaningful instructional change, teachers need additional skills and knowledge; that is, interim assessments can identify areas for improvement, but it is school staff who must use that information appropriately and implement instructional practices that will improve student performance (Blanc et al., 2010). Teacher interview data has suggested that teachers use interim assessments to gain information about their students' learning, but they do not typically use it to assess students' conceptual understanding (Goertz, Nabors-Olah, & Riggan, 2010). Empirical evidence on the nature, form, and content of the instructional changes teachers make in response to information gleaned from interim assessments is exceedingly sparse, and it is not clear that the changes made are always beneficial (Booher-Jennings, 2005; Clune & White, 2008; Plank & Condliffe, 2013; Shepard, Davidson & Bowman, 2011).

The effects of interim assessments have also been documented in experimental studies and the findings have been overall mixed. For instance, through a large-scale cluster randomized experiment, May and Robinson (2007) evaluated Ohio's Personalized Assessment Reporting System (PARS) for the Ohio Graduation Tests (OGT). PARS featured repeated assessment opportunities and provided reports on test outcomes and training for test data users. The authors compared 10th-grade student achievement between 51 treatment and 49 control schools during the pilot year and found that the impact of the first year of PARS on student achievement was not significant.

The impact of the work of the Center for Data-Driven Reform in Education (CDDRE) on student achievement was examined in a recent study (Carlson, Borman, & Robinson, 2011). The CDDRE intervention is a decision-making process that emphasizes instructional change based on interim assessment results (the CDDRE-designed 4Sight assessments) and other inputs. The authors analyzed data from a multistate district-level cluster randomized experiment to investigate the potential benefits of CDDRE. The results of the first year of the experiment indicated significant positive effects on mathematics scores, but the positive effects on reading scores did not reach statistical significance.

Follow-up studies have investigated the impact of CDDRE over a four-year period (Slavin, Cheung, Holmes, Madden, & Chamberlain, 2013). More than 600 elementary and middle schools were included in the analysis. Multilevel models were used to analyze the impact of CDDRE on the fifth- and eighth-grade mathematics and reading achievement. The results of the multilevel analysis showed significant positive effects on both mathematics and reading and grade levels in the fourth year. The effects were more pronounced in reading in elementary schools (see Slavin et al., 2013).

The impact of the Measures of Academic Progress (MAP) assessments on reading achievement has also been examined recently (Cordray, Pion, Brandt, & Molefe, 2012). Thirty-two elementary schools from five Illinois districts were randomly assigned to treatment and control conditions at Grades 4 and 5. The study found that MAP was implemented with moderate fidelity, but MAP teachers were not more likely to differentiate instruction than their non-MAP colleagues. The researchers found no statistically significant differences at either grade in reading achievement on the Illinois State Achievement Test (ISAT) or on the MAP composite score.

A more recent study also examined the impact of interim assessment programs on mathematics and reading achievement in Indiana (Konstantopoulos, Miller, & van der Ploeg, 2013). Results indicated that the treatment effects are positive, but not consistently significant. The treatment effects were smaller in lower grades (i.e., kindergarten to second grade), and larger in upper grades (i.e., third to eighth grade). Significant treatment effects were detected in Grades 3–8, especially in third- and fourth-grade reading and in fifth- and sixth-grade mathematics.

Methodology

Sample

A large-scale experiment was conducted in Indiana during the 2010–2011 academic year and included K–8 public schools that had volunteered to participate in the intervention in the spring of 2010. The design was a two-level cluster randomized design (see Boruch, Weisburd, & Berk, 2010). Students were nested within schools, and schools were nested within treatment and control groups. Schools were randomly assigned to a treatment or a control condition.

From a list of 157 schools that had volunteered in the spring of 2010 to participate in the interim assessments program in Indiana during the 2010–2011 school year, we first randomly selected a sample of 70 schools. Second, we randomly assigned these 70 schools to treatment and control conditions following a balanced design. Specifically, 36 schools were assigned to the treatment condition, and 34 schools were assigned to the control condition.

This is the initial sample of schools that were part of the random assignment process (random assignment sample). However, due to a nearly 20% attrition of schools because of school closures, refusal to participate in the experiment, or use of similar assessments in the recent past, the sample of schools that participated in the experiment (e.g., schools assigned to treatment group actually receiving treatment) was reduced to 55 (28 schools in the treatment condition and 27 schools in the control condition). This is the sample of schools that participated in the experiment.

Variables

Indiana state test (ISTEP+) scores were used as outcomes in Grades 3–8. Indiana does not offer a state test in Grades K–2, and thus we administered a Terra Nova test and used these scores as outcomes in Grades K–2. To put mathematics or reading scores across grades into a comparable metric, we standardized students' scores within each grade level (i.e., created z scores). The main independent variable denoted whether a school was assigned to the treatment (i.e., mCLASS, Acuity) or not. The treatment variable was coded as a dummy variable (one for schools assigned to, or which received, mCLASS or Acuity and zero otherwise); that is, the coefficient of the treatment is a standardized mean difference between treatment and control groups. The student level covariates included gender (a binary indicator for female students—male students being the reference category), age, race (multiple binary indicators for Black, Latino, and other race students—White students being the reference category), low SES that represents economic disadvantage (a binary indicator for free or reduced-price lunch eligibility—no eligibility being the reference category), special education status (a binary indicator for special education students—no special education status being the reference category), and limited English proficiency (LEP) status (a binary indicator for students with limited English proficiency—English proficiency being the reference category). The school level covariates were percent of female, minority, lower SES (eligible for free or reduced-price lunch), special education, and limited English proficiency students, as well as school urbanization categories (multiple binary indicators for rural, suburban, and small town—urban being the reference category). The school composition variables were aggregate measures of student covariates and were constructed by computing school specific means of student covariates.

Statistical Analyses

We used student and school data in Grades K–8. First, we conducted intention to treat (ITT) analyses using data from the initial sample of schools and students that were assigned to treatment or control groups regardless of participation in the experiment. The treatment is represented by a dummy that takes the value of one if a school is assigned randomly to the treatment group and zero otherwise (ITT treatment dummy). The data were analyzed as was intended by random assignment regardless of whether schools participated in the study. All these schools provided data on student achievement and covariates. This analysis should produce causal estimates of treatment effects because it capitalizes on random assignment of schools to treatment or control conditions. However, a potential caveat of the ITT analysis ignores information about schools that actually participated in the experiment (e.g., schools

randomly assigned to treatment actually received treatment). This analysis used school and student data from 70 schools and nearly 30,000 students.

Second, we conducted sensitivity analyses using data from the sample of participating schools and students (i.e., treatment on the treated analyses or TOT). The treatment is represented in this case by a dummy that takes the value of one if a school assigned randomly to the treatment group actually received the treatment and zero otherwise (TOT treatment dummy). The data included schools (and their students) that have stayed in the experiment and have complied with random assignment. This analysis used school and student data from 55 schools and nearly 20,000 students.

A potential caveat of analyses of the TOT sample of schools is that they may produce biased estimates of the treatment effect because of selection bias. Specifically, in our study 15 schools overall (seven control schools and eight treatment schools) did not participate in the experiment. If this attrition was differential (i.e., treatment schools that left the experiment are different on average than control schools that left the experiment), the treatment and control schools that remained in the experiment may not be similar. The analyses of the TOT sample of schools do not include schools that did not participate in the experiment, and thus the treatment effect may be positively or negatively affected.

Thus, to rectify this potential threat to the internal validity of the treatment effect, we used an instrumental variables (IV) approach. Specifically, we used random assignment in the initial sample of schools (i.e., ITT treatment) as an instrument to "clean" the TOT treatment and facilitate causal inferences for the TOT analyses (see Angrist & Pischke, 2009; Wooldridge, 2010). The key variable in this methodological procedure is the instrument, which should be orthogonal to selection processes in order to be valid. In experiments, it is straightforward to select an instrument because of the random assignment process; that is, ITT treatment, the random assignment in the initial sample of assigned schools (i.e., 70 schools), can be used to remove error from the TOT treatment and address possible selection bias. This analysis used school and student data from 70 schools and nearly 30,000 students.

To capture potential dependencies in the data (i.e., students nested within schools) we used two-level models with students at the first level and schools at the second level. Schools were treated as random effects, and the between-school variance of these random effects indicated differences in average mathematics or reading achievement across schools. First, we conducted several analyses using data across all grades (i.e., K–8). Second, we conducted separate analyses using Grade K–2 data only or Grade 3–8 data only to examine the effects of mCLASS or Acuity, respectively. Third, we also conducted single-grade analyses for each grade separately to gauge grade-specific effects. We conducted ITT, TOT, and IV analyses for each grade separately. Finally, in Grades 4–8 we also estimated treatment effects controlling for prior ISTEP+ achievement.

Models

The across-grade (i.e., K–8, K–2, 3–8) ITT or TOT analyses involved regressing mathematics or reading scores on the treatment variable and other student and school covariates. In particular, student achievement is a function of the treatment, student and school covariates,

and grade effects, namely

$$Y = f(Treatment,\ Student,\ School,\ Grade,\ error), \tag{1}$$

where Y is the outcome (mathematics or reading scores) and the error has a student and a school component, namely $error = (student,\ school)$. Details of this model are reported in the Appendix.

The across-grade IV analysis involved two steps. The outcome variable in the first step is a treatment dummy coded as one if a school stayed in the experiment and received the treatment, and zero otherwise (received treatment or RT dummy). That is, the 28 treatment schools and their students that received the treatment were coded as one, and the remaining 42 schools and their students were coded as zero. The 42 schools in the control group (zero category) included the 15 schools that did not participate in the experiment, but were part of the initial sample of 70 schools. This outcome variable (RT dummy) was regressed on initial random assignment (ITT treatment dummy) and all other covariates described in Equation (1). The outcome variable in this regression is represented by RT and the main independent variable is represented by $ITTTr$ (see Equation 2). It is common practice to use OLS in this first step because this step focuses only on the fitted or predicted values of the regression and not on the error term (Wooldridge, 2010). In fact, the error of the RT dummy is detached in this process. In the first step, the RT dummy is a function of the ITT treatment dummy, student and school covariates, and grade effects

$$RT = f(ITT\ Treatment,\ Student,\ School,\ Grade,\ error). \tag{2}$$

Details of this model are reported in the Appendix. The predicted values of the regression model in Equation (2) (i.e., the prediction part of the regression) were used as the new main independent variable in Equation (3). Thus, in the second step, student achievement is a function of the predicted values, student and school covariates, and grade effects, namely

$$Y = f(Predicted\ Values,\ Student,\ School,\ Grade,\ error). \tag{3}$$

Details of this model are reported in the Appendix.

Results

We initially checked whether random assignment was successful using observed school characteristics. This analysis included 70 schools for Grades K–8 or 3–8 analyses and 42 schools for Grade K–2 analyses. The goal was to identify potential observed variables where random assignment may not have been as successful as intended by the design. Specifically, we used t tests for independent samples to determine whether significant differences existed between the treatment and control for several school-level observed variables, including proportion female, minority, disadvantaged, special education, and limited English proficiency students, as well as prior school achievement. The results of this analysis are reported on Table 1. Specifically, Table 1 reports mean differences, their standard errors, p values of the t tests, and effect sizes. We used t tests for independent samples to determine whether significant differences existed between the two conditions for several school-level observed variables,

Table 1. Random assignment check using observed variables.

Variable	M_d	SE_d	*P* value	ES
Grades K–2: 42 Schools				
Proportion of Female Students	−0.042	0.029	0.167	−0.505
Proportion of Minority Students	−0.036	0.107	0.737	−0.103
Proportion of Economically Disadvantaged Students	0.068	0.072	0.357	0.288
Proportion of Special Education Students	0.013	0.017	0.433	0.241
Proportion of Limited English Proficiency Students	0.005	0.025	0.829	0.070
Grades 3–8: 70 Schools				
Proportion of Female Students	−0.017	0.019	0.372	−0.219
Proportion of Minority Students	0.000	0.085	0.998	−0.001
Proportion of Economically Disadvantaged Students	0.032	0.054	0.556	0.140
Proportion of Special Education Students	0.004	0.021	0.845	0.047
Proportion of Limited English Proficiency Students	0.001	0.015	0.926	0.022
Spring 2010 ISTEP+ Mathematics Scores	−9.350	9.500	0.328	−0.232
Spring 2010 ISTEP+ ELA Scores	−7.251	7.533	0.339	−0.227

Note. M_d = Difference between treatment and control group school means; SE_d = Standard error of the mean difference; ES = Effect size reported is Hedges's *g*.

including proportion female, minority, disadvantaged, special education, and limited English proficiency students, as well as prior school achievement. In this analysis, we also took into account the What Works Clearinghouse (WWC) standards about baseline equivalence of observed variables expressed as effect size estimates. The recommended effect size estimate is Hedges's *g*.

Overall, the results suggested that random assignment was successful (i.e., no systematic significant differences were detected between treatment and control with respect to observed school variables). The mean differences were consistently smaller than their standard errors, and thus all *p* values of the *t* tests were greater than 0.15. However, the effect size estimates for proportion of female students and proportion of economically disadvantaged students in Grades K–2 were 0.505 and 0.288, respectively, which, according to WWC standards, do not meet baseline equivalence. All other effect size estimates were overwhelmingly smaller than 0.25. All of these school variables were included in the regression models as statistical controls.

Then, we computed descriptive statistics, which are summarized in Table 2. Specifically, Table 2 reports sample sizes, means, and standard deviations for variables of interest. Forty-nine percent of students were females, 53% of the students were White, and 57% of the students were eligible for free or reduced-price lunch. The average student age was nearly 11 years. Nearly 25,000 students had ISTEP+ scores, and over 6,000 students had Terra Nova scores.

Estimates Across Grades

The treatment effect estimates produced by the ITT, TOT, and IV analyses are reported in Table 3 in the upper, middle, and lower panels, respectively. Because the outcomes are standardized and the treatment variable is a dummy, the coefficients of the treatment are standardized mean differences (i.e., mean differences in standard deviation units between treatment and control groups). Positive estimates indicate a positive treatment effect and negative estimates indicate higher performance in control schools. First, we report results of the ITT analyses shown in the upper panel of Table 3. The mathematics or reading estimates

Table 2. Descriptive statistics of variables of interest.

	N	Mean	SD
Terra Nova Reading Scores	6270	575.96	55.07
Terra Nova Mathematics Scores	6249	540.36	56.94
ISTEP+ ELA Scores	24743	503.96	64.80
ISTEP+ Mathematics Score	24868	519.19	76.60
Age (months)	36063	129.16	33.35
Female	36087	0.49	0.50
Race			
White	36027	0.53	0.50
Black	36027	0.27	0.45
Latino	36027	0.12	0.32
Other	36027	0.08	0.28
Limited English Proficiency	37559	0.04	0.20
Low SES: Free- or Reduced-Price Lunch	36102	0.57	0.50
Special Education	37559	0.19	0.39

Note. SD = standard deviation.

obtained from the Grade K–8 analysis were close to zero and statistically insignificant at the 0.05 level. The mathematics or reading estimates obtained from the Grade 3–8 analysis were very similar; that is, they were small and statistically insignificant at the 0.05 level. In contrast, the mathematics or reading estimates obtained from the Grade K–2 analysis were negative, greater than 0.20 standard deviations, and statistically significant at the 0.05 level. The K–2 results are surprising because mCLASS was designed to improve student performance.

Second, we report results of the TOT sensitivity/robustness analyses shown in the middle panel of Table 3. The TOT analyses provided estimates for a smaller number of schools that participated in the study and therefore some selection bias is possible. Nonetheless, results were, by and large, similar to the ITT estimates. All treatment effect estimates were small and statistically insignificant at the 0.05 level in Grades K–8 and 3–8. The estimates from the Grade K–2 analyses were negative and statistically significant at the 0.05 level both for mathematics and reading. In sum, the ITT and TOT estimates point to no treatment effects for Acuity and negative and significant treatment effects for mCLASS.

Third, we present results of the second stage IV analyses (Equation 3) shown in the lower panel of Table 3. These results are very similar to the ITT and TOT results. The mathematics or reading estimates obtained from the Grade K–8 analysis were small—close to zero and statistically insignificant at the 0.05 level. The mathematics or reading estimates obtained from the Grade 3–8 analysis were qualitatively similar and statistically insignificant at the 0.05 level. The mathematics or reading estimates obtained from the Grade K–2 analysis were identical to the ITT estimates.

In multilevel models, variance component estimates of the random effects (i.e., error terms) at different levels of the hierarchy are also informative (see Dedrick et al., 2009). Hence, we summarized estimates of the residual variance components of the random effects at the first and second levels in Table 4. Estimates produced from the ITT analyses are reported in the left panel of Table 4. All residual variances at the first or second levels were significantly different than zero. As expected, the largest proportion of the total residual variance in the outcomes was at the first level. Specifically, more than 80% of the total residual variance was at the student level. The residual variances of the second-level errors were also significant, but were typically smaller than 10% of the total residual variance. Still, the

Table 3. Estimates of treatment effects in mathematics and reading across grades.

	Mathematics		Reading	
	Estimate	SE	Estimate	SE
ITT Analysis				
Grades K–8				
Treatment	0.008	0.053	−0.034	0.045
Grades K–2				
Treatment	−0.221*	0.063	−0.194*	0.054
Grades 3–8				
Treatment	0.043	0.064	−0.013	0.054
TOT Analysis				
Grades K–8				
Treatment	−0.035	0.063	−0.059	0.054
Grades K–2				
Treatment	−0.197*	0.072	−0.205*	0.063
Grades 3–8				
Treatment	−0.006	0.082	−0.039	0.068
IV Analysis				
Grades K–8				
Treatment	0.011	0.074	−0.047	0.062
Grades K–2				
Treatment	−0.221*	0.063	−0.194*	0.055
Grades 3–8				
Treatment	0.065	0.096	−0.020	0.081

*$p < 0.05$; SE = standard error.

second-level variance estimates indicate significant variability in achievement between schools (i.e., school differences in achievement).

Estimates produced from the TOT analyses are reported in the right panel of Table 4. The results are overall similar to those produced from the ITT analyses. All residual variances at the first or second levels were significantly different than zero. Most of the residual variance in the outcomes was at the first level. The second-level residual variances of the random effects were typically smaller than 10% of the total residual variance, but still indicated significant differences between schools in achievement.

A key condition in the IV analysis is that the instrument (ITT treatment) should be significantly related to the *RT* dummy. This condition can be checked through the first stage regression model (see Equation 2). In particular, for this condition to hold, the coefficient of *ITTTr* in Equation (2) needs to be significantly different than zero. Then the instrument is strong; otherwise, marginal significance indicates that the instrument may be weak. When instruments are weak, the IV estimates may be unreliable (see Stock, Wright, & Yogo, 2002). The *t* statistic of the regression coefficient of the instrument (β_1 in Equation 2) should be greater than 3.20 and significant in order to have a strong instrument. This *t* statistic tests the null hypothesis of a zero coefficient for the instrument. The results of the first-stage regression (Equation 2) are reported in Table 5 and indicate overall that the instrument was strongly and significantly related to *RT*. In particular, all coefficients were positive and significant at 0.05, which shows that the instrument was significantly related to *RT*. In addition, all *t* statistics were greater than five, which further supports that the instrument is relevant and strong. Notice that in Table 5 we did not report results for Grades K–2 because in these three grades the *RT* and the instrument are the same (i.e., perfect correlation). This means that in Grades K–2, the ITT and the IV results are identical.

Table 4. Variance component estimates and standard errors.

	ITT				TOT			
	Mathematics		Reading		Mathematics		Reading	
	Estimate	*SE*	Estimate	*SE*	Estimate	*SE*	Estimate	*SE*
Grades K to 8								
Level-1 Variance	0.735*	0.006	0.742*	0.006	0.736*	0.008	0.739*	0.008
Level-2 Variance	0.043*	0.008	0.029*	0.006	0.042*	0.009	0.031*	0.007
Grades K to 2								
Level-1 Variance	0.828*	0.015	0.796*	0.014	0.830*	0.016	0.797*	0.015
Level-2 Variance	0.015*	0.006	0.010*	0.005	0.017*	0.007	0.013*	0.006
Grades 3 to 8								
Level-1 Variance	0.696*	0.006	0.718*	0.007	0.678*	0.008	0.703*	0.009
Level-2 Variance	0.061*	0.012	0.043*	0.009	0.068*	0.016	0.045*	0.011
Grade K								
Level-1 Variance	0.802*	0.027	0.765*	0.025	0.801*	0.027	0.763*	0.026
Level-2 Variance	0.095*	0.043	0.060*	0.029	0.101*	0.047	0.067*	0.033
Grade 1								
Level-1 Variance	0.802*	0.025	0.785*	0.024	0.796*	0.026	0.779*	0.025
Level-2 Variance	0.040*	0.018	0.040*	0.019	0.047*	0.022	0.040*	0.020
Grade 2								
Level-1 Variance	0.784*	0.024	0.766*	0.023	0.792*	0.025	0.771*	0.025
Level-2 Variance	0.037*	0.016	0.016*	0.010	0.037*	0.017	0.021*	0.012
Grade 3								
Level-1 Variance	0.777*	0.019	0.756*	0.018	0.773*	0.023	0.742*	0.022
Level-2 Variance	0.040*	0.013	0.031*	0.010	0.025*	0.012	0.018*	0.009
Grade 4								
Level-1 Variance	0.753*	0.017	0.766*	0.018	0.743*	0.022	0.744*	0.022
Level-2 Variance	0.070*	0.021	0.047*	0.015	0.072*	0.026	0.047*	0.019
Grade 5								
Level-1 Variance	0.625*	0.017	0.714*	0.019	0.658*	0.023	0.722*	0.026
Level-2 Variance	0.110*	0.039	0.038*	0.015	0.152*	0.069	0.054*	0.028
Grade 6								
Level-1 Variance	0.713*	0.016	0.722*	0.017	0.743*	0.023	0.743*	0.023
Level-2 Variance	0.061*	0.023	0.069*	0.025	0.093*	0.043	0.088*	0.039
Grade 7								
Level-1 Variance	0.651*	0.013	0.660*	0.014	0.578*	0.016	0.628*	0.017
Level-2 Variance	0.037*	0.016	0.021*	0.010	0.056*	0.030	0.035*	0.019
Grade 8								
Level-1 Variance	0.602*	0.012	0.647*	0.013	0.533*	0.015	0.611*	0.017
Level-2 Variance	0.053*	0.021	0.039*	0.016	0.072*	0.036	0.052*	0.027

*$p < 0.05$

Estimates Within Grades

We also conducted analyses within each grade separately (i.e., kindergarten through eighth grade) to determine grade-specific treatment effects. Again, all treatment estimates are standardized mean differences. First, we present results from the ITT analyses summarized in Table 6. In Grades 3–8, the majority of the treatment effect estimates was positive, except for some estimates in fourth and eighth grades. The estimates were typically close to zero and no estimate reached statistical significance at the 0.05 level. The estimates in fifth-grade mathematics, however, were positive and greater than 0.20 standard deviations. These effects are similar in magnitude to class-size effects reported in Project STAR studies (e.g., Nye, Hedges, & Konstantopoulos, 2000), which may indicate meaningful effects. In Grades K–2, all estimates of the treatment effect were negative, and in kindergarten and second-grade reading the estimates were statistically significant at the 0.05 level. In kindergarten, in

Table 5. First stage regression estimates, standard errors, and *T* tests.

	Mathematics			Reading		
	Estimate	*SE*	*T* test	Estimate	*SE*	*T* test
Grades K to 8						
Instrument Coefficient	0.723*	0.074	9.746	0.724*	0.074	9.782
Grades 3 to 8						
Instrument Coefficient	0.662*	0.083	7.978	0.664*	0.083	7.989
Grade 3						
Instrument Coefficient	0.684*	0.101	6.768	0.691*	0.100	6.896
Grade 4						
Instrument Coefficient	0.675*	0.102	6.643	0.677*	0.102	6.670
Grade 5						
Instrument Coefficient	0.552*	0.097	5.712	0.554*	0.097	5.737
Grade 6						
Instrument Coefficient	0.760*	0.113	6.699	0.762*	0.113	6.719
Grade 7						
Instrument Coefficient	0.641*	0.113	5.669	0.638*	0.114	5.584
Grade 8						
Instrument Coefficient	0.650*	0.115	5.633	0.649*	0.115	5.632

*$p < 0.05$

particular, the magnitude of the estimates was nearly 0.33 standard deviations, and in second grade the estimate in reading was just smaller than 0.20 standard deviations.

Table 7 summarizes the results from the TOT analyses for Grades K–8. In Grades K–2, the estimates of the treatment effect were negative. In kindergarten and second grade some of the estimates were statistically significant. In kindergarten, in particular, the estimates were large and nearly 0.33 standard deviations. The estimates were typically smaller in Grades 3–8 and all treatment effect estimates were statistically insignificant at the 0.05 level. Generally, the grade-specific analyses suggested no effects of Acuity on ISTEP+ scores, and negative effects of mCLASS on Terra Nova scores, especially in kindergarten and to some degree in second grade.

Table 6. Estimates of treatment effects in mathematics and reading by grade: ITT analysis.

	Mathematics		Reading	
	Estimate	*SE*	Estimate	*SE*
Kindergarten				
Treatment	−0.341*	0.159	−0.308*	0.129
Grade 1				
Treatment	−0.171	0.104	−0.106	0.104
Grade 2				
Treatment	−0.181	0.101	−0.192*	0.075
Grade 3				
Treatment	0.151	0.076	0.103	0.069
Grade 4				
Treatment	−0.014	0.095	−0.030	0.080
Grade 5				
Treatment	0.240	0.143	0.148	0.090
Grade 6				
Treatment	0.157	0.115	0.022	0.122
Grade 7				
Treatment	0.050	0.091	0.057	0.071
Grade 8				
Treatment	−0.004	0.106	0.009	0.092

*$p < 0.05$; *SE* = Standard error

Table 7. Estimates of treatment effects in mathematics and reading by grade: TOT analysis.

	Mathematics		Reading	
	Estimate	*SE*	Estimate	*SE*
Kindergarten				
Treatment	−0.304	0.174	−0.304	0.144
Grade 1				
Treatment	−0.168	0.120	−0.141	0.112
Grade 2				
Treatment	−0.138	0.109	−0.178	0.088
Grade 3				
Treatment	0.150	0.084	0.102	0.075
Grade 4				
Treatment	−0.096	0.127	−0.108	0.106
Grade 5				
Treatment	0.169	0.234	0.128	0.147
Grade 6				
Treatment	0.246	0.170	0.115	0.166
Grade 7				
Treatment	0.109	0.163	0.164	0.133
Grade 8				
Treatment	0.131	0.179	0.098	0.155

*$p < 0.05$; SE = Standard error

The results from the IV analyses are summarized in Table 8. These results are overall qualitatively similar to those reported in Tables 6 and 7. In Grades 3–8, all estimates were statistically insignificant at the 0.05 level. The estimates in fifth grade, both in mathematics and reading, were positive and typically greater than 0.25 standard deviations, which may indicate meaningful effects. The Grade 6 estimate in mathematics was also positive and nearly 0.20 standard deviations. The estimates in Grades K–2 were identical to those in

Table 8. Estimates of treatment effects in mathematics and reading by grade: IV analysis.

	Mathematics		Reading	
	Estimate	*SE*	Estimate	*SE*
Kindergarten				
Treatment	−0.341*	0.159	−0.308*	0.129
Grade 1				
Treatment	−0.171	0.104	−0.106	0.104
Grade 2				
Treatment	−0.181	0.101	−0.192*	0.075
Grade 3				
Treatment	0.221	0.111	0.149	0.100
Grade 4				
Treatment	−0.020	0.141	−0.045	0.119
Grade 5				
Treatment	0.435	0.262	0.267	0.164
Grade 6				
Treatment	0.207	0.152	0.029	0.160
Grade 7				
Treatment	0.078	0.142	0.090	0.112
Grade 8				
Treatment	−0.005	0.163	0.013	0.142

*$p < 0.05$; SE = Standard error

Table 6; that is, the estimates were negative, considerable, and, in kindergarten and second grade, statistically significant at the 0.05 level.

Finally, we ran sensitivity analyses in Grades 4–8, where we added prior ISTEP+ scores as another covariate in our linear models. The results of these analyses were, by and large, very similar to those reported in Tables 6, 7, and 8. All estimates were statistically insignificant in Grades 4–8, and there was no indication that Acuity impacted ISTEP+ scores. For simplicity, we do not report these results in tables.

Discussion

We examined the effects of interim assessments on student achievement using data from a large-scale experiment. Our study sheds more light on the effects of interim assessments on student learning, a timely issue in the United States where assessments are a popular school improvement strategy. Overall, the findings of the across-grade analyses are mixed. The K–8 ITT results suggest that the treatment effect is not statistically significant across grades. In Grades 3–8, we observe insignificant ITT effects both in mathematics and reading. The estimates were typically close to zero. By contrast, the findings in Grades K–2 are negative and typically statistically significant both in mathematics and reading. This finding is puzzling, and the magnitude of the negative effects is somewhat concerning because it suggests a disadvantage in early grades for students in schools that received the treatment.

In Grades K–8, the TOT results were similar to the ITT results, indicating no treatment effects. Similarly, in Grades 3–8 the treatment effects were insignificant. The effects were negative and significant, however, in Grades K–2. Finally, the findings of the IV analyses pointed to similar treatment effects; that is, overall the estimates are robust. Taken together, the results consistently show statistically significant but negative effects in Grades K–2 and statistically insignificant findings in later grades (i.e., 3–8).

The single-grade analyses yielded some interesting findings. For example, the fifth-grade ITT and IV results suggest positive but insignificant effects that are greater than 0.20 standard deviations in mathematics. The magnitude of these estimates is important because prior work has reported that the average annual gain from nationally normed tests in mathematics in fifth grade is nearly 0.40 standard deviations (see Hill, Bloom, Black, & Lipsey, 2008). This finding partially supports previous findings by Konstantopoulos et al. (2013), who found significant positive effects of interim assessments on mathematics achievement in fifth grade.

In kindergarten and second grade, the results revealed that students in treatment schools were at a disadvantage compared to students in control schools. The kindergarten ITT and IV estimates in particular, both in mathematics and reading, were negative, significant, and greater than 0.30 standard deviations. This negative effect in kindergarten is alarming because according to Hill et al. (2008), the average annual gain from nationally normed tests in mathematics in kindergarten is slightly greater than one standard deviation. The ITT and IV estimates in second-grade reading were also negative and significant and slightly smaller than 0.20 standard deviations. By comparison, the average annual gain from nationally normed tests in reading in second grade is nearly 0.60 standard deviations (see Hill et al., 2008). In both cases it seems that the disadvantage is nearly one third of the average annual gain and indicates a critical setback in mathematics and reading achievement. These findings are not in agreement with previous work that had reported the impact of interim

assessments in early grades as zero and insignificant (Konstantopoulos et al., 2013). In contrast, evidence from the vendor claims positive effects for mCLASS in mathematics in Indiana (Wang & Gushta, 2013).

The negative and significant estimates of mCLASS on student achievement deviate from what was expected. It is worrisome that these estimates were considerable in magnitude and sometimes even greater than 0.30 standard deviations. It is challenging to interpret the findings in Grades K–2 given the expectation of vendors and Indiana's leadership that these carefully selected interim assessment programs would improve student achievement.

Indiana, like most states, uses no statewide end-of-year accountability test in Grades K–2. The study administered the Terra Nova test in these three grades. The Terra Nova is designed and scored by CTB McGraw Hill, which also designs and scores the ISTEP+. There is considerable overlap in content coverage and format between the two tests, although the Terra Nova test is not explicitly aligned with Indiana's academic content standards. And, possibly more important, there were no accountability repercussions for schools from the Terra Nova results, unlike for ISTEP+ results in Grades 3–8. Student and staff motivation to do well on the Terra Nova was therefore likely to be minimal, both in control and treatment conditions. If mCLASS had a positive effect on the instruction that the Terra Nova test could capture, it should have been apparent in our results, but the evidence points to the contrary.

It may be argued in addition that if Terra Nova miss-measures mCLASS effects, then any effects the program may have will be underestimated. That interpretation, however, is not consistent with our results. Something about mCLASS or the manner in which schools worked with it depressed the Terra Nova results. To the extent that the Terra Nova scores reasonably measure K–2 performance, the results suggest that the mCLASS interim assessment was not a good strategy for improving student performance in the early grades in these schools at this time in Indiana. No intervention is without trade-offs, and it could be that the amount of time lost during the individually administered K–2 assessments is of more value to teachers than the information gained. It is also possible that teachers in the early grades are already deeply attuned to individual students' levels of knowledge and gain little from a formal test or a new assessment program that may interfere with their everyday routine. Last, if teachers misuse or misinterpret mCLASS results, or if they are unclear about how to modify instruction based on the signals mCLASS emits, then the probability of negative impacts increases. Instructional change is difficult to make, and much support, teacher training, and time are required (see Connor, Piasta, et al., 2009; Connor et al., 2011; Kennedy, 2005).

In our study, interim assessments were not coupled with ongoing professional development for teachers. In addition, specific suggestions about differentiated instruction were not provided to teachers. Previous work, however, has indicated that interim evaluations coupled with professional development and clear recommendations for implementing instruction effectively can improve literacy instruction (see Al Otaiba et al., 2008; Connor, Jakobsons, Crowe, & Meadows, 2009). For instance, Connor, Jakobsons, et al. (2009) have reported that effective literacy instruction in Reading First classes, guided by research and supported by appropriate teacher training, improved reading comprehension for first graders in Florida.

It is also possible that teachers in the control schools were using assessments that may have already been in place quite effectively; that is, teachers in the control schools did not have to change their classroom practices and may have continued enacting their usual instruction according to assessment systems already in place in their schools (Coyne, 2015).

In contrast, treatment teachers were being forced to modify the type of interim assessments they may have been using, and thus their routine was interrupted. This disruption could have resulted in a disadvantage because within a school year, treatment teachers would be less practiced with using the new assessment programs to modify their instruction. Unfortunately, we did not have an appropriate fidelity of implementation plan in place to know exactly what type of classroom practices and teaching materialized in control classrooms (see Coyne, 2015).

Our inability to detect Acuity effects on student achievement is somewhat surprising because Acuity was designed to be aligned both to Indiana standards and ISTEP+, and thus one would expect to observe treatment effects in these grades. Because the Grade 3–8 estimates are not significant, one could conclude that Acuity had no important effects on student achievement. In contrast, there is evidence of a negative mCLASS effect in Grades K–2. These results cast some doubt on the assumption that the intervention was the same in lower and upper grades or that the two assessment products should be viewed as one.

The estimates of the ITT, TOT, and IV analyses were similar overall, indicating no effects for Acuity in Grades 3–8 and negative effects for mCLASS in Grades K–2. The fact that the estimates were overall similar across different analyses points to the robustness of the effects.

One potential limitation of our study is the lack of fidelity of treatment implementation. Despite significant efforts, the study's ability to document the details of treatment implementation was unfortunately limited. IDOE, the two vendors, and the study authors each independently confirmed that schools received the mCLASS or Acuity treatments. In addition, IDOE and the vendors confirmed that more than 95% of the eligible students participated in the interim assessments during each assessment window. Direct and indirect trainings of staff and teachers were conducted as contracted. The study team's analyses of the mCLASS and Acuity online data systems confirmed treatment school administrative and instructional staff made use of the interim assessment data and tools, and that control schools did not. Most teachers in the treatment schools agreed the interim assessment results led them to make changes in their instructional practice, and one third characterized these changes as "major." Nonetheless, we do not have additional information about how teachers used these assessments tools to improve achievement and to what degree implementation matched what was intended by design.

In addition, it was not possible to include a classroom (or teacher) level in our models because we did not have access to such data. Including a teacher level would have informed us about between-teacher variability (net of between-student or between-school variability) and would have allowed us to include teacher variables in our models. Such data would also have allowed us to model potential heterogeneity in the treatment effect. Although interim assessments are school interventions, teachers are the ultimate users of these assessments, and it is possible that different teachers may implement the treatment in a variety of ways. This is a second limitation of our study.

A third possible limitation is that our modeling could not capture potential school district effects because such data have not been collected. Although IDOE expected that schools would be the deciding actors (and that within them teaching staff would be influential) about whether to adopt the interim assessments, it quickly became apparent that district leadership was often the lead actor. Any differences between locations where instructional staff jointly volunteered for the rollout and locations where instructional staff were told they would be using the rollout products are unaccounted for in our study.

It is critical to remember that the implementation of Indiana's system of diagnostic assessments did not take place in a vacuum. In Indiana, as in many other states, multiple efforts—some state-led, others regional, yet others within some districts—were underway to increase the use of "data-driven decision making" in schools. In our study, many teachers in control schools used a variety of ongoing, less formal, progress monitoring procedures to help guide their practice. Given the ubiquitous pressures for progress monitoring and data-based decisions, the results we obtained in this study could be considered conservative estimates of the impact of Indiana's Diagnostic Assessment Tools intervention. Presumably the impact estimates would have been larger had the control schools used no data or progress monitoring tools. In an end-of-year survey administered to treatment teachers, one third said they had made what they considered to be "major changes" in their instructional practice during the study year, driven by what they had learned from their use of data provided by the intervention.

Causal inferences are warranted in this study because of the quality of the field experiment and the use of IV methods. Our tests for baseline equivalence of observed school variables suggested that random assignment was successful. However, the external validity of the results may be limited. Specifically, our sample was drawn from a subset of Indiana elementary schools that volunteered (in the spring of 2010) to implement the Diagnostic Assessment Tools intervention. We do not know what motivated these schools to volunteer. Their motivations may differ from those that volunteered for the prior year's study. Thus, there is reason to be cautious about generalizing our results beyond these groups of schools that aspired to use technology-supported interim assessments in the Indiana contexts of 2009–2010 and 2010–2011. Furthermore, while our study examined the effects of interim assessments, it examined only two such interventions (mCLASS and Acuity), and between those two we see different results. The effects of interim assessments may vary under a variety of conditions.

Finally, we note that this was a one-year study. It is reasonable to accept that instructional improvements driven by individual teachers working to choose to learn to use new or different practices (with minimal support and little expert guidance) will take considerable time to coalesce into a smoothly running repertoire (cf. Bryk, Gomez, & Grunow, 2011). Longitudinal trials of significant length and procedural sophistication are needed.

Funding

This research was supported by a grant from the Institute of Education Sciences, U.S. Department of Education (R305E090005).

ARTICLE HISTORY

Received 18 February 2015
Revised 10 October 2015
Accepted 25 October 2015

EDITORS

This article was reviewed and accepted under the editorship of Carol McDonald Connor and Spyros Konstantopoulos.

References

Al Otaiba, S., Connor, C. M., Kosanovich, M., Schatschneider, C., Dyrlund, A. K., & Lane, H. (2008). Reading First kindergarten classroom instruction and students' phonological awareness and decoding fluency growth. *Journal of School Psychology*, *48*, 281–314.

Angrist, J., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, *91*, 444–472.

Angrist, J. D., & Pischke, J-F. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton, NJ: Princeton University Press.

Blanc, S., Christman, J. B., Liu, R., Mitchell, C., Travers, E., Bulkley, K. E., & Lawrence, N. R. (2010). Learning to learn from data: Benchmarks and instructional communities. *Peabody Journal of Education*, *85*, 205–225.

Booher-Jennings, J. (2005). Below the bubble: "Educational triage" and the Texas accountability system. *American Educational Research Journal*, *42*(2), 231–268.

Boruch, R., Weisburd, D., & Berk, R. (2010). Place randomized trials. In A. Piquero, & D. Weisburd (Eds.), *Handbook of quantitative criminology* (pp 481–502). New York, NY: Springer.

Bracey, G. W. (2005). *No Child Left Behind: Where does the money go?* (EPSL-0506-114-EPRU). Tempe: Education Policy Studies Laboratory, Arizona State University.

Bryk, A., Gomez, L., & Grunow, A. (2011). Getting ideas into action: Building networked improvement communities in education. In M. Hallinan (Ed.), *Frontiers in sociology of education* (pp. 127–162). New York, NY: Springer.

Carlson, D., Borman, G. D., & Robinson, M. (2011). A multi-state district-level cluster randomized trial of the impact of data-driven reform on reading and mathematics achievement. *Educational Evaluation and Policy Analysis*, *33*, 378–398.

Clune, W. H., & White, P. A. (2008). *Policy effectiveness of interim assessments in Providence public schools* (WCER Working Paper No. 2008-10). Madison: University of Wisconsin–Madison, Wisconsin Center for Education Research.

Connor, C. M., Jakobsons, L. J., Crowe, E., & Meadows, J. (2009). Instruction, differentiation, and student engagement in Reading First classrooms. *Elementary School Journal*, *109*(3), 221–250.

Connor, C. M., Morrison, F. J., Fishman, B., Giuliani, S., Luck, M., Underwood, P. S., … Schatschneider, C. (2011). Testing the impact of child characteristics × instruction interactions on third graders' reading comprehension by differentiating literacy instruction. *Reading Research Quarterly*, *46*(3), 189–221.

Connor, C. M., Piasta, S. B., Fishman, B., Glasney, S., Schatschneider, C., Crowe, E., … Morrison, F. J. (2009). Individualizing student instruction precisely: Effects of child × instruction interactions on first graders' literacy development. *Child Development*, *80*(1), 77–100.

Cordray, D., Pion, G., Brandt, C., & Molefe, A. (2012). *The impact of the Measures of Academic Progress (MAP) program on student reading achievement* (NCEE 2013-4000). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Science, U.S. Department of Education.

Coyne, M. D. (2015). Effectiveness of a beginning reading intervention: Compared with what? Examining the counterfactual in experimental research. In C. M. Connor, & P. H. McCardle (Eds.), *Advances in reading intervention: Research to practice to research* (pp. 221–230). New York, NY: Brookes Publishing.

Datnow, A., Park, V., & Wohlstetter, P. (2007). *Achieving with data: How high-performing school systems use data to improve instruction for elementary students*. San Diego: University of Southern California, Rossier School of Education, Center on Educational Governance.

Davidson, K., & Frohbieter, G. (2011). *District adoption and implementation of interim and benchmark assessments* (CRESST Report 806). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.

Dedrick, R. F., Ferron, J. M., Hess, M. R., Hogarty, K. Y., Kromrey, J. D., Lang, T. R., … Lee, R. S. (2009). Multilevel modeling: A review of methodological issues and applications. *Review of Educational Research*, *79*(1), 69–102.

Goertz, M. E., Nabors-Olah, L., & Riggan, M. (2010). *From testing to teaching: The use of interim assessments in classroom instruction* (CPRE Research Report #RR-65). Philadelphia, PA: The Consortium for Policy Research in Education.
Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives*, *2*(3), 172–177.
Indiana State Board of Education. (2006). *A long-term assessment plan for Indiana: Driving student learning*. Indianapolis, IN: Author.
Kennedy, M. M. (2005). *Inside teaching: How classroom life undermines reform*. Cambridge, MA: Harvard University Press.
Konstantopoulos, S., Miller, S., & van der Ploeg, A. (2013). The impact of Indiana's system of interim assessments on mathematics and reading achievement. *Educational Evaluation and Policy Analysis*, *35*(4), 481–499.
Luce, T., & Thompson, L. (2005). *Do what works: How proven practices can improve America's public schools*. Dallas, TX: Ascent Education Press.
May, H., & Robinson, M. A. (2007). *A randomized evaluation of Ohio's Personalized Assessment Reporting System (PARS)*. Philadelphia: University of Pennsylvania Consortium for Policy Research in Education.
Michael & Susan Dell Foundation. (2009). *Performance management report*. Austin, TX: Author.
Nye, B., Hedges, L. V., & Konstantopoulos, S. (2000). Effects of small classes on academic achievement: The results of the Tennessee class size experiment. *American Educational Research Journal*, *37*, 123–151.
Perie, M., Marion, S., Gong, B., & Wurtzel, J. (2007). *The role of interim assessments in a comprehensive assessment system: A policy brief*. Retrieved from http://www.achieve.org/files/TheRoleofInterimAssessments
Plank, S. B., & Condliffe, B. F. (2013). Pressures of the season: An examination of classroom quality and high-stakes accountability. *American Educational Research Journal*, *50*(5), 1152–1182.
Sawchuk, S. (2009, May 13). Testing faces ups and downs amid recession. *Education Week*, *28*, pp. 1, 16–17.
Shepard, L., Davidson, K., & Bowman, R. (2011). *How middle school mathematics teachers use interim and benchmark assessment data* (CSE Technical Report). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).
Slavin, R. E., Cheung, A., Holmes, G. C., Madden, N. A., & Chamberlain, A. (2013). Effects of a data-driven district reform model on state assessment outcomes. *American Educational Research Journal*, *50*, 371–396.
Stock, J. H., Wright, J. H., & Yogo, M. (2002). A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business and Economic Statistics*, *20*(4), 518–529.
Tomlinson, C. A. (2000). *Differentiation of instruction in the elementary grades*. Champaign: ERIC Clearinghouse on Elementary and Early Childhood Education, University of Illinois.
Wang, Y., & Gushta, M. (2013, September). *Improving student outcomes with mCLASS Math, a technology-enhanced CBM and diagnostic interview assessment*. Presented at the Society for Research on Educational Effectiveness, Washington, DC.
Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*. Cambridge, MA: MIT Press.

Appendix

ITT or TOT Analysis

The linear model for student *i* in school *j* is

$$y_{ij} = \beta_{00} + \beta_{10} Treatment_j + \mathbf{X}_{ij}\mathrm{B}_{20} + \mathbf{Z}_j\mathrm{B}_{30} + \mathbf{G}_{ij}\mathrm{B}_{40} + \upsilon_j + \varepsilon_{ij} \qquad (\mathrm{A}-1)$$

where *y* is the outcome (mathematics or reading scores), β_{00} is the constant term, β_{10} is the estimate of the treatment effect, *Treatment* is a binary indicator of being in the treatment

group or not, **X** is a row vector of student-level predictors such as gender, race, or low SES (indicating poverty), B_{20} is a column vector of regression estimates of student predictors, **Z** is a row vector of school-level predictors such as school composition and urbanization, B_{30} is a column vector of regression estimates of school predictors, **G** represents grade fixed effects (dummies), B_{40} is a column vector of grade fixed effects estimates, υ is a school-level residual, and ε is a student-level residual. The variance component of residual term υ captures the nesting of students within schools. Notice that in the ITT analyses the treatment is represented by the dummy ITT treatment and in the TOT analyses the treatment is represented by the dummy TOT treatment.

IV Analysis

In the first step of the analysis the regression model is

$$RT_{ij} = \beta_0 + \beta_1 ITTTr_j + \mathbf{X}_{ij}\mathrm{B}_2 + \mathbf{Z}_j\mathrm{B}_3 + \mathbf{G}_j\mathrm{B}_4 + \varepsilon_{ij}. \qquad (\mathrm{A}-2)$$

This equation shows that *RT* is divided into two components: the prediction part of the regression and the error. In particular, the error component of *RT* is removed and the fitted values (*FV* in Equation A-3) from Equation (A-2) were used as the new main independent variable in Equation (A-3). The model at the second step is

$$y_{ij} = \beta_{00} + \beta_{10} FV_j + \mathbf{X}_{ij}\mathrm{B}_{20} + \mathbf{Z}_j\mathrm{B}_{30} + \mathbf{G}_{ij}\mathrm{B}_{40} + \upsilon_j + \varepsilon_{ij}, \qquad (\mathrm{A}-3)$$

where the estimate β_{10} is the IV treatment effect and all other terms have been discussed previously. This method essentially creates a new TOT treatment variable that should be free of selection and therefore the IV estimate facilitates causal inferences of the TOT treatment effect (see Angrist, Imbens, & Rubin, 1996).