Article

Educational and Psychological Measurement
2016, Vol. 76(4) 587–608

DOI: 10.1177/0013164415606498
epm.sagepub.com
SAGE

Effects of Interim Assessments Across the Achievement Distribution: Evidence From an Experiment

Spyros Konstantopoulos[1], Wei Li[1], Shazia R. Miller[2], and Arie van der Ploeg[2]

Abstract

We use data from a large-scale experiment conducted in Indiana in 2009-2010 to examine the impact of two interim assessment programs (*mCLASS* and *Acuity*) across the mathematics and reading achievement distributions. Specifically, we focus on whether the use of interim assessments has a particularly strong effect on improving outcomes for low achievers. Quantile regression is used to estimate treatment effects across the entire achievement distribution (i.e., provide estimates in the lower, middle, or upper tails). Results indicate that in Grades 3 to 8 (particularly third, fifth, and sixth) lower achievers seem to benefit more from interim assessments than higher achieving students.

Keywords

interim assessments, low achievers, quantile regression, experiment, IV estimation

Arguably most school interventions have a dual objective. First, school interventions aim to improve student performance and learning for all students. Second, they intend to close achievement differences between higher and lower achievers by providing additional help to lower achievers. Among the enormous range of school

[1]Michigan State University, East Lansing, MI, USA
[2]American Institutes for Research, Chicago, IL, USA

Corresponding Author:
Spyros Konstantopoulos, Michigan State University, 461 Erickson Hall, East Lansing, MI 48824, USA.
Email: spyros@msu.edu

interventions that focus on improving academic performance, one of the most intuitive is the use of standardized assessments administered several times during the school year, variously known as benchmark, diagnostic, or interim assessments (Perie, Marion, Gong, & Wurtzel, 2007). Interim assessments are presumed to provide teachers and administrators with useful information about changes in students' knowledge and understanding of the material throughout the school year. In addition, interim assessments are hypothesized to lead to constructive feedback and differentiated instruction (Tomlinson, 2000); in turn, differentiated instruction that more closely meets the student's ability and knowledge, promises improved student outcomes.

Specifically, interim assessments are expected to help teachers identify areas of instructional need for each student by providing immediate, detailed insight on students' strengths and weaknesses. Through an iterative series of assessments, teachers are expected to use assessment-based student data to modify instruction according to student's abilities, knowledge, and learning needs. The principle is that the more closely instruction matches students' needs and capabilities, the greater the probability of improved learning. With frequent access to objective data about student performance, teachers can monitor student progress closely to make informed decisions about which instruction is more effective for which student. Teachers are expected to diagnose weaknesses and strive to promote positive changes in student learning by adjusting their instructional practices to what the assessment data at hand suggest.

One might expect interim assessments to have a stronger positive effect on low-achieving students than other students. Specifically, frequent provision of reliable student data should help teachers identify more clearly who the lower performing students are at each point in time. Given the current accountability requirements to have students meet benchmarks teachers' knowledge about who the low achieving students are is likely to compel them to focus more attention on addressing these students' areas of weakness. Also, if interim assessments help teachers enact classroom practices that more closely match students' learning needs, this should increase the likelihood of improving student achievement, especially for low achievers, whose learning needs are greater. In addition, teachers might respond to information gleaned from interim assessments to identify weaknesses in understanding, and address those weaknesses through reteaching important concepts and skills in areas. Reteaching would have a larger effect on students who did not grasp the material well the first time, who are likely to be low-achieving students. Finally, differentiated instruction may have more impact for lower performing students.

If differentiated instruction varies by student ability level (e.g., additional focus on low achievers) one would expect more pronounced effects of interim assessments at certain parts of the achievement distribution. For example, if enacted differentiated instruction focuses on low achievers' needs the effects would be more pronounced in the lower tail of the achievement distribution than the middle or the upper tail, and the treatment effect would vary by achievement level. However, if the enacted differentiated instruction enabled by interim assessments does not interact with student

ability level one would expect similar benefits for all students regardless of achievement level. In this case one would expect a uniform treatment effect across the entire achievement distribution (i.e., comparable benefits for low, medium, and high achievers).

Interim assessments may have positive effects on student achievement on average as intended by product developers and policy makers. However, they may also affect various levels of achievement in different ways. First, if high achievers benefit more than do low achievers from interim assessments, one would expect a larger effect of the intervention at the upper tail of the achievement distribution. If this hypothesis were true, interim assessments would not reduce but instead enlarge the achievement gap between high and low achievers. Second, if low achievers benefit more from interim assessments, one would expect a larger effect of the intervention at the lower tail of the achievement distribution. If this hypothesis were true, interim assessments would reduce the achievement gap between high and low achievers. Third, if interim assessments have the same effects across the achievement distribution one would expect comparable benefits for students at different achievement levels. If this hypothesis were true, interim assessments would have no effect on the achievement gap between high and low achievers because lower and higher scoring students would benefit equally from the intervention.

Although the literature has provided evidence about the average effects of interim assessments on overall student achievement, little is known about the effects of interim assessments on low achievers. To our knowledge, recent evidence about the effects of interim assessments in the tails of the achievement distribution and specifically on low achievers has not been published. The only evidence documented in the literature has been on the effects of formative assessments on the achievement gap (see Black & Wiliam, 1998). Although these authors reported that formative assessments benefited lower achieving students, they used a very general definition of formative assessments that does not focus on interim assessments, now broadly in use. The purpose of our study is to fill in that gap in the literature and examine the effects of interim assessments throughout the distribution of student achievement with a focus on low achievers (i.e., the lower tail of the achievement distribution).

We used data from a large-scale experiment conducted in the state of Indiana in the 2009-2010 school year. Indiana's interim assessments intended 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). That is, in Indiana, interim assessments were explicitly expected to have different effects across the distribution of achievement, which matches our hypothesis.

Because our study's goal was to explore the effects of interim assessments at different levels of achievement we utilized an appropriate estimation procedure known as quantile regression (see Hao & Naiman, 2007). This method works well for examining treatment effects across the achievement distribution by producing estimates at different quantiles (e.g., the middle or the upper or lower tails) of the achievement

distribution (see Koenker, & Bassett, 1978). Quantile regression arguably provides a more complete picture of treatment effects and is a frequently used method in sociology and economics (Hao & Naiman, 2007).

Literature Review

There is a growing body of work on the effects of assessment-based interventions on the overall student population. 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 objective was to monitor student progress through easily accessible data and change teacher instruction to improve student performance. The impact of the first year of PARS on student achievement was not significant, but PARS effects were evident for students who retook the OGT assessments.

Another recent experiment examined the impact of the work of the Center for Data-Driven Reform in Education (CDDRE) on student achievement (Carlson, Borman, & Robinson, 2011). CDDRE emphasized instructional change based on benchmark assessment results (the CDDRE-designed *4Sight* assessments). The results of the first year of the experiment indicated significant positive effects on mathematics scores, but the effect was not statistically significant in reading. The impact of CDDRE was also measured over a 4-year period in follow-up work (see Slavin, Cheung, Holmes, Madden, & Chamberlain, 2013). The results showed significant positive effects on both mathematics and reading in the fourth year.

Along the same lines a different study examined the impact of the Measures of Academic Progress (MAP) assessments on reading achievement (Cordray, Pion, Brandt, & Molefe, 2012). MAP is an assessment system that incorporates computer-adaptive assessments and training for teachers in differentiated instruction. 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. Statistically significant differences were not observed at either grade in reading achievement on the Illinois State Achievement Test or on the MAP composite score.

Moreover, the average effect of interim assessment programs on mathematics and reading achievement in Indiana was measured in a recent study (Konstantopoulos, Miller, & van der Ploeg, 2013). Results indicated that treatment effects are positive but not consistently significant. Significant average treatment effects were detected in third and fourth grade reading and in fifth and sixth grade mathematics.

Finally, meta-analytic reviews have also provided evidence about the magnitude of effects of formative assessments but not of interim assessments as discussed in this study. Nyquist (2003) reported average effects of 0.40 standard deviation (*SD*) units for formative assessments. Kluger and DeNisi (1996) conducted a meta-analysis about the effects of interventions that provide feedback on student achievement, and also reported an effect size of 0.40 *SD* units. Other reviewers of the formative

assessment literature (e.g., Black & Wiliam, 1998) have reported similarly strong effects. This level of impact is not undisputed however (see Dunn & Mulvenon, 2009; Hattie & Timperley, 2007) and more recent meta-analyses of formative assessments of various kinds at various levels point to smaller but still meaningful effect sizes, in the 0.20 to 0.30 SD's range (Kingston & Nash, 2011).

The *Diagnostic Assessment Tools* System in Indiana

Two assessment programs were selected by the Indiana Department of Education (IDOE): *mCLASS* in early grades (i.e., K-2) and *Acuity* in later grades (i.e., 3-8). *mCLASS* assessment's diagnostic probes are conducted face-to-face, where students and teachers work together. For English language arts the student performs language tasks while the teacher records characteristics of the work using a personal digital assistant (PDA).[1] Teachers are guided through the assessments by the PDA and can immediately view results and compare them to prior performance. For mathematics the assessments are conducted using paper and pencil, with results entered onto a computer database by the teacher.

In Grades 3 to 8, CTB/McGraw-Hill's *Acuity* provided Indiana with online assessments in reading and mathematics. These assessments are 30- to 35-item multiple-choice online tests that can be completed within a class period, usually in group settings. *Acuity* assessments are closely aligned to Indiana standards and are designed to forecast student performance on the Indiana state test, the ISTEP+.[2]

Methodology

Sample Data

The experimental design was a cluster randomized design where clusters (i.e., schools) were assigned randomly to a treatment and control condition (see Boruch, Weisburd, & Berk, 2010). The experiment was conducted in Indiana during the 2009-2010 academic year and included K-8 schools that had volunteered to be part of the intervention in the spring of 2009. We identified 116 schools that met four conditions of our experiment: (a) schools were planning to use both assessment programs, (b) schools have not used either program, (c) schools were not using similar products such as MAP in the prior school year, and (d) schools would not participate in Indiana's NCLB differentiated accountability pilot in 2009-2010.[3] To assure a geographically balanced sample schools were nested in four U.S. Census locales—urban, suburban, small town, and rural. Although our target sample was nearly 50 schools overall, because we had anticipated some attrition we randomly selected 70 schools. Ten of these 70 schools had used one or both vendors' products the prior year and another school closed; thus, 11 schools overall were excluded from the sample. The remaining 59 schools were randomly assigned to treatment or control conditions.

To facilitate participation we decided to use an unbalanced design and have a larger number of schools in the treatment group. That is, 35 schools were randomly assigned to the treatment condition, and 24 schools were randomly assigned to the control condition. One of the control schools closed and another control school did not provide any student data. The final number of schools included in the random assignment process that provided data was therefore 57. This is the treatment assigned sample of schools. Due to a nearly 10% attrition of schools because of school closures, refusal to participate in the experiment, or use of similar assessments in the recent past, the participating sample of schools was reduced to 50. This is the treatment received sample of schools. Specifically, 32 of the 35 treatment schools that were offered treatment actually received treatment, and 18 of the 24 schools randomly assigned to the control group actually served as control schools. Overall, approximately 25,000 students participated in the study during the 2009-2010 school year. The schools in the treatment condition received *mCLASS* and *Acuity*, and the training associated with each product. The control schools did not receive access to these assessments and their associated trainings, and operated under business-as-usual conditions.[4]

Dependent and Independent Variables

In Grades 3 to 8, the outcomes were mathematics or reading scores of Indiana's state test, ISTEP+, while in Grades K-2 the outcomes were Terra Nova scores in mathematics and reading.[5] To simplify interpretation of estimates we standardized achievement scores within grades (i.e., mean = 0 and $SD = 1$) using the grade-specific SD of each outcome. For across grades analyses we pooled scores (see statistical models section).

The main independent variable was the treatment (coded 1 for treatment schools who received *mCLASS and Acuity* and 0 otherwise). The student covariates were 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), SES (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 percentage of female, minority, low SES (eligible for free or reduced price lunch), and LEP students.

Statistical Models

First we conducted analysis on the sample of the 57 schools that were randomly assigned to treatment and control conditions to estimate the effect of the Intention to

Treat (ITT). This estimate should be unbiased under the assumption that random assignment was successful. Second, we conducted sensitivity analyses using the treatment received sample of 50 schools to estimate the effect of the Treatment on the Treated (TOT). The TOT analysis tests the robustness of the estimates across the different samples. The total number of schools included in the ITT analyses was 57 and the total number of schools included in the TOT analyses was 50.

The analyses of the ITT sample of schools assume that the treatment schools that did not participate in the experiment actually received the treatment. The analyses of the TOT sample of schools may produce biased estimates of the treatment effect because of possible selection bias. Specifically, in our study four control schools and three treatment schools did not participate in the experiment (but provided student and school data). Suppose that school attrition is not random for treatment and control conditions. For example, low-achieving schools that are assigned randomly to the control group refuse to participate in the experiment because they want to receive the treatment (e.g., these schools expect that the treatment will improve their achievement considerably). Similarly, high-achieving schools that are assigned to the treatment group refuse to participate in the experiment because they anticipate that the treatment may disrupt teacher instruction and ultimately have a negative effect on achievement. In this scenario, detecting a treatment effect may not necessarily suggest that the treatment is effective. An alternative explanation is that because of selection in the treatment and control groups the treatment effect was overestimated. The analyses of the TOT sample of schools do not include these schools and thus, the treatment effect may be biased.

To address the ''caveats'' of the ITT and TOT analyses and especially the potential threat to the internal validity of the treatment estimates we used an Instrumental Variables (IV) approach coupled with quantile regression (see Abadie, Angrist, & Imbens, 2002). The key variable in this methodological procedure is the instrument, which should be orthogonal to any unobserved variables (i.e., the error term) in order to be valid. In experiments, it is straightforward to select an instrument because of random assignment. That is, random assignment in the assigned sample of schools (i.e., 57 schools) can be used to take into account possible selection bias. The sample size of schools in this two-step analysis is 57. The first step of this approach uses a logistic regression model where the outcome is binary (i.e., 1 for schools assigned to treatment that actually received the treatment and 0 otherwise). The predictors include random assignment (1 for schools assigned randomly to treatment and 0 otherwise) which served as the instrument, and student and school predictors (as defined in the variables section). The unit of analysis in this first step is the student. The second step of this approach is a quantile regression (see Equation 1 below). The predicted values from the logistic model in Step 1 are used in Step 2 to construct weights that account for the propensity of schools (and therefore students) that were assigned to and actually received the treatment. Intuitively treatment schools (and therefore students) that complied with their random assignment have a higher weight. This method is described in detail by Abadie et al. (2002). We used STATA to conduct

the IV quantile regression analysis; in particular we used the *ivqte*[6] command (see Frolich & Melly, 2010). The unit of analysis in the quantile regression is the student. The IV analysis was conducted by pooling data across grades (i.e., K-8, K-6, 3-8, and 3-6 data).

For the ITT and TOT samples we first conducted analyses using data across all grades (i.e., K through 8) to estimate treatment effects for both *mCLASS* and *Acuity*. Second we conducted analyses using K-2 data only or Grades 3 to 8 data only to estimate *mCLASS* or *Acuity* effects separately. The K-2 analyses included 44 participating schools instead of 50, because only 44 schools administered the Terra Nova test at the end of the 2009-2010 school year. We also conducted analyses using K-6 or Grades 3 to 6 data, because only few schools had enrolled seventh and eighth graders. Single grade analyses were also conducted to examine grade-specific treatment effects for Grades K-6 that had sufficient data. Prior ISTEP+ scores were available for Grades 4 to 6, and thus we also ran models that included prior student achievement as a covariate.

We regressed reading or mathematics scores on the treatment variable and student and school covariates. Grade fixed effects (i.e., dummies) were included in the model only in across grades analyses to control for potential grade differences. The analysis involved computing differences in achievement between students in treatment and control schools across the achievement distribution (e.g., lower, middle, or upper tails). We used quantile regression to estimate the interim assessment effect at various points on the achievement distribution (see Buchinsky, 1998; Koenker & Bassett, 1978). Specifically, we examined the treatment effect at the lower tail (e.g., 10th and 25th quantiles or percentiles), the middle (50th quantile or percentile), and the upper tail (e.g., 75th and top 90th quantiles or percentiles) of the achievement distribution.

At each quantile the across-grade model we used was the following:

$$y_i = \beta_0 + \beta_1 Treatment_j + \boldsymbol{X}_i\boldsymbol{B}_2 + \boldsymbol{Z}_j\boldsymbol{B}_3 + \boldsymbol{G}_i\boldsymbol{B}_4 + \varepsilon_i, \quad (1)$$

where y is the outcome variable (mathematics or reading scores), β_0 is the constant term, β_1 is the estimate of the treatment effect, *Treatment* is a binary indicator of being in a treatment or a control group, $\boldsymbol{X}$ is a row vector of student predictors, $\boldsymbol{B}_2$ is a column vector of regression estimates of student predictors, $\boldsymbol{Z}$ is a row vector of school predictors, $\boldsymbol{B}_3$ is a column vector of regression estimates of school predictors, $\boldsymbol{G}$ represents grade fixed effects (dummies), $\boldsymbol{B}_4$ is a column vector of grade fixed effects estimates, and ε is a student error. The subscript i indicates student. Grade effects were not included in the single grade analyses. Notice that in the ITT analyses the treatment dummy indicates schools that are randomly assigned to treatment or not, while in the TOT or IV analyses the treatment dummy indicates schools actually receiving the treatment or not. The standard errors of all quantile estimates were corrected for possible clustering effects (i.e., between-school variability) and nonconstant variation. The quantiles are like percentiles and can be interpreted as such. In the IV analysis the quantile regression employed at the second stage is similar to that described in Equation (1). The only difference is that in the IV analysis weights are

Table 1. Random Assignment Check Using Observed Variables.

Variable	M_d	SE_d	p value	ES
Grades K to 2: 57 Schools				
Proportion of female students	−0.008	0.010	.442	−0.233
Proportion of minority students	0.009	0.080	.914	0.029
Proportion of disadvantaged students	0.007	0.053	.897	0.033
Proportion of special education students	0.016	0.030	.603	0.117
Proportion of limited English proficiency students	0.025	0.014	.084	0.391
Grades 3 to 8: 57 Schools				
Proportion of female students	−0.006	0.010	.569	0.168
Proportion of minority students	0.016	0.078	.839	0.054
Proportion of disadvantaged students	0.009	0.052	.856	0.046
Proportion of special education students	0.011	0.031	.730	0.078
Proportion of limited English proficiency students	0.025	0.014	.079	0.399
Spring 2009 math scores	2.672	6.298	.673	0.115
Spring 2009 ELA scores	0.343	5.453	.950	0.017

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 *g*.

incorporated in the quantile regression to take into account the probability of treatment schools that complied with random assignment and received the treatment.

Results

First, we report results produced from analyses that checked whether random assignment was successful using observed school characteristics. This analysis included 57 schools. The objective was to identify observed variables where random assignment was not as successful as intended by design. We used *t* tests for independent samples to determine whether significant differences existed between treatment control groups for several school-level observed variables including proportion female, minority, disadvantaged, special education, limited English proficiency students as well as prior school achievement. Table 1 summarizes the results of this analysis and reports mean differences, their standard errors, *p* values of the *t* tests and effect sizes. This analysis took into account the What Works Clearinghouse standards about baseline equivalence of observed variables expressed as effect size estimates. We used the recommended effect size estimate, Hedges' g, which is expressed in SD units.

These results suggested that random assignment was by and large successful (i.e., no systematic significant differences were detected between treatment and control groups in observed school variables). The majority of the *p* values of the *t* tests were greater than 0.50. However, the *p* values for proportion of LEP students in a school were smaller than 0.10, but still greater than 0.05. The effect size estimates for proportion of LEP students in a school in Grades K-2 or Grades 3 to 8 were 0.391 and 0.399, respectively. According to the What Works Clearinghouse standards these

Table 2. Descriptive Statistics of Variables of Interest.

	N	Mean	*SD*
TerraNova reading score	7,640	535.36	59.24
TerraNova mathematics score	7,667	566.54	55.29
ISTEP English language arts score	13,287	481.58	64.42
ISTEP mathematics score	13,307	495.84	72.22
Age (months)	25,477	112.33	25.30
Female	12,339	0.48	0.50
Male	13,185	0.52	0.50
Race			
White	19,470	0.77	0.42
Black	3,776	0.15	0.36
Latino	1,123	0.04	0.21
Other	1,058	0.04	0.20
Limited English proficiency (LEP)	701	0.03	0.16
Non-LEP	24,899	0.97	0.16
Free or Reduced-Price Lunch (FRPL)	13,501	0.53	0.50
Non-FRPL	12,067	0.47	0.50
Special education (SE)	1,580	0.06	0.24
Non-SE	24,020	0.94	0.24

effect sizes do not meet baseline equivalence. That is, baseline equivalence may not have been successful for the observed variable proportion of LEP students in a school. All other effect size estimates were overwhelmingly smaller than 0.25. All the school variables reported in Table 1 were included in our regression models as statistical controls.

Table 2 reports sample sizes, means, and *SD*s of variables of interest. Forty-eight percent of the students were females, 77% of the students were White, and 53% of the students were eligible for free or reduced price lunch. The average student age was 9 years. More than 13,000 students had ISTEP+ scores and about 7,600 students had Terra Nova scores.

Across Grade Analyses

ITT Estimates. The treatment effect estimates are mean differences in SD units between treatment and control groups and positive estimates indicate a positive treatment effect. The results of the ITT analyses are reported in Table 3. The ITT estimates should be unbiased because there were produced from the sample of schools and students that were initially randomly assigned to treatment and control groups. Across quantiles and across grades all treatment effect estimates both in mathematics and reading scores were positive. The estimates of the Grade K-8 analysis were on average around one-tenth of a SD in mathematics, but they were smaller in reading. In mathematics, the treatment estimate at the 10th quantile was statistically significant at the .05 level and nearly one seventh of a *SD*. The estimates of the grade K-6

Table 3. Quantile Regression Estimates of Treatment Effects in Mathematics and Reading Achievement: ITT Analysis.

	Mathematics					Reading				
	10th	25th	50th	75th	90th	10th	25th	50th	75th	90th
Grades K to 8										
Treatment effect	0.149*	0.136	0.112	0.118	0.117	0.067	0.058	0.056	0.021	0.007
SE	0.073	0.070	0.062	0.073	0.087	0.060	0.048	0.053	0.054	0.047
Number of schools	57					57				
Number of students	20,792					20,795				
Grades 3 to 8										
Treatment effect	0.203*	0.194*	0.175*	0.157*	0.179*	0.114*	0.065	0.060	0.028	0.014
SE	0.074	0.075	0.074	0.076	0.061	0.057	0.046	0.042	0.049	0.050
Number of schools	57					57				
Number of students	13,274					13,254				
Grades K to 6										
Treatment effect	0.155	0.136*	0.116	0.113	0.099	0.076	0.067	0.064	0.036	0.024
SE	0.084	0.065	0.080	0.067	0.085	0.041	0.047	0.040	0.059	0.060
Number of schools	57					57				
Number of students	20,107					20,107				
Grades 3 to 6										
Treatment effect	0.214*	0.205*	0.176*	0.155*	0.172*	0.128*	0.082	0.072	0.047	0.030
SE	0.063	0.071	0.063	0.065	0.081	0.060	0.056	0.047	0.043	0.047
Number of schools	57					57				
Number of students	12,589					12,566				

Note. *SE* = standard error; ITT = intention to treat.
*$p \leq .05$.

analysis were similar. In mathematics, the 10th quantile estimate was significant at the .10 level, and the 25th quantile estimate was significant at the .05 level. These results suggest that in mathematics low achievers may have benefited more by interim assessments than other students and thus our hypothesis is partly supported.

The estimates produced by Grade 3 to 8 and 3to 6 analyses point to significant positive effects in mathematics across the achievement distribution. Treatment effects in these grades appear to be uniform across the mathematics achievement distribution. The estimates in the lower tail nonetheless were larger in magnitude and nearly one fifth of an *SD*. Arguably in education such effects are not trivial. In reading, the 10th quantile estimate was significant indicating a positive treatment effect for the lowest achieving students. All other estimates were not statistically significant at the .05 or .10 levels.

TOT Estimates. The results of the TOT analyses are reported in Table 4. The TOT analysis provided estimates based on the smaller number of schools and students that participated in the study. The results were overall very similar to those reported in Table 3, which points to the robustness of the estimates. All but two treatment effect estimates were positive for Grade K to 8, K to 6, 3 to 6, and 3 to 8 analyses. The magnitude of the effects was generally larger than those reported in Table 3 however. Again, the strongest evidence about effects was in mathematics and the larger estimates were obtained in the lower tails of the achievement distribution. Some of these estimates for mathematics were slightly larger than one fifth of an *SD*, a considerable effect.

The estimates in reading were smaller and mostly insignificant except the 10th quantile estimates in Grades 3 to 6 or 3 to 8. The reading estimates of the Grade K-2 analysis were very small and not different than zero. In fact, all grade K-2 estimates were insignificant both in mathematics and reading. By and large the estimates presented in Tables 3 and 4 point to positive treatment effects that are not consistently significant. The estimates were typically larger in mathematics than in reading. Also, the TOT estimates were larger than the ITT estimates. More important for our study, there is also some evidence that treatment effects were more pronounced for low achievers than for students at higher achievement levels. However, the lower tail estimates were not significantly different than the upper tail estimates.

IV Estimates. Table 5 summarizes the results of the IV analysis. The IV analyses provided estimates based on the sample of schools (and their students) randomly assigned to treatment or control groups (i.e., 57 schools). Overall, the estimates of the treatment effect were very similar to those reported in Tables 3 and 4, which points to the robustness of the estimates. All treatment effect estimates were positive and the magnitude of the effects was similar to those reported in Tables 3 and 4. There is consistent evidence about treatment effects in mathematics but not in reading. The larger estimates were observed in the lower tails of the achievement distribution. In Grades 3 to 8 some of the estimates in mathematics were larger than one fifth of an *SD*. In reading the estimates were smaller and typically insignificant except the 10th quantile estimate in Grades 3 to

Table 4. Quantile Regression Estimates of Treatment Effects in Mathematics and Reading Achievement: TOT Analysis.

	Mathematics					Reading				
	10th	25th	50th	75th	90th	10th	25th	50th	75th	90th
Grades K to 8										
Treatment effect	0.158*	0.144	0.118	0.121	0.111	0.071	0.057	0.053	0.023	0.017
SE	0.065	0.077	0.082	0.071	0.085	0.055	0.062	0.057	0.055	0.068
Number of schools	50					50				
Number of students	19,859					19,861				
Grades K to 2										
Treatment effect	0.023	0.027	0.019	0.016	−0.029	−0.018	−0.010	0.026	−0.022	0.001
SE	0.093	0.097	0.078	0.101	0.130	0.088	0.103	0.074	0.072	0.101
Number of schools	44					44				
Number of students	7,517					7,540				
Grades 3 to 8										
Treatment effect	0.227*	0.206*	0.189*	0.176*	0.185*	0.135*	0.075	0.062	0.028	0.024
SE	0.070	0.064	0.068	0.079	0.086	0.067	0.044	0.039	0.055	0.057
Number of schools	50					50				
Number of students	12,342					12,321				
Grades K to 6										
Treatment effect	0.171*	0.145*	0.123*	0.119	0.104	0.079	0.071	0.063	0.041	0.034
SE	0.067	0.065	0.062	0.080	0.075	0.054	0.051	0.049	0.059	0.054
Number of schools	50					50				
Number of students	19,174					19,173				
Grades 3 to 6										
Treatment effect	0.239*	0.220*	0.192*	0.175*	0.183*	0.150*	0.093	0.072	0.055	0.047
SE	0.078	0.062	0.065	0.066	0.076	0.059	0.051	0.046	0.048	0.051
Number of schools	50					50				
Number of students	11,657					11,633				

Note. *SE* = standard error; TOT = Treatment on the Treated.
*$p \leq .05$.

Table 5. Quantile Regression Estimates of Treatment Effects in Mathematics and Reading Achievement: IV Analysis.

	Mathematics					Reading				
	10th	25th	50th	75th	90th	10th	25th	50th	75th	90th
Grades K to 8										
Treatment effect	0.156*	0.136*	0.113	0.118	0.116	0.064	0.056	0.051	0.023	0.024
SE	0.078	0.069	0.067	0.064	0.078	0.058	0.054	0.051	0.054	0.060
Number of schools	57					57				
Number of students	20,792					20,795				
Grades 3 to 8										
Treatment effect	0.218*	0.203*	0.177*	0.160*	0.185*	0.107	0.062	0.058	0.031	0.030
SE	0.081	0.073	0.072	0.063	0.078	0.061	0.058	0.047	0.051	0.049
Number of schools	57					57				
Number of students	13,274					13,254				
Grades K to 6										
Treatment effect	0.165*	0.143*	0.112	0.114	0.104	0.071	0.066	0.059	0.038	0.038
SE	0.073	0.061	0.064	0.068	0.077	0.056	0.055	0.054	0.049	0.058
Number of schools	57					57				
Number of students	20,107					20,107				
Grades 3 to 6										
Treatment effect	0.235*	0.210*	0.177*	0.159*	0.178*	0.124*	0.078	0.071	0.052	0.055
SE	0.081	0.063	0.070	0.065	0.076	0.060	0.057	0.047	0.048	0.050
Number of schools	57					57				
Number of students	12,589					12,566				

Note. *SE* = standard error; IV = Instrumental Variables.
*$p \leq .05$.

6. There is some evidence that treatment effects were more pronounced for low achievers than for high achievers, however, the lower tail estimates were not significantly different than the upper tail estimates.

Single Grade Analyses

IV Estimates. To further explore the potential benefits of interim assessments we also conducted analyses within each grade (i.e., kindergarten through sixth grade).[7] The IV single grade analysis estimates were by and large similar to the ITT and TOT single grade analysis estimates. Therefore, in this subsection we report the IV estimates only for simplicity. Results from the IV analysis are summarized in Table 6. In kindergarten, first, and second grades the estimates of the treatment effect were small and not significantly different from zero. In reading many of the estimates especially in kindergarten were negative. In Grades 3 to 8, the overwhelming majority of the treatment effect estimates were positive, except for some of the sixth grade estimates in reading. In Grade 3, mathematics and reading, the 10th quantile estimate was positive and significant at .05. The 10th quantile reading estimate was nearly twice as large as the median estimate. The 25th quantile mathematics estimate was significant at the .10 level. In Grade 5, mathematics, all quantile estimates were positive and significant at .05. The Grade 6 mathematics estimates in the median and lower tail were large but statistically insignificant. All other estimates across grades and quantiles were not significant at the .05 or .10 levels. The larger estimates were observed in Grade 3 reading and in Grades 5 and 6 in mathematics. Some of these estimates in Grade 5 or 6 mathematics were as large as a fourth or a third of a SD, which arguably are potentially important effects.

In Grades 4 through 6, we also had data about prior achievement scores and thus we conducted analyses including prior scores as statistical controls in the regression models for sensitivity analyses purposes. The results of the IV analysis are summarized in Table 7. Again, these estimates are very similar to the ITT and TOT estimates. Overall, we observed the same pattern of results as in Table 6. The majority of the treatment effect estimates was positive, except for some of the fifth and sixth grade estimates in reading. In Grade 3 reading, the 10th, 50th, and 75th quantile estimates were positive and significant at .05. The estimates in the lower tail were larger than other estimates. The Grade 3 mathematics estimates were not significant. In Grade 5 mathematics, all quantile estimates were positive, significant at .05, and comparable in magnitude. All other estimates across grades and quantiles were not significant at the .05 level. The larger estimates were observed in Grade 3 reading and in Grade 5 mathematics.

Discussion

We examined the effects of two interim assessment programs on mathematics and reading achievement at various quantiles using data from a large-scale cluster randomized experiment in the state of Indiana. We focused on whether the use of interim assessments has stronger effects on lower achieving students. Because Indiana used

Table 6. Grade-Specific Quantile Regression Estimates of Treatment Effects in Mathematics and Reading Achievement: IV Analysis.

	Mathematics					Reading				
	10th	25th	50th	75th	90th	10th	25th	50th	75th	90th
Grade K										
Treatment effect	0.092	0.020	0.032	0.008	0.003	−0.037	−0.087	−0.041	−0.117	−0.140
SE	0.132	0.126	0.125	0.176	0.227	0.138	0.124	0.133	0.140	0.199
Number of schools	39					39				
Number of students	2,386					2,404				
Grade 1										
Treatment effect	0.045	0.031	−0.028	−0.014	0.037	−0.067	0.004	0.020	0.020	0.080
SE	0.150	0.098	0.088	0.093	0.114	0.149	0.121	0.089	0.093	0.117
Number of schools	38					38				
Number of students	2,473					2,471				
Grade 2										
Treatment effect	0.137	0.010	0.027	0.023	−0.064	0.073	0.059	0.103	0.009	−0.092
SE	0.101	0.094	0.083	0.078	0.129	0.095	0.072	0.059	0.077	0.095
Number of schools	44					44				
Number of students	2,658					2,665				
Grade 3										
Treatment effect	0.208*	0.176	0.157	0.101	0.131	0.215*	0.109	0.095	0.047	0.030
SE	0.093	0.093	0.092	0.090	0.104	0.092	0.063	0.054	0.067	0.085
Number of schools	57					57				
Number of students	3,741					3,737				

(continued)

Table 6. (continued)

	Mathematics					Reading				
	10th	25th	50th	75th	90th	10th	25th	50th	75th	90th
Grade 4										
Treatment effect	0.128	0.093	0.091	0.045	0.087	0.115	0.123	0.093	0.060	0.043
SE	0.121	0.083	0.101	0.108	0.138	0.093	0.067	0.063	0.059	0.057
Number of schools	57					57				
Number of students	3,739					3,730				
Grade 5										
Treatment effect	0.279*	0.262*	0.240*	0.290*	0.392*	0.107	0.059	0.087	0.102	0.154
SE	0.116	0.102	0.086	0.105	0.139	0.098	0.075	0.072	0.077	0.086
Number of schools	56					56				
Number of students	3,509					3,502				
Grade 6										
Treatment effect	0.276	0.313	0.287	0.198	0.117	0.005	−0.080	−0.109	−0.203	−0.135
SE	0.253	0.179	0.151	0.154	0.172	0.115	0.103	0.121	0.148	0.134
Number of schools	26					26				
Number of students	1,600					1,597				

Note. *SE* = standard error; IV = Instrumental Variables.
*$p \leq .05$.

Table 7. Grade-Specific Quantile Regression Estimates of Treatment Effects in Mathematics and Reading Achievement: IV Analysis That Controls for Prior Scores.

	Mathematics					Reading				
	10th	25th	50th	75th	90th	10th	25th	50th	75th	90th
Grade 4										
Treatment effect	0.107	0.056	0.060	0.037	0.082	0.139*	0.111	0.111*	0.092*	0.080
SE	0.100	0.087	0.092	0.109	0.113	0.065	0.066	0.037	0.043	0.054
Number of schools	57					57				
Number of students	3,604					3,572				
Grade 5										
Treatment effect	0.251*	0.184*	0.209*	0.203*	0.264*	−0.033	−0.028	−0.028	−0.029	0.032
SE	0.081	0.082	0.082	0.087	0.111	0.071	0.051	0.049	0.055	0.073
Number of schools	56					56				
Number of students	3,375					3,357				
Grade 6										
Treatment effect	0.062	0.091	0.119	0.114	0.187	−0.052	−0.095	−0.024	−0.076	0.009
SE	0.168	0.152	0.161	0.168	0.217	0.119	0.104	0.081	0.091	0.100
Number of schools	26					26				
Number of students	1,505					1,498				

Note. *SE* = standard error; IV = Instrumental Variables.
*$p \leq .05$.

two assessment systems, *mCLASS* for Grades K-2 and *Acuity* for Grades 3 to 8, we also examined the effects of the two systems separately. We collected data that should facilitate causal inferences about the treatment effect across the achievement distribution.

The findings in Grades 3 to 8 overall suggest that *Acuity* had a positive and significant impact in various quantiles of the mathematics achievement distribution. The magnitude of the effects was consistently greater than one sixth of a SD. In contrast, in reading only the 10th quantile estimate was positive and significant. The findings in Grades K-2 overall suggest that the effect of *mCLASS* on mathematics or reading scores across the achievement distribution was small and not statistically different than zero.

The treatment effect estimates were typically larger for low achievers and in certain grades significant. Therefore, the hypothesis that interim assessments may promote low achievers performance more than for other students was partly supported. However, the evidence was not consistent. In particular, it seems that *mCLASS*, the intervention in Grades K-2, did not affect mathematics or reading scores significantly at any level of the achievement distribution. In contrast, *Acuity*, the intervention in Grades 3 to 8, seems to have affected mathematics and reading achievement positively and in some instances considerably (e.g., Grades 3 through 6) with stronger effects for low-achieving students.

We conducted single grade analyses to explore whether the impacts for low-achieving students were stronger in particular grades, but we did not find any strong patterns emerging from the grade level by performance-level analysis. There was some evidence that the effects for low achievers were more pronounced in Grades 3 to 8, but overall the lower tail estimates were not significantly different than the upper tail estimates.

The ITT, TOT, or IV estimates were overall similar qualitatively, which indicates that the estimates are robust. There was some albeit weak indication that the treatment was more pronounced for low achievers than other students in some grades. These effects were observed both in mathematics and reading in Grades 3 to 6. However, because the evidence is not consistent, we are not able to conclude definitively that interim assessments can reduce the achievement gap and improve achievement for low achievers much more than for other students.

Although were able to detect significant treatment effects in Grades 3 to 8, we were unable to detect any significant or meaningful treatment effects in Grades K-2. One possibility is that *mCLASS* is not as effective as *Acuity*. Another possibility is that perhaps the effects of these assessment programs are not as visible in early grades. A third possibility is that the effects of *mCLASS* were not adequately captured by the outcome we used in Grades K-2. Specifically, in Grades K-2 the outcome measure was Terra Nova scores which we administered because Indiana's state wide test was not available in these grades. Terra Nova tests are general tests about student performance and contrary to ISTEP+, were not specifically aligned with Indiana's standards or curricula. In addition, *mCLASS* was not designed to be aligned with Terra Nova tests. Therefore, potential *mCLASS* effects may not have been manifested with Terra Nova scores because of misalignment. Finally, the ISTEP+ used

for Grades 3 to 8 is an accountability test with high stakes for both students and schools, whereas the Terra Nova given to the K-2 students is not. This difference may have resulted in the 3 to 8 teachers modifying their instruction to focus more attention to the interim assessment results than the K-2 teachers.

Our study has potential limitations. One possible limitation is the lack of fidelity of treatment implementation. 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. 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'' (van der Ploeg, 2012). Still, we are uncertain about how teachers used these assessment tools to improve student performance and about whether actual implementation matched what was intended by design.

Another potential limitation of our study is that we were unable to include classroom or teacher variables in our models due to lack of data. Including teacher variables would have allowed us to control for teacher variability in our models. In the same vein, we also could not capture potential school district effects due to lack of data. Although IDOE expected schools to be the deciding actors about whether to adopt interim assessments, district leadership was often the lead actor. Any differences between locations where instructional staff jointly volunteered for the roll out and locations where instructional staff were told they would be using the rollout products are not captured in our models.

Claims about causal effects are reasonable in this study because of the quality of the field experiment and the IV methodology. The generality of our results is less obvious. Our sample was drawn from a subset of Indiana elementary schools that volunteered to implement the school intervention. Thus, our results may generalize to schools that aspired to use technology-supported interim assessments that year but not necessarily to all schools in Indiana or to schools in other states or countries. Still our results should have higher external validity than those produced from smaller, convenience samples.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by a grant from the Institute of Education Sciences, U.S. Department of Education (R305E090005).

Notes

1. Due to federal requirements of the funding sources that Indiana used, schools were required to pay for the PDAs. In 2008-2009 Wireless Generation (Now Amplify) used Palm PDAs, but in subsequent years the software was ported to a variety of Android and Apple smartphones and tablet computers.
2. CTB/McGraw-Hill is also the testing vendor for the ISTEP+.
3. Indiana required these schools to use *mCLASS* and *Acuity*.
4. IDOE guaranteed full access to *mCLASS* and *Acuity* in 2010-2011 for control schools (i.e., delayed treatment) and study funds paid for vendors' to provide more intensive trainings of staff and support personnel.
5. Indiana's ISTEP+ does not extend below third grade. Thus, we administered the Terra Nova test in Grades K-2. This test is developed and maintained by CTB/McGraw Hill's educational assessment unit, which also develops and maintains the ISTEP+. Domain and conceptual overlap between the two assessment batteries is considerable.
6. Stata code is available on request from the authors.
7. Too few sampled schools enrolled students in seventh and eighth grades and thus these data were omitted form the analyses because they were not in accord with the experimental design.

References

Abadie, A., Angrist, J., & Imbens, G. (2002). Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earnings. *Econometrica*, *70*, 91-117.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Educational Assessment: Principles, Policy & Practice*, *5*(1), 7-74.

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.

Buchinsky, M. (1998). Recent advances in quantile regression models: A practical guideline for empirical research. *Journal of Human Resources*, *33*, 89-126.

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.

Cordray, D., Pion, G., Brandt, C., & Molefe, A. (2012). *The Impact of the measures of Academic Progress (Map) Program on student reading achievement*. Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, & U.S. Department of Education.

Dunn, K. E., & Mulvenon, S. W. (2009). A critical review of research on formative assessment: The limited scientific evidence of the impact of formative assessment in education. *Practical Assessment Research and Evaluation*, *14*(7), 1-11.

Frolich, M., & Melly, B. (2010). Estimation of quantile treatment effects with STATA. *STATA Journal*, *3*, 423-457.

Hao, L., & Naiman, D. Q. (2007). *Quantile regression*. Thousand Oakes, CA: Sage.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*, 88-112. doi:10.3102/003465430298487

Indiana State Board of Education. (2006). *A long-term assessment plan for Indiana: Driving student learning*. Indianapolis, IN: Author.

Kingston, N., & Nash, B. (2011, Winter). Formative assessment: A meta-analysis and a call for research. *Educational Measurement: Issues and Practice*, *30*, 28-37.

Kluger, A., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, *119*, 254-284.

Koenker, R., & Bassett, G. (1978). Regression quantiles. *Econometrica*, *46*, 33-50.

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*, 481-499.

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.

Nyquist, J. (2003). *Reconceptualizing feedback as formative assessment: A meta-analysis* (Unpublished master's thesis). Vanderbilt University, Nashville, TN.

Perie, M., Marion, S., Gong, B., & Wurtzel, J. (2007). *The role of interim assessments in a comprehensive assessment system: A policy brief*. Dover, NH: National Center for the Improvement of Educational Assessment.

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.

Tomlinson, C. A. (2000). *Differentiation of instruction in the elementary grades*. Champaign: ERIC Clearinghouse on Elementary and Early Childhood Education, University of Illinois.

van der Ploeg, A. J. (2012). *Interim assessment: Yes, it is complicated*. Paper presented at the National Conference on Student Assessment, Council of Chief State School Officers, Minneapolis, MN. Retrieved from https://ccsso.confex.com/ccsso/2012/webprogram/Presentation/Session2994/van%20der%20Ploeg%2C%20CCSSO%20NCSA%202012.pdf