Abstract Title Page

Title: Effects of Interim Assessments on the Achievement Gap: Evidence from an Experiment

Authors and Affiliations:
Spyros Konstantopoulos – Michigan State University
Wei Li – Michigan State University
Shazia Miller – American Institutes for Research
Arie van der Ploeg – American Institutes for Research

Abstract Body

Background:

Motivated by the passage of the No Child Left Behind (NCLB) Act, all states operate accountability systems that measure and report school and student performance annually. The NCLB accountability mandate further resulted in a plethora of assessment-based school interventions that targeted to improve student performance (Bracey, 2005; Sawchuk, 2009). Among the assessment-based solutions offered to improve student performance are periodic assessments variously known as interim assessments (Perie, M., Marion, S., Gong, B., & Wurtzel, J., 2007). Such assessments are administered typically three or four times during the school year providing information on students' understanding of the material. These systems provide resources designed to help teachers use assessment-based evidence to make better instructional decisions and differentiate instruction to meet students' needs.

Instructional change is considered the essential link between interim assessments and student performance. Specifically, interim assessments are hypothesized to lead to constructive feedback and differentiated instruction (Tomlinson, 2000), which are commonly believed to mediate the relationship between assessment and academic performance. 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. With frequent access to objective data about student performance, teachers can monitor student progress closely. In turn, this ongoing evidence about student performance will guide teachers' choices about which instruction is more effective for which student.

Because of the key component of differentiated/individualized instruction interim assessments should in principle be especially beneficial for low-achievers. Through interim assessments low-achievers should be easily identified and via appropriate differentiated instruction they could improve at least as much as higher achieving students. In addition, another key element of interim assessments is that teachers conduct meaningful follow-ups with students who struggle with certain tasks. These follow-ups would naturally incorporate re-teaching important concepts and skills in particular areas where low-achievers may have specific problems and need support.

Purpose:

The purpose of this study is to examine the effects of interim assessments on the achievement gap. We examine the impact of interim assessments throughout the distribution of student achievement with a focus on the lower tail of the achievement distribution. Specifically, we investigated the effects of two interim assessment programs (i.e., *mCLASS* and *Acuity*) on mathematics and reading achievement for high- median- and low-achievers. We use data from a large-scale experiment conducted in the state of Indiana in the 2009-2010 school year. Quantile regression is used to analyze student data. To our knowledge the literature thus far has not documented clearly the effects of interim assessments at different levels of achievement, and for low-achievers in particular. The only evidence we could find was that formative assessments may help low-achievers more than other students (see Black & William, 1998). In this study we fill in that gap in the literature.

Setting:

The study was a large-scale experiment conducted in Indiana during the 2009-2010 academic year and included K-8 public schools that had volunteered to participate in the intervention in the spring of 2009.

Participants:

From a stratified (by school urbanicity) pool of 116 schools we randomly selected 70 schools. Ten of the 70 schools had used one or both assessment programs the prior year and were excluded from the pool. Two other schools closed and another school did not provide any student data. Thus, our final sample included 57 schools, 35 in treatment and 22 in control condition. Overall, nearly 20,000 students participated in the study during the 2009-2010 school year.

Intervention:

In grades K-2, the *mCLASS* assessment's diagnostic probes are conducted face-to-face, where students and teachers work together. For reading and English language arts (ELA), the student performs language tasks while the teacher records characteristics of the work using a personal digital assistant (PDA). The *mCLASS* mathematics assessments are conducted using paper and pencil, with results entered onto a computer database by the teacher. CTB/McGraw-Hill's *Acuity* provides Indiana with online assessments in reading and mathematics for grades 3-8. The assessments are 30- to 35-item multiple-choice online tests that can be completed within a class period, usually in group settings. These assessments are closely aligned to Indiana standards and also aligned to some degree to the Indiana state test.

Research Design:

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 conditions. Schools were randomly assigned to a treatment (interim assessment) or a control condition. The schools in the treatment condition received *mCLASS* and *Acuity*, and the training associated with each program. The control schools operated under business-as-usual conditions.

Data Analysis:

We used student and school data in grades K-8. The total number of schools included in this analysis was 57. The outcomes were mathematics or reading scores of ISTEP+ (the Indiana state test) in grades 3-8, and the main independent variable was the treatment (mCLASS or Acuity coded as one for treatment schools and zero otherwise). In grades K-2 the main dependent variable was Terra Nova scores in mathematics and reading (administered by our study team). We conducted analyses using data across all grades (i.e., K through 8) for both mCLASS and Acuity, as well as analyses using grade 3-8 data (Acuity) only. Other analyses were conducted using grade K-6 data or grade 3-6 data since too few schools had enrolled seventh and eighth graders. We included student and school covariates in our regression models. The student covariates were gender, age, race, SES, special education status, and limited English proficiency status. The school level covariates were percent of female, minority, low SES, and limited English proficiency students.

An appropriate estimation procedure to investigate empirically whether interim assessments have uniform or differential effects across the achievement distribution is quantile regression. This method produces estimates in the middle and the tails of the achievement distribution and provides a more complete picture of the treatment effect (see Buchinsky 1998; Koenker & Bassett 1978). The standard errors of the regression estimates were corrected for potential clustering effects. We examined the treatment effect at the lower tail (e.g., 10th and 25th quantiles), the middle (50th quantile), and the upper tail (e.g., 75th and top 90th quantiles) of the achievement distribution. The quantiles are like percentiles and can be interpreted as such.

Results:

Table 1 reports sample sizes, means, and standard deviations of variables of interest. Forty eight percent of the students were females, 77 percent of the students were white, and 53 percent of the students were eligible for free or reduced price lunch. The average student age was nearly nine years (112 months). Nearly 13,000 students had ISTEP+ scores and about 7,500 students had Terra Nova scores.

The treatment effect estimates are mean differences in standard deviation units between treatment and control groups. Positive estimates indicate a positive treatment effect. The results of the main analysis are reported in Table 2. Across quantiles and across grades all treatment effect estimates both in mathematics and reading scores were positive. The estimates produced by 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 0.05 level and nearly one-seventh of a SD. The estimates of the grade K-6 analysis were similar. In mathematics, the 10th quantile estimate was significant at the 0.10 level, and the 25th quantile estimate was significant at the 0.05 level. These results suggest that in mathematics low-achievers may have benefited more by interim assessments than other students. These results are supportive of our hypothesis.

However, the estimates produced by grade 3-8 and 3-6 analyses point to significant positive effects in mathematics across quantiles. 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 a SD. Arguably in education such effects are not trivial. In reading, the 10th percentile estimates were significant providing support about a differential positive treatment effect for low-achievers. All other estimates were not statistically significant at the 0.05 or 0.10 levels.

Conclusions:

Overall, the findings suggest that the treatment effect was positive, but not consistently significant across all grades. Significant treatment estimates were observed in the grade 3-8 analysis in mathematics. The estimates were typically larger for low-achievers and in some cases significant. These results are consistent in terms of the sign of the effect (i.e., positive), but inconsistent in terms of statistical significance. We observed positive, statistically significant effects for grades 3-8 especially in mathematics. It seems that *Acuity* affected mathematics and reading achievement positively and in some instances considerably in grades 3-6. The notion that interim assessments may promote achievement more for low-achievers than other students was partially supported. However, the evidence is not systematic, and thus we were not able to conclude definitively that interim assessments could reduce the achievement gap and improve achievement for low-achievers more than for other students.

References

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

Boruch, R., Weisburd, D., & Berk, R. (2010). Place Randomized Trials. In A. Piquero & David Weisburd (Eds.), *Handbook of Quantitative Criminology* (pp 481-502). New York: , Springer.

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

Buchinsky, M. (1998). Recent Advances in Quantile Regression Models: A Practical Guideline for Empirical Research. *Journal of Human Resources* 33: 89–126.

Koenker, R., & Bassett, G. (1978). Regression Quantiles. *Econometrica* 46: 33–50.

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; Washington, DC: Achieve, Inc; and Washington DC: The Aspen Institute.

Sawchuk, S. (2009). Testing faces ups and downs amid recession. *Education Week 28* (31 May 13), pp. 1, 16-17.

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

Tables

Table 1. Desriptive Statistics

	N	Mean	SD
TerraNova Reading Score	7640	535.36	59.24
TerraNova Mathematics Score	7667	566.54	55.29
ISTEP English Language Arts Score	13287	481.58	64.42
ISTEP Mathematics Score	13307	495.84	72.22
Age (months)	25477	112.33	25.30
Female	25524	0.48	0.50
Race			
White	25427	0.77	0.42
Black	25427	0.15	0.36
Latino	25427	0.04	0.21
Other	25427	0.04	0.20
Limited English Proficiency	25600	0.03	0.16
Free or Reduced-Price Lunch	25568	0.53	0.50
Special Education	25600	0.06	0.24

Table 2. Quantile Regression Estimates of Treatment Effects in Mathematics and Reading Achievement

	Mathematics					Reading				
Quantile variable	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	20792					20795				
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	13274					13254				
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	20107					20107				
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	12589					12566				

*p ≤ .05; Note: SE = Standard Error