Educational Evaluation and Policy Analysis
December 2013, Vol. 35, No. 4, pp. 481–499
DOI: 10.3102/0162373713498930
© 2013 AERA. http://eepa.aera.net

The Impact of Indiana's System of Interim Assessments on Mathematics and Reading Achievement

Spyros Konstantopoulos
Michigan State University

Shazia Rafiullah Miller
Arie van der Ploeg
American Institutes for Research

Interim assessments are increasingly common in U.S. schools. We use high-quality data from a large-scale school-level cluster randomized experiment to examine the impact of two well-known commercial interim assessment programs on mathematics and reading achievement in Indiana. Results indicate that the treatment effects are positive but not consistently significant. The treatment effects are smaller in lower grades (i.e., kindergarten to second grade) and larger in upper grades (i.e., third to eighth grade). Significant treatment effects are detected in Grades 3 to 8, especially in third- and fourth-grade reading and in fifth- and sixth-grade mathematics.

Keywords: *interim assessments, student achievement, randomized control experiment*

Since the passage of the No Child Left Behind (NCLB) Act in 2001, all states operate accountability systems that measure and report school and student performance annually. The NCLB accountability mandate resulted in an abundance of assessment-based school interventions that targeted improved student performance (Bracey, 2005; Sawchuk, 2009). Among the assessment-based solutions offered to improve student performance are periodic assessments variously known as benchmark, diagnostic, or interim assessments (see Perie, Marion, Gong, & Wurtzel, 2007). Typically, these are administered three or four times during the school year and also provide resources designed to help teachers use assessment-based evidence to make better instructional decisions and differentiate instruction.

Many educational leaders in states and districts throughout the United States view interim assessments to be promising levers for increasing student achievement (Davidson & Frohbieter, 2011). The belief is that if objective data on student performance are available more frequently, then teachers possess evidence that they can use to better diagnose changes in student learning and adjust instructional practices—and do so before state accountability tests are administered. Thus, such assessment-based interventions are perceived as data-driven, likely to improve ongoing classroom instruction, and provide important feedback to students, teachers, and administrators. Support for this line of reasoning is widespread in the literature (see, for example, Datnow, Park, & Wohlstetter, 2007; Luce & Thompson, 2005; Michael & Susan Dell Foundation, 2009).

However, empirical evidence for these beliefs and practices remains problematic. Some quasiexperimental evidence on the impact of formative assessment, that is to say, evaluative information acquired in the course of classroom activity and used immediately, has been considered strong, with effect sizes often in excess of .40 standard deviation (*SD*) units (e.g., Black & Wiliam, 1998; Heritage, 2010). However, this level of impact is not undisputed (see Dunn & Mulvenon, 2009; Hattie & Timperley, 2007), and recent meta-analyses of formative assessments of various kinds at various levels also point to smaller effect sizes, in the .20 to .30 range (Kingston & Nash, 2011; Nyquist, 2003).

The few well-executed quasiexperimental studies of interim assessments described in the literature have produced inconclusive findings. For instance, Henderson, Petrosino, Guckenburg, and Hamilton (2007) examined pilot data for Massachusetts' quarterly benchmark assessments and found no statistically significant or substantively important differences between program and comparison schools. Another study used an interrupted time-series design to investigate the impact of Boston's Formative Assessment Student Thinking in Reading (FAST-R); on the Massachusetts' State Test and the separately administered Stanford Achievement Test, results were generally positive but not statistically significant (Quint, Sepanik, & Smith, 2008). Similarly, Faria and colleagues (2012) analyzed data from more than 100 schools in four Council of Great City Schools districts and reported small and marginally significant effects (less than .15 *SD*s) for teachers' use of interim assessments on student achievement.

Recently, two randomized controlled trials have reported estimates of the impact of interim assessment programs. First, Carlson, Borman, and Robinson (2011) reported very small (nearly .06 *SD*s) but significant impacts of the *4Sight* quarterly assessment program on mathematics achievement in Grades 3 through 8, but not on reading. Second, Cordray, Pion, Brandt, and Molefe (2012) conducted a cluster randomized experiment using the Northwest Evaluation Association (NWEA) *Measures of Academic Progress* (MAP) in fourth and fifth grades in schools in Illinois districts and found no impact on reading performance. Although these experiments varied in many ways, including units of analysis, assessments, and organizational level, their results were similar.

Nonetheless, there is a clear need to evaluate rigorously schools' use of interim assessments at scale and to examine the consistency of this practice across organizational levels and subject matter. We designed and conducted a cluster randomized experiment to test the hypothesis that schools in which teachers who have regular and repeated access to objective data that monitors student progress during an academic year and use it to guide their choices about day-to-day instruction will produce students who perform better on state assessments. We used data from nearly 60 schools in Indiana randomly assigned to treatment or control conditions. Because the data are from a well-designed randomized experiment, our estimates should have high internal validity and justify causal inferences about the intervention effects (Shadish, Cook, & Campbell, 2002). In addition, because our sample included schools in 31 districts widely distributed across a single state, our estimates should have higher external validity than those obtained from convenience or localized samples. Given the current level of interest in interim assessments as levers to accelerate school improvement, the results of this study should be informative to education practitioners, researchers, and policymakers.

We proceed as follows. First, we describe the intervention that is the focus of this experiment 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 third section, we first describe the sample and the data we used, and then we outline the analytic procedures we used. The fourth section presents the results and the fifth section discusses our findings and their implications, as well as the limitations of our analysis.

The Indiana Diagnostic Assessment Tools Intervention

In 2006, the Indiana Legislature charged the Indiana State Board of Education (ISBE) and the Indiana Department of Education (IDOE) to develop a long-term plan for a new assessment system that would measure student growth from year to year, as well as provide diagnostic information to teachers multiple times within a school year to improve ongoing instruction and ultimately facilitate student learning. Undergirding this policy decision was the premise that instruction should decrease the differences between students' current and intended knowledge. More frequent assessments were intended to regularly measure student knowledge, provide critical feedback, and document the size and nature of any differences between demonstrated and intended knowledge, and document changes in accumulated learning for students and schools over time. The plan was 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" (ISBE, 2006, pp. 11–12).

The plan that ISBE and IDOE developed in response to this charge stipulated that the new within-year assessments be voluntary. In schools that chose to use them, IDOE would cover costs. The plan tasked IDOE to ensure alignment of test content to Indiana standards and grade-level expectations. The new system was to be fully online. Through a standard public agency request for information, request for proposals, and negotiated bidding process, IDOE identified and purchased two commercial products: for Grades K–2, Wireless Generation's mCLASS, and for Grades 3 to 8, CTB/McGraw-Hill's Acuity. Both vendors agreed to work with Indiana staff and teachers to align their assessments, instructional resources, and training curricula to Indiana's content standards and instructional scope and sequences.[1] From IDOE's perspective, this pairing of vendors produced a single intervention that it referred to as Diagnostic Assessment Tools. This system of periodic diagnostic assessments is consistent, because students throughout Indiana take the same assessments; periodic, because students are tested at the same time points during the school year statewide; and diagnostic, because the assessments produce evidence from which teachers infer individual learning needs and seek appropriate instructional adjustments. A staggered, multiple-year rollout was planned with schools or districts volunteering in early spring for participation in the following academic year.

With this adoption, Indiana became the first state to implement technology-supported interim assessments statewide. IDOE began the rollout of the Diagnostic Assessment Tools in the fall of 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 2 to 3 days of training in late summer. These trainings were followed in fall after the first testing window by an additional vendor-conducted training focused on data use. Teachers trained in the summer received a supply of materials to train colleagues. They were expected to conduct two to three training sessions at their schools during the first 6 months of the program. Additional schools volunteered to use the Tools in each of the next several years with similar training experiences. IDOE staff expects that, essentially, all elementary schools and students statewide will be using the Tools by 2013 to 2014.

The mCLASS component of the Tools provides diagnostic measures in literacy and numeracy to teachers of K–2 students. Specifically, mCLASS:Reading3D, which comprises Dynamic Indicators of Basic Early Literacy Skills (DIBELS), alerts teachers to problems in students' learning of basic literacy and helps teachers identify and track students' error patterns, reading strategies, and comprehension via running reading records. DIBELS diagnostic probes are conducted face-to-face, student and teacher working together; there are multiple probes, each composed of a small number of tasks. The probes used in each testing window change as students progress. The student performs tasks while the teacher records and scores the work. This takes place using a personal digital assistant (PDA).[2] Teachers are guided through the assessment process by the PDA and, within the PDA interface, they can immediately view results and compare them with prior performance. In addition, at any point, teachers are able to monitor individual student progress in the classroom using short one-on-one, 1-minute probes and then see those results linked to previous results graphically on the PDA screen. In numeracy, mCLASS:Math helps teachers to identify students at risk of not acquiring proficiency in early mathematics skills and to understand students' mathematical thinking. The math assessments are paper and pencil, with the results entered into a computer by the teacher. Item selection varies across test windows, in line with expectations drawn from the Indiana state standards. Detailed individual and group reports, as well as ad hoc queries, are available to the classroom teacher and other authorized personnel via a Web interface for all mCLASS assessments. The system also includes a variety of additional progress monitoring tools for teachers as well as instructional resources.

CTB/McGraw-Hill's Acuity provides Indiana with online assessments in reading and mathematics for Grades 3 to 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 aligned to Indiana standards and the state test ISTEP$^+$ (Indiana Statewide Testing for Educational Progress–Plus). The Acuity assessments are of two types, diagnostic and predictive.

The diagnostic assessments focus on identifying specific student needs and then appropriately targeting and personalizing instruction. The tests are given at four windows, equally spaced throughout the school year, and are reported as number-right scores, with teachers given access to item wording and response choices. Item content follows the pacing of the state-recommended scope and sequence. The first assessment tests skills from the prior grade; subsequent tests match the within-grade scope and sequence of the state standards more carefully. The online reports to teachers provide the text of question and response options along with student results.

The Acuity predictive assessments are designed to forecast student performance on the Indiana state test, the ISTEP$^+$, and are reported in the standard score scale of that test.[3] There are three testing windows, with the third falling just before the state test is taken. The first testing window focuses on content typically taught at the beginning of the academic year; the next two windows incorporate more content typically taught later in the school year. The text of questions and response options used in the predictive assessments are not accessible to teachers in the online reports, because they are drawn from the same pool from which ISTEP$^+$ items are drawn.

For both assessment types, the online support system provides teachers with reports of performance against Indiana standards and objectives with individual and group summaries in various formats. Acceptable, advanced, and poor performances by individual students are emphasized in reports with green, yellow, and red highlights. In addition, Acuity permits 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 also available and may be assigned directly from Acuity's computerized displays as teachers desire. Teacher access to reports and queries is immediate.

Indiana made a deliberate, complex, and multifaceted decision to offer its schools Diagnostic Assessments Tools in the expectation that teachers would use these Tools to customize instruction to individual student skills and needs, and that they would use the results from the tests—along with personal knowledge of their students, the local curriculum, and state standards—to adjust their instruction to close the gaps between what individual students had shown they knew and what they needed to know. The state's expectation was that by adding meaningful detail to teachers' awareness of students' current performance relative to prior performance, instruction would closely match student needs and current and intended knowledge gaps would be reduced. These expectations are consistent with a long history of research and interpretation in educational science (e.g., Coe, 2002; Heritage, 2010; Kluger & DeNisi, 1996; Sadler, 1989; Stiggins, 2002; Tobias, 1987).

Related Literature

The uses of interim assessments in schools and their effects on student achievement have only recently begun to be documented in the literature. Some researchers have argued that it is unclear what expectations district and school staffs have for these assessments or that these assessments have been sufficiently realized (Bulkley, Christman, Goertz, & Lawrence, 2010). In particular, data provided by leaders, principals, and teachers who have implemented interim assessments suggest that by and large the value they see in such assessments is the hope of establishing a link between district policy and what and how teachers teach. In other studies, teacher focus group and survey data indicate that teachers believe that interim assessments can be useful in redesigning lessons, modifying instruction, and preparing students for standardized testing (Clune & White, 2008).

However, to achieve meaningful instructional change, principals and teachers need additional skills and knowledge. Interim assessments identify areas for improvement, but it is school staff who must use that information to identify and implement instructional strategies and practices that will bring about the targeted improvements (Blanc et al., 2010). Teacher interview data have 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 (Nabors-Olah, Lawrence, & Riggan, 2010). Evidence also suggests that only teachers who attempt to gauge students' conceptual understanding are likely to change instructional practices (Goertz, Nabors-Olah, & Riggan, 2010). In addition, teachers typically are not well trained to understand technical aspects of assessment tools, which limit their ability to use assessment data effectively. Principals and teachers may also have difficulty distinguishing among the types of assessments and their appropriate uses (Li, Marion, Perie, & Gong, 2010).

The impact of interim assessments has also been documented in recent work. In a large-scale cluster randomized experiment, May and Robinson (2007) evaluated Ohio's Personalized Assessment Reporting System (PARS) for the Ohio Graduation Tests (OGT). Although not strictly speaking an interim assessment program, PARS did feature repeated assessment opportunities and provided online reports on test outcomes and training for test data users. The researchers compared 10th-grade student achievement between 51 treatment and 49 control schools during the pilot year and found that the impact of the 1st year of PARS on student achievement was not significant. However, among students who did not pass the exam on the first try, students in the PARS group (treatment) were far more likely to attempt the test a second time. Furthermore, of the students retaking the exam, students in the PARS group received significantly higher scores than students in the control group.

The impact of the work of the Center for Data-Driven Reform in Education (CDDRE) on student achievement was examined in a recent study by Carlson and colleagues (2011). The CDDRE intervention is a data-driven decision-making process that emphasizes instructional change driven by benchmark assessment results (the CDDRE-designed *4Sight* assessments). The authors analyzed data from a multistate district-level cluster randomized experiment to investigate the potential benefits of CDDRE. The sample included more than 500 schools in 56 districts in seven states. The results of the 1st year of the experiment indicated statistically significant positive effects on mathematics scores, but the positive effects on reading scores were not statistically significant.

A follow-up study investigated the impact of CDDRE over a 4-year period (Slavin, Cheung, Holmes, Madden, & Chamberlain, 2011). A total of 391 elementary schools and 217 middle schools were included in the analysis. Multilevel models were used to analyze the impact of CDDRE on Grade 5 and 8 student achievement in mathematics and reading. The researchers created matched comparison groups to examine the effects of treatment on treated (TOT) using ANCOVA. The experiment and the matched design analyses yielded effects that were generally small in both grades. The results of the experimental analysis showed significant positive effects on mathematics and reading and grade levels in the 4th year, whereas the results of the matched design analysis indicated strong positive results for reading only.

Most recently, the impact of MAP assessments on reading achievement was examined (Cordray et al., 2012). MAP is a product of NWEA that is a widely used and commercially available system incorporating computer-adaptive assessments and training for teachers in differentiating instruction. Thirty-two elementary schools from five Illinois districts were randomly assigned to treatment and control conditions at Grades 4 and 5 (with one grade per school assigned to treatment and the other to control). More than 170 teachers and nearly 4,000 students were included in the analyses. The study found that

MAP was implemented with moderate fidelity. Nonetheless, 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.

Method

Data

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. The design was a two-level cluster randomized design (see Boruch, Weisburd, & Berk, 2010, for a discussion on these designs). Students were nested within schools, and schools were nested within treatment and control groups. Random assignment took place at the school level; that is, schools were randomly assigned to treatment and control conditions. Because the intervention was a whole-school intervention and the program implementation took place at the school level, we used a clustered randomized design to achieve a close match between research design and practice.

Indiana first announced the availability of the Diagnostic Assessment Tools in early spring 2008. Schools were asked to volunteer to participate. A similar process was followed in early spring 2009, 2010, 2011, and 2012. In each case, the list of volunteers closed when it appeared that funding for the year had been exhausted. This study's sample was drawn from the queue of 264 schools volunteering in early spring 2009 for mCLASS and 421 schools for Acuity for the 2009–2010 school year, most volunteering for both products.

Working with IDOE, we identified within this queue a set of 116 schools that met the conditions of our experiment. Our conditions were that (a) the school would use mCLASS and Acuity, (b) the school was not a previous user of either product, (c) the school was not a user in the prior school year of similar products such as NWEA's MAP, and (d) the school would not be a participant in Indiana's NCLB differentiated accountability pilot in 2009–2010.[4] The first requirement was intended to match the spirit of Indiana's conception of the Diagnostic Assessment Tools as a single unitary intervention. The remaining requirements were intended to reduce contamination and to assure the voluntary nature of school's participation. To assure a geographically balanced sample statewide, schools were stratified in four U.S. Census locales—urban, suburban, small town, and rural. From this nested pool, we randomly selected 70 schools. Communication with vendors determined that 10 of the 70 schools had used one or both vendors' products the prior year and thus these schools were removed from the pool. District reorganization at the end of the 2008–2009 school year closed one school. The final pool numbered 59 eligible schools and each was randomly assigned to a treatment or control condition.

Our initial objective was a balanced design with 25 schools randomly assigned to treatment and another 25 assigned to control schools. However, to facilitate participation, we decided to use an unbalanced design and targeted a larger number of schools for the treatment group (i.e., 30 treatment and 20 control schools). The final sample included 59 schools—35 randomly assigned to the treatment condition, whereas the remaining 24 schools were randomly assigned to the control condition. Of the 35 treatment schools, 31 participated in the experiment, and of the 24 control schools, 18 participated in the experiment for the whole year. The total number of participating schools was 49. Overall, some 20,000 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. The control schools did not receive access to these assessments and their associated trainings, continuing to operate under business-as-usual conditions.[5] However, business-as-usual today in many Indiana schools, as elsewhere in the nation, includes an active role for data-driven decisions informed by progress monitoring data. As argued by Goren (2012), it is difficult to avoid data in today's American schools. At the time of our study, IDOE was implementing a statewide response to intervention initiative, offering data interpretation workshops throughout the state. Our surveys of reading and mathematics teachers in study schools found that 88% of the teachers in control schools said that they used assessment data to monitor student progress. Three quarters said that they placed at least a moderate emphasis on customizing instruction to students based on monitoring evidence. Nor was it necessarily the case that in treatment schools mCLASS or Acuity was the only source of assessment data: About one fifth of teachers in control and treatment schools confirmed that they had access to Renaissance Learning's STAR assessments or other data tools.

Measures

The outcomes were mathematics and reading ISTEP$^+$ (Indiana's state test) scores in Grades 3 to 8. In Grades K–2, the main dependent variable was Terra Nova scores in mathematics and reading.[6] The main independent variable indicated whether a school received the treatment (i.e., mCLASS and Acuity) or did not receive the treatment. The treatment variable was coded as a dummy variable taking the value of one for schools that received mCLASS and Acuity and zero otherwise. 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), socioeconomic status (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 (English language learner [ELL]) 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, lower SES (eligible for free or reduced price lunch), and limited English proficiency students. The school variables were aggregate measures of student covariates and were constructed by computing school means of student covariates.

Statistical Analysis

We used student and school data for Grades K–8. We conducted analyses on the initial number of schools that were randomly assigned to conditions (intention to treat or ITT) and on the participating schools (TOT). To capture the dependency 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 effects indicated differences in average mathematics or reading achievement across schools. We conducted several analyses using data across all grades (i.e., K–8), using K–2 grade data, and 3 to 8 grade data separately. We also conducted within-grade analyses for each grade separately.

In each case, we regressed mathematics or reading scores on the treatment variable (which was at the school level) that was coded as a binary indicator, and other student and school covariates. The regression model for student *i* in school *j* is

$$\begin{aligned} y_{ij} = \beta_{00} + \beta_{10}\text{Treatment}_j + \mathbf{X}_{ij}\mathbf{B}_{20} + \mathbf{Z}_j\mathbf{B}_{30} \\ + \mathbf{G}_{ij}\mathbf{B}_{40} + \upsilon_j + \varepsilon_{ij}, \end{aligned} \tag{1}$$

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 that represents treatment assignment, **X** is a row vector of student-level predictors, $\mathbf{B}_{20}$ is a column vector of regression estimates of student predictors, **Z** is a row vector of school-level predictors, $\mathbf{B}_{30}$ is a column vector of regression estimates of school predictors, **G** represents grade fixed effects (dummies), $\mathbf{B}_{40}$ is a column vector of grade fixed effects estimates, υ is a school-level residual, and ε is a student-level residual. The variance of υ captures the nesting of students within schools.

The analysis described in Equation 1 was replicated for Grades K–2, Grades 3 to 8, and Grades 3 to 6 separately to determine mCLASS and Acuity effects, respectively. The analyses for Grades K–2 estimated only TOT effects because the outcomes (i.e., Terra Nova scores) in these grades were available only for participating schools. We also conducted within-grade analyses for Grades K–6. In Grades 7 and 8, the data were scarce, and thus these grades are not included in the within-grade analyses. In the within-grade analyses, the grade dummies were omitted from the model. For Grades 4, 5, and 6, we also ran models that included prior student achievement as a covariate. This analysis was not possible to conduct in other grades given that prior scores were not available or the data were very scarce.

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; see Glick & White, 2003; Konstantopoulos, 2006). The standardization creates comparable indexes of achievement across grades under the assumption that the tests are vertically equated (Holland & Rubin, 1982; Indiana Department of Education & CTB/McGraw-Hill, 2010; Kolen & Brennan, 1995). We standardized mathematics or reading scores within each grade using the grade-specific sample *SD* of the outcome and then pooled all scores across grades together for the analyses. The sample *SD*s within each grade were very similar, and thus the within-grade standardization seemed reasonable.

We also conducted sensitivity analyses. In particular, because the schools in our sample were located in urban, suburban, small town, and rural areas, it is possible that the treatment effect varied by school location—urbanicity. To address this issue, we conducted analyses within school urbanicity categories—that is, we repeated the analyses described in Equation 1 four times within each school urbanicity category (i.e., urban, suburban, small town, and rural) separately.

Results

Tests for Random Assignment

The strength of a randomized experiment is that successful random assignment ensures that the schools (and as a result, the students in these schools) in the treatment and control conditions have on average equivalent observed and unobserved characteristics. The idea is to create equivalent groups of schools and students, on average, across conditions. When random assignment holds, differences in characteristics across treatment types are only due to chance and should not be systematic (Shadish et al., 2002). Random assignment should effectively eliminate preexisting differences between schools and students in the treatment and control conditions. We checked whether random assignment was successful, empirically, using observed school characteristics. Note, however, that such tests do not discredit random assignment altogether but can merely identify observed variables where random assignment may not have been as successful as intended.

The preliminary analyses involved tests that checked whether random assignment of schools to conditions was successful for several observed variables at the school level. The random assignment indicated ITT. We used *t* tests for independent samples to determine whether significant differences existed between the two conditions for several school-level observed variables, including proportion female, minority, disadvantaged, special education, and limited English proficiency students, as well as prior school achievement. In addition, we followed the What Works Clearinghouse (WWC) standards and reported the results about baseline equivalence in effect size estimates format. We used Hedges's *g* to compute the effect size estimates. The results of these analyses are reported in Tables 1 and 2. In particular, these two tables report mean differences between treatment and control groups, the standard errors of these mean differences, the *p* value of the *t* test, and the effect size estimates (in the last column).

The results for the sample of schools that we used for random assignment are reported in Table 1. The results for the schools that ended up participating in the study are reported in Table 2. The *p* values of the *t* tests in Tables 1 and 2 indicated that random assignment was overall successful. The results in Tables 1 and 2 were essentially similar, that is, there were no significant differences between treatment and control conditions with respect to the variables of interest. The mean differences were typically smaller than their standard errors and never larger than two times their standard errors. As a result, all *p* values were larger than .05, which is the standard level of statistical significance. Thus, these results overall seem to be in accord with the principle of random assignment. However, the *p* values for students with limited English language proficiency were close to

TABLE 1

Check of Random Assignment on Observed Variables of Interest: All Schools

Variable	M_d	SE_d	*p* Value	Effect size
Grades K–8: 58 schools				
Proportion of female students	–.006	.010	.563	–.173
Proportion of minority students	.016	.079	.845	.053
Proportion of disadvantaged students	.042	.054	.434	.202
Proportion of special education students	.011	.014	.440	.203
Proportion of limited English proficiency students	.026	.014	.078	.406
Grades 3 to 8: 57 schools				
Proportion of female students	–.006	.010	.580	–.170
Proportion of minority students	.009	.081	.916	.029
Proportion of disadvantaged students	.048	.055	.384	.228
Proportion of special education students	.011	.014	.440	.203
Proportion of limited English proficiency students	.025	.014	.087	.393
Spring 2009 math scores	3.626	6.754	.594	.148
Spring 2009 ELA scores	1.130	5.362	.834	.059

Note. M_d = difference between treatment and control group school means; SE_d = standard error of the mean difference; *p* Value = *p* value of the *t* test; the effect size reported is Hedges's *g*; ELA = English language arts.

TABLE 2

Check of Random Assignment on Observed Variables of Interest: Participating Schools

Variable	M_d	SE_d	*p* Value	Effect size
Grades K–8: 49 schools				
Proportion of female students	–.001	.010	.921	–.033
Proportion of minority students	–.014	.094	.886	–.044
Proportion of disadvantaged students	.029	.062	.642	.131
Proportion of special education students	.008	.013	.551	.181
Proportion of limited English proficiency students	.022	.016	.175	.327
Grades K–2: 44 schools				
Proportion of female students	–.002	.011	.842	–.066
Proportion of minority students	.018	.099	.860	.055
Proportion of disadvantaged students	.040	.066	.550	.177
Proportion of special education students	.002	.014	.911	.035
Proportion of limited English proficiency students	.027	.019	.151	.386
Grades 3 to 8: 49 schools				
Proportion of female students	–.001	.010	.921	–.033
Proportion of minority students	–.014	.094	.886	–.044
Proportion of disadvantaged students	.029	.062	.642	.131
Proportion of special education students	.008	.013	.551	.181
Proportion of limited English proficiency students	.022	.016	.175	.327
Spring 2009 math scores	3.821	7.862	.630	.148
Spring 2009 ELA scores	1.869	6.365	.771	.091

Note. M_d = difference between treatment and control group school means; SE_d = standard error of the mean difference; *p* Value = *p* value of the *t* test; the effect size reported is Hedges's *g*; ELA = English language arts.

being statistically significant in Table 1. Thus, it is unclear that for this variable baseline equivalence was achieved. In addition, following the WWC standards when the effect sizes are between .05 and .25, statistical adjustment is required in the regression models. This rule applies to the majority of the variables in our study, which we have included as covariates in

TABLE 3

Estimates of the Treatment Effect in Mathematics and Reading Achievement: ITT Analysis

	Model I				Model II			
	Mathematics		Reading		Mathematics		Reading	
Variable	Estimate	*SE*	Estimate	*SE*	Estimate	*SE*	Estimate	*SE*
Grades K–8								
Treatment effect	.074	.083	.014	.074	.127	.069	.078	.050
Number of schools	57		57		57		57	
Number of students	20,428		20,436		19,199		19,205	
Grades 3 to 8								
Treatment effect	.141	.088	.039	.075	.192*	.073	.112*	.051
Number of schools	57		57		57		57	
Number of students	12,784		12,765		12,251		12,233	
Grades 3 to 6								
Treatment effect	.134	.088	.042	.075	.187*	.073	.118*	.054
Number of schools	57		57		57		57	
Number of students	12,137		12,115		11,632		11,610	

Note. Model I includes the treatment only; Model II adds student and school variables, and grade dummy variables. ITT = intention to treat; *SE* = standard error.
*$p \leq .05$.

our statistical models. For one variable (proportion of limited English proficiency students), however, the effect size is greater than .25, and thus, according to WWC, baseline equivalence for that variable may not have been met. We have also included that variable in our regression models.

The results reported above were strengthened by results obtained from our use of the Surveys of Enacted Curriculum (SEC; Blank, Porter, & Smithson, 2001) for reading and mathematics teachers in Grades 2 and 5 in treatment and control schools. These instruments have been shown to capture variation in instructional practice with respect to topics taught and cognitive demand (Porter, 2002). There was no initial expectation that control and treatment teachers would teach differently during the study year because there were no observable differences in the academic content and cognitive demand of their instructional practice the prior year—that is, the survey revealed treatment and control schools were similar in terms of instructional intent at the beginning of the year.

Analyses Across Grades

In the primary analyses, we used two-level models to estimate the treatment effects. The regression estimates were mean differences in *SD* units between treatment and control groups. Positive estimates indicated a positive treatment effect. The results of the ITT analyses are reported in Table 3. All treatment effect estimates for mathematics and reading scores were positive, which suggest an overall positive treatment effect. The estimates obtained from Grade K–8 analysis were on average around one tenth of an *SD* in mathematics but smaller in reading. All estimates, however, were not statistically significant at the .05 level. The estimates from Grade 3 to 8 analysis were also positive, typically larger in magnitude, and reached statistical significance in mathematics and reading at the .05 level when student and school covariates were included in the model. The significant treatment effect was slightly smaller than one fifth of an *SD*) in mathematics and nearly one tenth of an *SD* in reading. The reading achievement estimates were insignificant in Grades K–8 and 3–8 analyses. The estimates from Grade 3 to 6 analysis were very similar to those from Grade 3 to 8 analysis and suggested a positive and significant treatment effect.

The results of the TOT analyses are reported in Table 4. The TOT analyses provided estimates for the smaller number of schools that participated in the study and, therefore, selection bias

TABLE 4

Estimates of the Treatment Effect in Mathematics and Reading Achievement: TOT Analysis

	Model I				Model II			
	Mathematics		Reading		Mathematics		Reading	
Variable	Estimate	*SE*	Estimate	*SE*	Estimate	*SE*	Estimate	*SE*
Grades K–8								
Treatment effect	.149	.088	.051	.083	.187*	.070	.098	.055
Number of schools	49		49		49		49	
Number of students	19,101		19,109		17,931		17,937	
Grades K–2								
Treatment effect	.029	.119	.046	.112	.084	.100	.101	.091
Number of schools	44		44		44		44	
Number of students	7,644		7,671		6,948		6,972	
Grades 3 to 8								
Treatment effect	.231*	.092	.083	.084	.265*	.074	.139*	.055
Number of schools	49		49		49		49	
Number of students	11,457		11,438		10,983		10,965	
Grades 3 to 6								
Treatment effect	.222*	.092	.087	.084	.258*	.074	.146*	.054
Number of schools	49		49		49		49	
Number of students	10,810		10,788		10,364		10,342	

Note. Model I includes the treatment only; Model II adds student and school variables, and grade dummy variables. TOT = treatment on treated; *SE* = standard error.
*$p \leq .05$.

is not impossible. Nonetheless, results were for the most part similar to those reported in Table 3. All treatment effect estimates were positive as in Table 3, but the magnitude of the effects was larger for Grade K–8, 3 to 8, and 3 to 6 analyses than those reported in Table 3. In the K–8 analyses, when student and school covariates were included in the model, the treatment effect estimate in mathematics was significant and slightly smaller than one fifth of an *SD*. The reading scores estimate was smaller and insignificant. The estimates from Grade K–2 analyses were also positive but insignificant for mathematics and reading. The estimates for reading scores were larger than those for mathematics. The treatment effects from Grade 3 to 8 and 3 to 6 analyses were positive and significant, and nearly one fourth of an *SD* in mathematics. When student and school covariates were included in the model, the treatment effect estimate in reading was also significant and nearly one eighth of an *SD*. In sum, the estimates presented in Tables 3 and 4 point to positive treatment effects that are not consistently significant. The estimates are typically larger for mathematics than for reading scores. In addition, the TOT estimates are larger than the ITT estimates. By and large, it appears that the treatment is more pronounced in Grades 3 to 8, which would suggest an Acuity effect.

The results produced by the school urbanicity categories analyses were overall similar to those reported above. These results are reported in Table 5. Specifically, Table 5 reports ITT and TOT estimates of the treatment effect separately for rural, urban, suburban, and small town schools. Student and school variables, as well as grade dummies, were included in the model. Overall, the treatment effects did not seem to vary much by school urbanicity, and the results were similar to those reported for all schools. The ITT results for rural schools were qualitatively similar to the results reported in Tables 3 and 4. This was expected because more than 50% of the schools were in rural areas. The ITT results for small town schools, however, were different, and all estimates were statistically insignificant. In urban and suburban schools, the K–8 results were significant for mathematics and reading achievement. Thus, there is only

TABLE 5

Treatment Effect Estimates by School Urbanicity: ITT and TOT Analysis

	ITT				TOT			
	Mathematics		Reading		Mathematics		Reading	
Variable	Estimate	*SE*	Estimate	*SE*	Estimate	*SE*	Estimate	*SE*
Rural								
Grades K–8								
Treatment effect	.151	.075	.067	.037	.212*	.065	.083*	.041
Number of schools	31		31		27		27	
Number of students	8,990		8,999		8,488		8,496	
Grades K–2								
Treatment effect					.076	.107	.041	.077
Number of schools					23		23	
Number of students					3,177		3,191	
Grades 3 to 8								
Treatment effect	.255*	.086	.101*	.046	.305*	.079	.135*	.053
Number of schools	31		31		27		27	
Number of students	5,813		5,808		5,311		5,305	
Grades 3 to 6								
Treatment effect	.218*	.085	.106*	.045	.297*	.079	.143*	.050
Number of schools	31		31		27		27	
Number of students	5,612		5,608		5,110		5,105	
Urban								
Grades K–8								
Treatment effect	.255*	.086	.347*	.086	.263*	.087	.353*	.086
Number of schools	9		9		9		9	
Number of students	3,970		3,974		3,970		3,974	
Grades K–2								
Treatment effect					.305*	.128	.460*	.128
Number of schools					9		9	
Number of students					1,725		1,734	
Grades 3 to 8								
Treatment effect	.245*	.115	.267*	.116	.250*	.117	.272*	.117
Number of schools	9		9		9		9	
Number of students	2,245		2,240		2,245		2,240	
Grades 3 to 6								
Treatment effect	.221	.118	.268*	.118	.225	.119	.270*	.119
Number of schools	9		9		9		9	
Number of students	2,053		2,049		2,053		2,049	
Suburban								
Grades K–8								
Treatment effect	.216*	.041	.104*	.040	.212*	.043	.078	.042
Number of schools	8		8		7		7	
Number of students	3,637		3,639		3,238		3,241	
Grades K–2								
Treatment effect					.294*	.080	.380*	.079
Number of schools					6		6	
Number of students					1,102		1,101	
Grades 3 to 8								
Treatment effect	.220*	.044	.085*	.043	.212*	.046	.055	.044
Number of schools	8		8		7		7	
Number of students	2,535		2,538		2,136		2,140	
Grades 3 to 6								
Treatment effect	.189*	.048	.106*	.047	.178*	.051	.072	.049
Number of schools	8		8		7		7	
Number of students	2,309		2,306		1,910		1,908	

(continued)

TABLE 5 (CONTINUED)

	ITT				TOT			
	Mathematics		Reading		Mathematics		Reading	
Variable	Estimate	*SE*	Estimate	*SE*	Estimate	*SE*	Estimate	*SE*
Small town								
Grades K–8								
Treatment effect	–.093	.058	–.068	.055	–.079	.066	–.086	.063
Number of schools	9		9		6		6	
Number of students	2,602		2,593		2,235		2,226	
Grades K–2								
Treatment effect					.132	.117	–.101	.116
Number of schools					6		6	
Number of students					944		946	
Grades 3 to 8								
Treatment effect	–.087	.071	.029	.066	–.099	.083	.026	.077
Number of schools	9		9		6		6	
Number of students	1,658		1,647		1,291		1,280	
Grades 3 to 6								
Treatment effect	–.087	.071	.029	.066	–.099	.083	.026	.077
Number of schools	9		9		6		6	
Number of students	1,658		1,647		1,291		1,280	

Note. Model includes student and school variables, and grade dummy variables. ITT = intention to treat; TOT = treatment on treated; *SE* = standard error.
$^*p \leq .05$.

weak evidence about variability of the treatment effect across school urbanicity categories. The TOT results were overall similar to the ITT results. The TOT K–2 results, however, were positive and significant in suburban and urban schools.

Analyses Within Grades

We also conducted analyses within each grade separately (i.e., kindergarten through sixth grade) to determine grade-specific treatment effects.[7] Results from the ITT analyses are summarized in Table 6 for Grades 3 through 6. The overwhelming majority of the treatment effect estimates were positive, except for some estimates in kindergarten, first, and sixth grades. The estimates, however, were not significant at the .05 level in most grades. When student and school covariates were included in the model, the treatment effect estimates in reading were significant and nearly one seventh of an *SD* in the third and fourth grades. The treatment effects were positive and significant in the fifth and sixth grades in mathematics, especially when covariates were included in the model. The estimates in mathematics were consistently larger than one fourth of an *SD*, which is a considerable effect.

The results from the TOT analyses for Grades K–6 are summarized in Table 7. Overall, the results were similar to those reported in Table 6. In kindergarten, first, and second grades, the estimates of the treatment effect were smaller and not different from zero. The estimates were larger in Grades 3 to 6 and the treatment effect was statistically significant in the third grade when covariates were included in the model in mathematics and reading. This result was replicated in the fifth grade. The treatment effect estimate in sixth grade mathematics was significant when covariates were included in the model. Similarly, the treatment effect estimate in fourth grade reading was significant when covariates were included in the model. The estimates in fifth- and sixth-grade mathematics were nearly one third of an *SD*, which is a considerable effect. In reading, the estimates were slightly smaller than one fifth of an *SD*. Generally, the within-grade analyses yielded

TABLE 6

Within-Grade Estimates of the Treatment Effect for Mathematics and Reading Achievement: ITT Analysis

	Model I				Model II			
	Mathematics		Reading		Mathematics		Reading	
Variable	Estimate	*SE*	Estimate	*SE*	Estimate	*SE*	Estimate	*SE*
Grade K								
Treatment effect	−.013	.144	−.089	.131	.029	.107	−.071	.1345
Number of schools	38		38		37		37	
Number of students	2,441		2,460		2,229		2,247	
Grade 1								
Treatment effect	−.031	.092	−.025	.094	.096	.092	.113	.095
Number of schools	38		38		38		38	
Number of students	2,510		2,510		2,290		2,288	
Grade 2								
Treatment effect	.001	.128	.016	.116	.067	.101	.107	.081
Number of schools	44		44		44		44	
Number of students	2,693		2,701		2,429		2,437	
Grade 3								
Treatment effect	.060	.095	.061	.086	.133	.088	.156*	.064
Number of schools	57		57		57		57	
Number of students	3,608		3,604		3,442		3,439	
Grade 4								
Treatment effect	.087	.111	.058	.079	.136	.092	.135*	.057
Number of schools	57		57		57		57	
Number of students	3,583		3,575		3,441		3,432	
Grade 5								
Treatment effect	.255*	.106	.085	.092	.293*	.097	.130	.066
Number of schools	56		56		56		56	
Number of students	3,401		3,394		3,275		3,268	
Grade 6								
Treatment effect	.253	.163	−.159	.130	.323*	.154	−.008	.100
Number of schools	26		26		26		26	
Number of students	1,545		1,542		1,474		1,471	

Note. Model I includes the treatment only; Model II adds student and school variables. ITT = intention to treat; *SE* = standard error.

*$p \leq .05$.

some interesting findings, which suggest some positive and significant treatment effects in Grades 3 to 6.

Finally, the results for Grades 4, 5, and 6 that included prior scores as a covariate are reported in Table 8. The results were similar to those reported in Tables 6 and 7 for reading in Grade 4 and mathematics in Grade 5. However, Grade 6 results for mathematics are different (i.e., not significant). It should be noted that the sample size in Grade 6 is much smaller and thus the prior scores effect may have been different.

Discussion

Our findings overall suggest that the treatment effect was positive but not consistently significant across all grades (K–8). The treatment effect was smaller in lower grades (i.e., kindergarten to second grade) and larger in upper grades (i.e., third grade to eighth grade). Significant treatment effects were observed in Grade 3–8 analysis in mathematics, especially with covariates included in the model. This is in congruent with findings reported by Carlson et al. (2011) who found significant treatment

TABLE 7

Within-Grade Estimates of the Treatment Effect for Mathematics and Reading Achievement: TOT Analysis

	Model I				Model II			
	Mathematics		Reading		Mathematics		Reading	
Variable	Estimate	*SE*	Estimate	*SE*	Estimate	*SE*	Estimate	*SE*
Grade K								
Treatment effect	−.013	.145	−.089	.131	.030	.146	−.070	.135
Number of schools	38		38		37		37	
Number of students	2,441		2,460		2,229		2,247	
Grade 1								
Treatment effect	.036	.123	.053	.125	.096	.093	.112	.095
Number of schools	38		38		38		38	
Number of students	2,510		2,510		2,290		2,288	
Grade 2								
Treatment effect	.002	.128	.016	.116	.068	.100	.107	.081
Number of schools	44		44		44		44	
Number of students	2,693		2,701		2,429		2,437	
Grade 3								
Treatment effect	.121	.098	.114	.094	.190*	.090	.190*	.068
Number of schools	49		49		49		49	
Number of students	3,224		3,220		3,079		3,076	
Grade 4								
Treatment effect	.161	.121	.103	.089	.187	.098	.166*	.060
Number of schools	49		49		49		49	
Number of students	3,200		3,192		3,074		3,065	
Grade 5								
Treatment effect	.338*	.115	.111	.104	.366*	.104	.155*	.074
Number of schools	48		48		48		48	
Number of students	3,018		3,011		2,903		2,896	
Grade 6								
Treatment effect	.297	.172	−.187	.136	.377*	.167	−.062	.100
Number of schools	24		24		24		24	
Number of students	1,368		1,365		1,308		1,305	

Note. Model I includes the treatment only; Model II adds student and school variables. TOT = treatment on treated; *SE* = standard error.

**p* ≤ .05.

effects in mathematics for another interim assessment system. The effects of the TOT analysis were more pronounced and typically significant in Grades 3 to 8 for mathematics and reading especially when covariates were included in the model. Taken together, the results are consistent in terms of the sign of the effect (i.e., positive) but inconsistent in terms of statistical significance. In Grades K–2, however, mCLASS did not affect mathematics or reading achievement significantly.

The within-grade analyses revealed that in fifth and sixth grade, the effect of the treatment on mathematics was positive, significant, and not trivial. The treatment effect was consistently as large as one fourth of an *SD* and indicated an important annual gain in mathematics achievement (see Hill, Bloom, Black, & Lipsey, 2008). The third- and fourth-grade estimates in reading were significant but smaller in magnitude than those in mathematics. The estimates of the TOT analyses were larger in magnitude and nearly one third of an *SD* in mathematics but smaller in reading. Given that the estimates are not significant in kindergarten through second grade, one would conclude that mCLASS had no important effect on student achievement. In contrast, it seems that there is some, although somewhat inconsistent, evidence of a positive Acuity effect in third through sixth grades. These

TABLE 8

Grade 4 to 6 Estimates of the Treatment Effect in Mathematics and Reading Achievement Controlling for Prior Scores

	ITT				TOT			
	Mathematics		Reading		Mathematics		Reading	
Variable	Estimate	*SE*	Estimate	*SE*	Estimate	*SE*	Estimate	*SE*
Grade 4								
Treatment effect	.095	.089	.139*	.044	.147	.093	.153*	.047
Number of schools	57		57		49		49	
Number of students	3,031		3,008		2,690		2,670	
Grade 5								
Treatment effect	.176*	.086	–.024	.060	.254*	.089	–.011	.068
Number of schools	55		55		47		47	
Number of students	2,765		2,754		2,421		2,410	
Grade 6								
Treatment effect	.192	.144	.017	.078	.259	.156	–.023	.075
Number of schools	26		26		24		24	
Number of students	1,279		1,275		1,123		1,121	

Note. Model includes student and school variables and prior scores; ITT = intention to treat; TOT = treatment on treated; *SE* = standard error.

*$p \leq .05$.

results cast some doubt on the assumption that the intervention was the same in lower and upper grades or that the intervention should be viewed as a unitary one. It is noteworthy, however, that the K–2 outcome measure was not a state accountability test as opposed to the ISTEP$^+$ used in Grades 3 to 8.

The estimates produced from the ITT and the TOT analyses were overall similar qualitatively. The TOT estimates, however, were larger and significant in more grades than the ITT estimates. Still, it is difficult to know whether the TOT estimates suffer for selection bias, because the two types of estimates are not that different. In addition, the tests we used to examine the degree to which random assignment was successful did not indicate any significant preexisting differences between treatment and control schools for either the ITT or the TOT analyses. The tests, however, were based on school means and, therefore, it is possible that differences could not be detected because of lower statistical power.

We also ran sensitivity analyses that omitted seventh- and eighth-grade data from the K–8 analyses, but the estimates were very similar to the ones reported here. This was expected because the data from these grades were scarce. For Grades 4, 5, and 6, we also ran models that included prior student achievement as a covariate and the results were similar to those reported here. We also conducted analyses using data from Grades 3 to 6 only, as well as using data within each school locale type (e.g., urban, suburban), and the results were very similar to those reported in the “Results” section. Other analyses were conducted including and excluding students who joined participating schools while the treatment was being implemented or students who left the experiment while it was in progress. The results of these analyses were very similar to those reported here, indicating that student movement in and out of the participating schools did not affect the treatment estimates materially. Overall, it appears that the estimates are robust and show positive treatment effects in Grades 3 to 6.

In making sense of these results, it is critical to remember that the implementation of Indiana’s system of diagnostic assessments did not take place in a vacuum. As alluded to above, in Indiana, as in many other states at this time, 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 and classrooms. Most teachers in control and treatment 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 this light, the findings for the middle grades become particularly impressive, given that the treatment effects may be underestimates. And, in an end-of-year survey administered to treatment teachers, one third said that 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.

In the earlier grades, despite their positive signs, the estimates were quite small. It is probable that schools and teachers stressed the importance of doing well on the ISTEP$^+$ to students in Grades 3 to 8—it is after all the state accountability test and determined a student's proficiency and a school's accountability rating. The same was not true of the Terra Nova, which was introduced just for this study only and had no accountability repercussions. Another possibility is that the alignment of the mCLASS and the Terra Nova is less strong than that between Acuity and the ISTEP$^+$.

Causal inferences are warranted in this study because of the quality of the field experiment. However, the external validity of the results may be limited. Specifically, our sample was drawn from a subset of Indiana elementary schools that volunteered to implement the Diagnostic Assessment Tools intervention. Unfortunately, we do not know what motivated these schools to volunteer. Thus, there is reason to be cautious about generalizing our results beyond this group of schools that aspired to use technology-supported interim assessments in this Indiana context 2009–2010.

Authors' Note

The opinions expressed are those of the authors and do not represent the position or policies of the U.S. Department of Education, or of the Indiana Department of Education (IDOE). The authors would like to thank Wes Bruce and Michele Walker of the IDOE for their unwavering support of this study; and David Stevenson of Wireless Generation and Linda Bond and David Needles of CTB/McGraw-Hill for their support and encouragement. We would also like to extend our appreciation to our smart and challenging advisors: Tom Cook, Larry Hedges, Leslie Nabors-Olah, and Jeff Wayman; our outstanding study team: Jared Eno, Cheng-Hsien Li, Wei Li, Feng Liu, Coby Meyers, Brenna O'Brien, Andrew Swanlund, Bing Tong, and Anne Traynor; and two EEPA (*Educational Evaluation and Policy Analysis*) reviewers.

Declaration of Conflicting Interests

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

Funding

The authors 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. Detailed mappings of mCLASS measures and scales to Indiana content standards and to expected scope-and-sequence statewide may be found at http://www.doe.in.gov/achievement/assessment/mclass-k-2 and for Acuity at http://www.doe.in.gov/achievement/assessment/acuity-grades-3-8-algebra-i-english-10
2. Because of the requirements of the federal funding sources that Indiana used, schools were required to pay for the personal digital assistants (PDAs). This was an obstacle for a few schools. In 2008 to 2009, Wireless Generation used Palm PDAs; in the subsequent years, the software was ported to a variety of Android and Apple smart phones and tablet computers.
3. CTB/McGraw-Hill is also the developer of and vendor for the ISTEP$^+$ (Indiana Statewide Testing for Educational Progress–Plus). No detailed statistical analyses on the accuracy of Acuity's ISTEP$^+$ predictions have been released so far.
4. Indiana required NCLB (No Child Left Behind) differentiated accountability pilot schools to use mCLASS and Acuity. School-grade spans vary: Middle schools do not operate Grades K–2 and therefore could not participate in mCLASS. Our sampling design followed Indiana's design and intent: whole-

school implementations and use of both vendors' products.

5. Control schools agreed to accept a delayed treatment. Indiana Department of Education (IDOE) guaranteed them full access to mCLASS and Acuity in 2010 to 2011 and study funds paid for vendors to provide more intensive trainings of staff and support personnel.

6. Indiana's ISTEP$^+$, like most state tests, does not extend below Grade 3. The study therefore administered the Terra Nova. We selected this instrument because it is developed and maintained by CTB/McGraw-Hill's assessment unit, which also develops and maintains the ISTEP$^+$.

7. Estimates are not presented for Grades 7 and 8 because too few sampled schools enrolled students in these grades.

References

Black, P., & Wiliam, D. (1998). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan, 80*, 139–144.

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.

Blank, R., Porter, A., & Smithson, J. (2001). *New tools for analyzing teaching, curriculum, and standards in mathematics and science* (Report from Survey of Enacted Curriculum Project, National Science Foundation REC98-03080). Washington, DC: Council of Chief State School Officers.

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.

Bulkley, K. E., Christman, J. B., Goertz, M. E., & Lawrence, N. R. (2010). Building with benchmarks: The role of the district in Philadelphia's Benchmark Assessment System. *Peabody Journal of Education, 85*, 186–204.

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, October). *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.

Coe, R. (2002). The role and impact of performance feedback in schools. In A. Vischer & R. Coe (Eds.), *School improvement through performance feedback* (pp. 3–26). Lisse, Netherlands: Swetz & Zeitlinger.

Cordray, D., Pion, G., Brandt, C., Molefe, A., & Toby, M. (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 Sciences, U.S. Department of Education.

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.

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 & Evaluation, 14*(7), 1–11.

Faria, A., Heppen, J., Li, Y., Stachel, S., Jones, W., Sawyer, K., . . . Palacios, M. (2012). *The use of interim assessment data in urban schools: Links among data use practices and student achievement*. Washington, DC: American Institutes for Research.

Glick, J. E., & White, M. J. (2003). The academic trajectories of immigrant youths: Analysis within and across cohorts. *Demography, 40*, 759–783.

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.

Goren, P. (2012, February). Data, data, and more data—What's an educator to do? *American Journal of Education, 118*, 233–237.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*, 88–112.

Henderson, S., Petrosino, A., Guckenburg, S., & Hamilton, S. (2007). *Measuring how benchmark assessments affect student achievement* (Issues and Answers Report, REL 2007–No. 039). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Northeast and Islands.

Heritage, M. (2010). *Formative assessment: Making it happen in the classroom*. Thousand Oaks, CA: Corwin Press.

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*, 172–177.

Holland, P. W., & Rubin, D. B. (1982). *Test equating*. New York, NY: Academic Press.

Indiana Department of Education & CTB/McGraw-Hill. (2010). *Indiana statewide testing for educational progress, plus: ISTEP$^+$ 2010 technical report*. Indianapolis, IN: Author.

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

Kingston, N., & Nash, B. (2011). Formative assessment: A meta-analysis and a call for research. *Educational Measurement: Issues and Practice, 30*(4), 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.

Kolen, M. J., & Brennan, R. L. (1995). *Test equating: Methods and practices*. New York, NY: Springer.

Konstantopoulos, S. (2006). Trends of school effects on student achievement: Evidence from NLS:72, HSB:82, and NELS:92. *Teachers College Record, 108*, 2550–2581.

Li, Y., Marion, S., Perie, M., & Gong, B. (2010). An approach for evaluating the technical quality of interim assessments. *Peabody Journal of Education, 85*, 163–185.

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.

Nabors-Olah, L., Lawrence, N. R., & Riggan, M. (2010). Learning to learn from benchmark data: How to analyze results. *Peabody Journal of Education, 85*, 226–245.

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.

Porter, A. (2002). Measuring the content of instruction: Uses in research and practice. *Educational Researcher, 31*, 3–14.

Quint, J., Sepanik, S., & Smith, J. (2008). *Using student data to improve teaching and learning: Findings from an evaluation of the Formative Assessments of Student Thinking in Reading (FAST-R) program in Boston elementary schools*. New York, NY: MDRC.

Sadler, D. R. (1989). Formative assessment and the design of instructional strategies. *Instructional Science, 18*, 119–144.

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

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin.

Slavin, R. E., Cheung, A., Holmes, G. C., Madden, N. A., & Chamberlain, A. (2011). *Effects of a data-driven district reform model*. Baltimore, MD: Johns Hopkins University's Center for Research and Reform in Education.

Stiggins, R. (2002). Assessment crisis: The absence of assessment for learning. *Phi Delta Kappan, 83*, 758–767.

Tobias, S. (1987). Learner characteristics. In R. Gagne (Ed.), *Instructional design: Foundations* (pp. 207–231). Hillsdale, NJ: Lawrence Erlbaum.

Authors

SPYROS KONSTANTOPOULOS is associate professor of measurement and quantitative methods in the Department of Counseling, Educational Psychology, and Special Education at Michigan State University. His research interests include experimental design, program evaluation, and teacher and school effects.

SHAZIA RAFIULLAH MILLER is a managing director at American Institutes of Research and leads the State and Local Evaluation Center. She specializes in conducting rigorous evaluation studies with practical implications. A major professional focus is data driven decision making.

ARIE VAN DER PLOEG is a principal researcher at the American Institutes for Research. He is interested in the application of quantitative data to decision making in educational policy and practice.

Manuscript received August 13, 2012
Revision received June 13, 2013
Accepted June 17, 2013