Investigating Teacher Use of Practice Tests for Formative Purposes

Sarah M. Bonner
Hunter College, CUNY

ABSTRACT: This paper describes an intensive professional development program focused on supporting teachers at high-need urban high schools in use of practice tests for formative assessment purposes. Teachers engaged in professional development while offering intensive summer school to students with histories of failure on state tests in science and mathematics. Specific skills that teachers developed included item analysis of practice tests, planning around assessment results, and providing feedback to students. The study used a quasi-experimental pretest/posttest design with an intervention and a comparison group, analysis of classroom artifacts, follow-up surveys, and long-term follow-up with teachers. Students in the summer programs experienced large gains in performance, and teachers reported more frequent and varied use of tests for formative purposes, even one-and-one-half years after the end of training.

This material is based upon work supported by the National Science Foundation's Math and Science Partnership program under Grant No. 0412413

KEYWORDS: *assessment, formative evaluation, inservice teacher education, item analysis, secondary education test preparation*

This study inquired into the practices of high school teachers in urban public schools who were engaged in professional development around interpreting assessment results during intensive summer institutes for teachers and students in the science, technology, engineering, and math (STEM) disciplines. Funded by the National Science Foundation, a large urban university and several urban public schools partnered to conduct the summer institutes beginning in 2005. At the same time that the professional development team worked to improve teacher practices during the institutes, high school students with prior histories of failure in state-required mathematics and science content areas attended classes. This study reports on results from the summer of 2007, when professional development was focused on formative assessment. High school teachers worked in small teaching teams to analyze student work samples daily, and were intensively supported in using item analysis of student responses to practice tests for formative purposes, e.g., to identify student learning weaknesses and plan appropriate feedback. While professional development was offered to teachers of multiple STEM courses, the focus of this report is on the teachers and students in introductory levels of high school mathematics and biology.

The summer institutes were venues for teacher professional development, but their proximal and most obvious goal was to help students pass state examinations. Particularly in the case of high school introductory mathematics and biology, passing state examinations is a gateway to more advanced course choices, high school graduation, access to college, and broad career options. Students in the summer institute programs were those at high risk of school dropout or failure: low achievers with histories of school failure, and low-income and minority students. Unless students in these vulnerable groups pass through the gateway of state examinations, they will be unable to compete in most professional and academic contexts in American society today. Starting in 2005, the summer institutes were highly successful in meeting this proximal goal; in some cases as many as 100 percent of students in a class successfully passed the state tests, and the success rate rarely fell below 80 percent, much higher than in other summer programs.

My purpose was to assess the degree to which teachers in the summer institutes that included professional development in formative assessment changed their formative practices, how they did so, and the effect of their practices on student learning. This study took advantage of a naturally occurring quasi-experimental condition in the summer of 2007, when a version of the summer institute was offered at two sites without the component of intensive teacher professional development. The following research questions were addressed: a) Did teachers who received the professional development use practice tests and feedback differently than teachers without the professional development? b) Did students whose teachers received professional development have better performance outcomes, compared to students whose teachers did not? c) For teachers who received the professional development, what specific techniques related to formative assessment did they adopt, and how did they implement those techniques? d) Did teachers who received the professional development show sustained, more frequent use of practice tests for formative purposes after the conclusion of training?

Background

Research suggests that students learn more when their teachers use formative assessments to interpret student work in terms of what it says about student achievement and when teachers use their interpretations to provide clear descriptive feedback tied to specific and clearly communicated learning objectives (Black & Wiliam, 1998; Rodriguez, 2004). These effects of quality formative assessment practices are often greatest for low-achieving students. In order to interpret student work accurately, teachers need to be skilled at identifying a range of causes that may contribute to student errors. For instance, a mathematics teacher highly skilled at formative assessment would be aware of a breadth of ways students might organize and structure mathematical learning. Furthermore, teachers need to plan and implement actions based on the results of their assessments. Assessment-based actions should provide students feedback, which should be frequent, interactive, and informational (Bangert-Drowns, Kulik, Kulik, & Morgan, 1991; Butler & Winne, 1995; Hattie & Timperley, 2007).

Little quantitative, empirical research has been done about how and under what circumstances teachers develop good formative learning environments in their classrooms (Blumenfeld, 1992). Most of the few published descriptions of systematic formative assessment practices in the content areas have been conducted in the United Kingdom, often in primary schools (Torrance & Pryor, 2001; Wiliam, Lee, Harrison, & Black, 2004). In mathematics, some research relevant to analysis of student work has been done by researchers engaged in inquiry into the construct of pedagogical content knowledge. Pedagogical content knowledge, also known as and referred to henceforth as content knowledge for teaching, is that part of the knowledge domain of a teacher that "embodies the aspects of content most germane to its teachability" (Shulman, 1986). Most conceptualizations of content knowledge for teaching specifically reference the ability to interpret student work (e.g., Floden & McCrory, 2007; Hill, Rowan, & Ball, 2005). Many of the empirical studies related to content knowledge for teaching are in the domain of mathematics. One expert-novice study in this area (Leinhardt & Greeno, 1986) found that experts allocate more time for feedback. A larger study found content knowledge for teaching (with teacher analysis of student work as a component) had significant effects on first- and third-grade student mathematics achievement (Hill et al., 2005).

Under certain circumstances, practice tests can be used for such formative assessment purposes. Claims are often made that practice testing results in a net shrinking of the curriculum and a focus on superficial, short-term learning. However, in most published reports, when practice tests have been included in the classroom formative assessment environment, positive effects on student learning have been found. A large meta-analysis (Bangert-Drowns et al., 1991) reviewed the evidence of the effects of frequent class testing and showed that, in general, performance improved with frequent testing and increased with increased frequency up to about once or twice a week. The expected effect of practice tests on student learning should be considered in light of type of learning outcome, frequency of testing, intervening test format, immediacy and quality of feedback, individual student and class characteristics, and characteristics of teachers. Martinez and Martinez (1992) found that frequent testing led to improved learning, especially for less experienced teachers. A positive effect of formative testing was found for teacher candidates in special education who were administered a short formative quiz after lectures (Schloss, Smith, & Posluzsny, 1990). Those who were quizzed performed significantly better than those candidates for whom no quiz was administered across multiple posttest measures. Kang, McDermott, and Roediger (2007) also found that an intervening test prior to a summative assessment was associated with greater performance, especially when corrective feedback was used. The testing effect appeared to be greater than mere exposure to the material in a nontesting format (through reading passages, for instance) would explain (McDaniel, Anderson, Derbish & Morrisette, 2007).

Rubenstein (2004) explained ways that practice tests specifically address the causes of student poor performance on large-scale assessments. He attributed poor test performance to five problem areas: poor test strategies, lack of problem-solving skill, lack of practice, lack of test-taking stamina, and lack of basic skills. Use of practice tests and test preparation techniques that address any of these areas may improve test performance in the short term, while some types of test preparation, for instance those that remediate basic skills rather than those that focus on test strategizing, may extend performance improvements beyond a single outcome measure. In the intervention reported in this study, practice tests of mixed short-answer and multiple-choice formats were administered weekly to high school students who were fairly homogeneous in skill level on the tested content. Practicing with tests that were highly similar in format, length, and content to the official state examinations was intended to build student stamina and awareness of content learning expectations. The professional development model stressed feedback in the form of teacher modifications to learning plans and specific instruction around student strengths and weaknesses. Basic skills in the content area rather than strategic skills were the emphases of classroom instruction.

Methods

Because the professional development program was not willing to engage in a true experimental design that would require disentangling facets of its complex program package, I sought to take advantage of naturally occurring variability in implementation of professional development over time and between sites, as well as use retrospective self-report of teacher practices before and after the intervention. I elaborated this quasi-experimental pretest/posttest design with an intervention and a comparison group through analysis of classroom artifacts, follow-up surveys, and long-term follow-up with teachers.

Summer Institutes: Teacher Professional Development and Comparison Sites

The two professional development (PD) summer institute sites were located at urban college campuses and included the following components: all-day summer school of five weeks' duration, professional development support for teachers in analyzing student work and using practice tests to guide instruction, administration of weekly practice tests, and tutors in classes. Two comparison (no-PD) sites held summer institutes but provided little teacher professional development. These sites were located at high school campuses and had a shorter school day. At the PD sites, students and teachers engaged in multiple activities related to formative assessment of student work from which multiple artifacts were collected. At all sites, PD and no-PD, students took a complete three-hour retired form of the state examination every Thursday for the duration of the summer institute. However, only at the PD sites were teachers frequently reinforced in analyzing the results from these practice exams, and given facilitated planning time specifically intended to allow them to plan feedback and instructional responses to the results.

Participants

High school teachers of Math A, Living Environments, Math B, and Chemistry who were involved in the summer institutes in 2007 participated in the study. Math A was the name then-current in the state for the basic course in integrated Algebra and Geometry and its associated state examination, both of which are required for high school graduation in the state. Living Environments is the name for the basic course in biology and its associated state examination; passing at least one science-related course and examination is required for high school graduation in the state, and most students meet the requirement through taking the Living Environments course and examination. Math B and Chemistry are more advanced courses; they are not required for regular graduation, nor are their examinations state-mandated.

Participating teachers were employed at high-need, under-performing schools, and had classes that experienced low rates of success on state examinations during the regular school year. Forty-one teachers provided most of the data for this study; the majority (71%) were teachers of Math A or Living Environments. Thirty-five teachers participated at one of the PD sites, while six taught at a no-PD site. The 41 teachers were 51 percent female and 51 percent white non-Hispanic. Teachers at the PD and no-PD sites were similar in terms of gender and ethnic background, years of experience, and degree of professional training.

In summer 2007, 552 students participated in the summer institutes. Students were enrolled at schools served by the program, which were in general low-performing schools. Students had either attempted the state examination in the content area and failed it, or failed to earn credit in the academic course with which the state examination was aligned. Most students met both criteria. Of 546 students responding to an item about gender, 47 percent were female. Of 483 students responding to an item about ethnic background, 45 percent were African-American and 45 percent were Latino/a. There were no significant differences relating to gender between students at the PD and no-PD sites; however, proportionally larger numbers of African-American students attended one of the no-PD sites. The composition of student groups was similar across all courses offered in the summer institutes. The 161 PD site students for whom pretest scores were available had on average lower scale test scores on entry into the summer institute (M = 44.52, SD = 12.07) compared to 72 students at the no-PD sites (M = 49.82, SD = 14.93), and the difference was statistically significant (t = -2.875, p < 0.1).

Measures

Measures for teachers included a short questionnaire on experience, professional training, and other background variables, and a survey of classroom practices and expectations about student learning. Three dimensions were examined: Inquiry Methods, Use of Practice Tests, and Expectations for Students. Each of the first two dimensions was measured with a scale composed of a group of statements, developed by the author, to which the teacher responded on a five-point Likert-type scale. On the pre-institute survey, teachers were asked to reflect on their regular, school-year classroom experiences; on the post-institute survey, teachers were asked to reflect on the summer institute experience. Although teacher use of Inquiry Methods was not a targeted area of professional development in the summer institute in 2007, Inquiry Methods was examined because they had been targeted areas in previous summers and were still core to the professional development model. Expectations for Students was measured with items adapted from the Motivated Strategies for Learning Questionnaire (MSLQ), developed by Pintrich, Smith, Garcia, and McKeachie (1993). Item numbers, sample items, and Cronbach's alpha for each scale are reported in Table 1. The internal consistency evidence suggests the scales showed sufficient internal consistency to be used to measure differences between groups.

Table 1
Survey Scale Characteristics

Scale	Sample item	*n* of Items	α (teacher)	α (student)
Inquiry Methods	My lab activities motivated students' curiosity to learn.[a]	4	0.85	0.81
Use of Practice Tests	I used information from students' practice tests to show students how their learning was progressing.[b]	5	0.85	0.86
Expectations for Students	I believe all students in this course will be capable of learning the basic concepts taught in the course. [a]	3	0.76	0.71

Note: [a] Response categories ranged from Strongly Agree (5) to Strongly Disagree (1).
Note: [b] Response categories ranged from Very Often (5) to Almost Never (1).

Students completed similar surveys at the end of the summer institute. Validity of teacher's self-report of classroom instructional practices on surveys was thus checked through triangulation with student reports. Items on the student surveys paralleled items on the teacher surveys, with appropriately modified stems, e.g., "My teacher used practice tests to show me how my learning was progressing." Student versions of these scales were similar in reliability to the teacher scales, as also reported in Table 1.

Students enrolled in the summer institute at both the PD and no-PD sites were pre- and posttested on achievement in their summer school content area, using (for the pretest) a retired form of the state examination in the content area of their summer coursework, and (for posttest) the official state test.

Artifacts from the PD sites included Excel spreadsheets documenting the teachers' item analyses of the weekly practice tests; teacher documentation of analysis of student classroom seatwork; teacher lesson plans following administration of practice tests, which would show whether and how the teachers planned feedback responses to assessment results; and digital video of classroom instruction. As part of professional development practices during the summer institutes, teachers were routinely videotaped; therefore, videotapes of teaching during lessons that followed practice test analysis were available for review, to assess whether and how teachers implemented feedback. These videotapes were minimally invasive as teachers in the PD program rapidly become inured to the constant presence of videographers and were unaware of the multiple ways in which the records were to be analyzed. Altogether, the artifacts provided unobtrusive evidence to validate teacher self-reports, enabled inquiry into the relationship between self-perception and practice, and provided illustrative examples of effects of the intervention.

Open-ended survey and interview questions were administered to teachers at both types of sites at multiple time-points, as described below. These were intended to reveal examples of teacher thinking regarding use of practice tests and analysis of student work and information about teachers' self-reported long-term incorporation of formative use of practice tests into their classroom instruction.

Data Collection

Data collection was complicated during the summer institutes due to multiple sites and the intensive pace of the program, especially at the no-PD sites, which were loosely administered by high school site supervisors. Only teachers at the PD sites received and responded to surveys about prior practices (pre-surveys). Teachers and students at both types of sites were, however, surveyed on the final day of the summer institute. Four months after the end of the summer institute, teachers were surveyed again on their formative practices. Finally, a number of teachers (*n* = 18) responded to interviews one-and-one-half years after the summer institute, providing a glimpse of long-term retention and use of the skills they had acquired.

Results

Paired-sample *t* tests were used to evaluate possible differences between PD site teacher perceptions when reflecting on regular school year instruction (pre-institute) and summer institute instruction (post-institute), as shown in Table 2. Teachers at the PD sites (*n* = 35) reported differences in how they used practice tests during the summer institute compared with their practices during the school year, reporting more use of practice tests to give feedback to students, plan whole group instruction, and differentiate instruction during the summer institute. Differences between school year and summer institute perceptions about use of Inquiry Methods, a dimension of pedagogy that was not directly targeted by professional development, were not statistically significant. Comparing Expectations for Students at the beginning and end of the summer institute, teachers at PD sites reported statistically lower Expectations for Students at the end of the summer institute (see Table 2).

Table 2
Paired Sample *t*-Test Results for PD Site Teacher Perceptions

Scale	Reference	*n* pairs	*M* (*SD*)	*t*	*d*
Practice Tests	Reflecting on school year	35	4.09 (0.73)	2.29*	0.47
	Reflecting on summer institute		4.43 (0.70)		
Inquiry Methods	Reflecting on school year	23	3.65 (0.63)	1.86	0.67
	Reflecting on summer institute		4.04 (0.57)		
Expectations for Students	Before summer institute	35	4.15 (0.69)	-4.55**	-0.74
	At end of summer institute		3.58 (0.83)		

NOTE: * p < .05
NOTE: **p < .01

Because of very unequal sample sizes and low variability in responses for teachers at the no-PD site who responded to surveys at the end of the institute (suggesting a possible response set), comparisons of survey responses between teachers at the PD and no-PD sites are not reported. Evidence supporting the inference that teachers at the PD sites made greater use of practice tests was found in student surveys. Student responses to post-institute surveys tended to support teacher perceptions that use of feedback and practice tests was frequent and varied at both types of sites. Students at the PD sites (*n* = 172) reported greater use of feedback on practice tests (*M* = 4.51, *SD* = 0.50), compared with students (*n* = 186) at the no-PD sites (*M* = 4.24, *SD* = 0.66), and the differences were statistically significant (Welch's *t* = 4.34, *p* < .01, *d* = 0.46).

Data indicated that students at the PD sites experienced somewhat better outcomes (Table 3). Combining scale scores across content areas, for those students for whom pre-institute test scores were available, students who had attended summer institutes at the PD sites (*n* = 153) made significantly greater gains in test scores compared with those who had attended summer institutes at the no-PD sites (*n* = 57) (*t* = 3.98, *p* < .01).

Table 3
Student Test Results by Content Area and Site Type

Content Area	Site	*n*	First Practice *M*(*SD*)	Official *M*(*SD*)	*t*	*d*
Math A	PD	54	50.06 (8.46)	63.65 (9.35)	8.30**	1.50
	No-PD	41	54.15 (9.31)	61.85 (10.22)	4.28**	0.78
Living Environments	PD	58	42.09 (14.12)	70.52 (9.17)	17.75**	2.37
	No-PD	16	39.00 (21.12)	68.25 (13.57)	7.42**	1.61

Note: **p < .01

The passing rate on the official state examination in August for students at the PD sites was 87 percent, while the passing rate at the no-PD sites was 80 percent. Both rates compare favorably with passing rates from traditional summer schools. The average official state scale score was also significantly higher for students at the PD sites (*t* = 2.27, *p* = .02).

Information about teacher practices was obtained through analysis of teacher responses(n = 42) to open-ended surveys at the end of the summer institute; this included the 35 teachers from the PD sites who also completed the scales reported above, one PD site teacher who only completed the open-ended items, and six no-PD site teachers. All teachers at both types of sites claimed that during the summer institute they "looked at actual student work to understand strengths and weaknesses in student thinking" at least daily, with 17 percent claiming they looked at student work more frequently than on an hourly basis. Teachers stated that they examined student work during virtually all summer institute activities. Approximately 95 percent of teachers reported they engaged in looking at student work during two or more activities in the day.

Teachers were asked for which purposes they reviewed or analyzed student work on practice tests, both during the school year and during the summer institute. During the summer institute, the least frequently reported purpose for reviewing practice tests was to assign grades (reported by 24 percent of teachers). This contrasts with their self-reported purposes for reviewing tests and quizzes during the school year, when the most frequently reported purpose was to assign grades (reported by 93% of teachers). Reported school year and summer institute use of tests and quizzes to understand students' strengths and weaknesses was very similar, with 83 to 86 percent of teachers reporting they used assessments in this way at each time. Thirty-six teachers reported use of tests and quizzes to plan instruction during the school year, while thirty teachers claimed they used practice test results to plan instruction during the summer institute.

Reflecting on their experiences during the summer institute, teachers at the PD sites reported more specific teaching responses to results of practice test analysis, compared with teachers not receiving PD. For instance, one teacher reported that during the summer institute, "We planned the lessons out, especially on Monday after the results (from the practice tests) were in. We broke our students down in groups, where they worked on different level of questions and different topics." Another teacher stated, "We did an item analysis and concentrated on the question types which most students scored low on … to direct instruction. We also looked at the method of problem solving from each student on the long-answer questions. Common mistakes were addressed in class and individual mistakes were addressed by the tutors who also looked over the … exams."

Analysis of teaching artifacts of PD site summer institute teachers revealed mixed results about teacher practice. Teachers were highly consistent in preparing Excel spreadsheets with student responses to practice test items, which were color-coded according to correct vs. incorrect responses to each item. On the other hand, teacher documentation of analyses of student daily seatwork was extremely varied in content specificity and attention to individual learners. For instance, one teacher invariably broke student work down by the individual student, noting each student's specific learning needs as evidenced by his/her work, and doing so even for those students who were working on the selected objective with relative proficiency. This teacher not only considered cognitive aspects of learning, but physical ones ("Student X needs to get more rest") and metacognitive ones ("Student Y is improving focus"). In contrast, some other teachers' daily analysis of seatwork was routine and repetitive. For example, one teacher included in every analysis the rote recommendation to "place students in mixed-ability pairs" and "give time with manipulatives."

Review of videos of instruction for classes immediately following analysis of practice tests yielded similarly varied use of feedback. In many cases, verbal feedback to the whole class, if given at all, was limited to praise, e.g., "Your scores keep getting better—you're on the right path to passing!" In some classes, awards were given each week in a ceremony for most improvement and highest score. Teachers emphasized continuous improvement towards passing, and recognition of students who had reached passing levels on practice tests.

Some teachers used the practice test results primarily to differentiate instruction and kept to their syllabus for whole group instruction. As they reported in follow-up surveys, some teachers relied on tutors to use the information from the analyzed practice tests for differentiated instruction, "We would pass on the areas of deficiency to the tutors to individualize instruction." Some also developed individualized assignments for students on the basis of their practice test interpretations: "Students would be given question sets on Mondays on the topics they had missed (on the practice test) from the Friday before."

In a few cases, teachers changed instruction considerably on the basis of the practice test results. One teacher reported that his/her class had been working on exponential and logarithmic functions, with disappointing results on the practice tests. Based on the test results, the teaching team decided to try a different approach to the topic. In a video segment from one class, a teacher reported to her class that the teaching team's analysis of the practice test results had revealed that few students were attempting the constructed response sections of the tests, which require verbal justifications for problem-solving steps. The teacher explained to the class that they were unlikely to pass the tests without earning at least partial credit on these types of items, but that verbal justification even of an incorrect solution was valuable because it was evidence of mathematical thinking. Therefore the teacher introduced a new instructional approach where students solved problems in a two-column format, writing the mathematical step in the left-hand column and its justification in the right-hand column.

Of the fourteen teachers who responded to requests to be interviewed one-and-one-half years after the 2007 summer institute, only one had been at a site without professional development. That teacher reported current use of practice tests, but described no methods of using the test results to plan whole group instruction, differentiate, decide teaching modifications, or motivate students. Of the thirteen PD site teachers who responded, most reported using practice tests in their regular classrooms one to three times per semester, with one teacher reporting weekly practice testing. All the interviewed PD site teachers stated that they continued to analyze their assessment results, although not always in a formal manner. Seven of these teachers reported that they continued to perform item-level data analysis on their practice test and other formative assessment results. One reported that he/she continued to use practice tests to show progress towards learning goals for motivation. All the PD site teachers reported that they gave their students feedback by going over student tests, identifying skills where students are successful and focusing instruction on misconceptions, common mistakes, and skills that need improvement. One teacher reported being "much more careful" than before training about reviewing and responding to student work. Multiple teachers reported that differentiating or individualizing based on the practice test results was challenging without the support of tutors, although one or two had developed classroom solutions that partly addressed those challenges.

Discussion

This study investigated ways that teachers in an intensive summer institute with high school students used and perceived the use of practice tests for formative assessment purposes. Professional development relating to the interpretation and use of practice tests was strongly implemented at two sites, where teachers were extensively trained in specialized techniques such as item analysis and engaged in daily analysis of student work. At two other sites, teachers were left largely to their own instructional devices, although they also administered weekly practice tests.

My first research question asked whether teachers who received the professional development used practice tests and feedback differently after training. Teachers reported that they used practice tests more often and in more varied ways after professional development than they did during the previous school year. The students' perceptions that teachers at the PD sites used feedback and practice tests more frequently and in more varied ways than other teachers substantiated this interpretation.

My second research question asked whether students whose teachers received professional development had better performance outcomes, compared to students whose teachers did not. While students at both PD and no-PD sites made strong gains in test scores and passed the state tests at high rates, students at the sites where teachers were receiving professional development made greater gains and passed at higher rates.

With my third research question, I sought to identify the specific techniques teachers adopted that related to formative assessment and how they implemented those techniques. Analysis of teacher self-reports and teaching artifacts from the PD sites showed that there was considerable variability in how teachers, all of whom received the same training, analyzed student work and provided feedback. There was variability in depth and specificity of analyses and variability in approaches to feedback, ranging from nontask-related feedback (e.g., praise) to item-specific whole-class feedback, feedback to tutors for differentiation, and feedback to teacher teams for making large changes to instructional approaches. The ways that PD site teachers took up skills related to formative assessment were likely moderated by individual differences in motivation and cognition and other factors such as competing demands on time.

The fourth research question asked whether teachers who received the professional development showed sustained, more frequent use of practice tests for formative purposes. PD site teachers who responded to long-term follow-up questions reported sustained, frequent use of varied formative assessments, including practice tests. Virtually all the PD site teachers who were interviewed provided examples of classroom adoption of at least one of the following skills targeted by the summer institute professional development: practice testing, item analysis of assessment results, feedback to students, planning instruction based on formative assessment results, and differentiating instruction based on assessment results. For some of the teachers who received this training, it is clear that there were both immediate and long term dramatic effects.

However, with little support during the school year, teachers struggled to find time to differentiate, give detailed feedback to students, and give students experience with a full three-hour practice test to build stamina and comfort with testing conditions. Some ways that teachers found to adapt their professional learning to their school environments included setting up work stations in their classrooms that students could move through for help on specific skills, limiting feedback to praise or extrinsic rewards, and administering only the multiple-choice sections of practice tests. In absence of continuing PD support, some teachers reported motivational, cognitive, and affective challenges around using practice tests results in their planning and instruction.

Some evidence was found that student achievement gains were at least partly attributable to teacher behaviors and due to more than a testing effect or the effect of tutoring. The program had the strongest impact on student achievement when the professional development component was strong, and weaker impacts without professional development, even when practice tests and tutoring were parts of the program. In summer 2008, when practice testing and tutoring continued to be used, but again without significant teacher professional development around formative assessment, achievement gains were again lower. This suggests that teacher training in interpretation and use of practice test results as one component of intensive instruction was associated with student learning gains.

Because this was not an experimentally controlled study, other explanations for performance outcomes are plausible. For instance, the overall performance gains common to both PD and no-PD sites in 2007 may be partly explained by the availability of tutors. However, it should be noted that the tutoring component of the program was inextricably linked to the use of practice tests in the summer institute; program tutors were themselves learners at far from expert levels of performance. Tutors' guidance for individual work with students was based on teacher interpretation of practice test results and analysis of other student work. Other plausible factors that may account for differences in outcomes between the two types of sites include duration of the school day and educational climate of the summer institute. The summer institutes at the no-PD sites were very similar to traditional school-year classes; the main differences between the no-PD summer institutes and traditional school being the presence of tutors and practice tests. There was no evidence that teachers used more inquiry methods in the summer institutes than in their regular classrooms. The student learning gains at the no-PD sites appear to speak in part to the power of the effect of repeated testing, as well as the usefulness of tutors.

Overall, while it is difficult to make causal claims about the relationship between student learning gains and teacher use of practice tests for formative purposes, these results suggest that in urban schools where many students have prior history of course failure in mathematics and/or science, providing opportunities for students to be exposed to full-length practice tests and training teachers in specific techniques to analyze and respond to test results may help improve student achievement.

It is likely that the results of this intervention would extend to types of formative assessment other than practice tests. In this particular intervention, practice tests had the merits of providing motivation due to state graduation requirements, a yardstick with a common metric so that students and teachers could easily monitor student progress towards goals, and a high degree of relatedness between test content and state standards and course curricula. However, the practice gained by teachers in attending to and analyzing student work, and responding through swift instructional feedback, could be extended to a variety of other assessment techniques. Long-term follow-up evidence in this study supports the possibility that such analytic skills generalize within teacher practices to impact how a range of student work is viewed. Few teachers interviewed one-and-one-half years after the intervention reported administering full-length practice tests; however, several of them reported using item analysis on student quizzes and examining student responses to open-ended homework and quiz questions to guide instructional decisions such as pacing and composition of student groups.

Given the difficulty of minimizing error and controlling for extraneous variables in a complex urban teaching and learning intervention, the main goal of which was to help at-risk students advance towards graduation and better chances of success in society, this report raises as many questions for further investigation as it answers. The gains made by students and the impacts on practice reported by teachers are sufficiently compelling to justify continued, more rigorously controlled investigation into when and how practice tests are most effectively used, the types of learning that are impacted most strongly by different approaches to practice testing, which kinds of students are most impacted, and the kinds of professional support teachers need to sustain their own learning and quality practice.

References

Bangert-Drowns, R. L., Kulik, C. C., Kulik, J. A., & Morgan, M. (1991). The instructional effect of feedback in test-like events. *Review of Educational Research, 61*, 213-238.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education, 5,* 7-74.

Blumenfeld, P. C. (1992). Classroom learning and motivation: Clarifying and expanding goal theory. *Journal of Educational Psychology, 84*, 272-281.

Butler, D. L., & Winne, P. H. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of Educational Research, 65*, 245-281.

Floden, R., & McCrory, R., (2007). *Mathematical knowledge for teaching algebra: Validating an assessment of teacher knowledge.* Paper presented at the annual meeting of the Association of Mathematics Teacher Educators, Irvine, CA.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*, 81-112.

Hill, H. C., Rowan, B., & Ball, D. H. (2005). Effects of teachers' mathematical knowledge for teaching on student achievement. *American Educational Research Journal 42*(2), 371-406.

Kang, S. H., McDermott, K. B., & Roediger, H. L. (2007). Test format and corrective feedback modify the effect of testing on long-term retention. *European Journal of Cognitive Psychology*, 19, 528-558.

Leinhardt, G., & Greeno, J. G. (1986). The cognitive skill of teaching. *Journal of Educational Psychology 78*(2), 75-95.

Martinez, J. G. R., & Martinez, N. C. (1992). Re-examining repeated testing and teacher effects in a remedial mathematics course. *British Journal of Educational Psychology, 62*, 356-363.

McDaniel, M. A., Anderson, J. L., Derbish, M. H., & Morrisette, N. (2007). Testing the testing effect in the classroom. *European Journal of Cognitive Psychology*, *19*, 494-513.

Pintrich, P. R., Smith, D. A. F., Garcia, T., & McKeachie, W. J. (1993). Reliability and predictive validity of the *Motivated Strategies for Learning Questionnaire* (MSLQ). *Educational and Psychological Measurement, 53,* 801-803.

Rodriguez, M. C. (2004). The role of classroom assessment in student performance on TIMMS. *Applied Measurement in Education, 17*, 1-24.

Rubenstein, J. (2004). Test preparation: What makes it effective? In J. E. Wall, & G. R. Walz (Eds.), *Measuring up: Assessment issues for teachers, counselors and administrators* (pp. 395-415). Washington, DC: ERIC Counseling and Student Services Clearinghouse.

Schloss, P. J., Smith, M. A., & Posluzsny, M. (1990). The impact of formative and summative assessment upon test performance of science education majors, *Teacher Education and Special Education Majors, 13*, 3-8.

Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher 15*(2), 4-14.

Torrance, H., & Pryor, J. (2001). Developing formative assessment in the classroom: Using action research to explore and modify theory. *British Educational Research Journal, 27*, 615-631.

Wiliam, D., Lee, C., Harrison, C., & Black, P. (2004). Teachers developing assessment for learning: Impact on student achievement. *Assessment in Education Principles Policy and Practice, 11*(1), 49-65.