# Investigating the Efficacy of a Professional Development Program in Formative Classroom Assessment in Middle School English Language Arts and Mathematics

This project was made possible through a grant from the Institute for Educational Sciences (IES) Teacher Quality Research Grant Program Award Number R305M050270B to the South Carolina Department of Education. The authors would like to acknowledge the contributions of the following persons who were involved in the project at various times: B.J. Miller, Susan Lottridge, Shelley Ragland, Pamela Kaliski, Jaime Cid, Dawn Mazzie, Pamela Gowan, and Lara Osleson.

M. Christina Schneider
*CTB/McGraw-Hill*

J. Patrick Meyer
*University of Virginia*

**Background:** Teachers who engage in formative classroom assessment using practices that accurately measure student learning should be better positioned to diagnose the instructional needs of their students and to act on that information. For this reason, there has been increased interest in formative classroom assessment in recent years. Although some researchers have found indications that some assessment practices may raise student achievement, evaluations of professional development programs designed to increase teacher assessment skill have not uniformly found differences in student performance. However, few studies of this type have been performed.

**Purpose:** The purpose of this study was to determine the efficacy of a professional development program in formative classroom assessment on teacher assessment knowledge and their students' achievement.

**Setting:** The professional development program was implemented in state-identified, low performing middle schools from November 2005 through April 2008.

**Intervention:** Researchers investigated a professional development program for teachers designed to increase their skill in creating and using assessments to support student learning. In Year 1, the professional development was implemented by an assessment coach in the treatment schools. No professional development was provided in the control schools. In Years 2 and 3, levels of treatment were investigated such that the professional development was implemented by an assessment coach or a relatively untrained facilitator.

**Research Design:** Year 1 involved a multi-site, cluster randomized trial where schools were randomly assigned to either the treatment or control group. Year 2 and Year 3 involved a quasi-experimental design.

**Data Collection and Analysis:** Researchers collected pretest and posttest teacher measures and analyzed the data using a split-plot ANOVA each year. Summative large scale assessment data was collected for students and analyzed using hierarchical linear modeling (HLM).

**Findings:** Findings from this study indicate that the professional development program increased

teacher assessment skill regardless of whether the program was implemented with a trained assessment coach or a relatively untrained facilitator. However, students of teachers participating in the professional development tended to demonstrate lower achievement than a matched set of students whose teachers did not receive the professional development. Implications for how teachers use assessment data to guide reteaching are discussed.

***Keywords:*** *formative assessment; classroom assessment; student achievement*

When teachers collect and analyze diverse types of evidence regarding individual student learning and use that information to either adjust instruction or provide feedback to students they are engaging in formative classroom assessment practices (Brookhart, Moss, & Long, 2008). Formative classroom assessment can have different connotations depending upon the framework being implemented. For example, Wiliam, Lee, Harrison, and Black (2004) defined formative classroom assessment as teacher questioning, comment only feedback, sharing grading criteria, and student self- and peer feedback in their study. These techniques tend to be embedded in the instructional sequence of the teacher. Brookhart (2010) described using formative feedback on summative classroom assessments. In this situation, the teacher uses the formative practice after the instructional sequence to provide students feedback on how to better target the expected outcome on a revision or on a subsequent assignment. Teachers may also use summative assessments to determine that reteaching may be necessary. Although formative classroom assessment practices can have differing proximities to the instructional sequence, what should identify an assessment practice as informative is whether or not the teacher and student effectively use the information to improve student achievement on the intended learning target (Brookhart, 2009; Nichols, Meyers, & Burling, 2009; Shepard, 2009).

Teachers who engage in formative classroom assessment using assessment construction practices that accurately measure student learning should be better positioned to diagnose the instructional needs of their students and to act on that information. Although teachers spend 20% to 30% of their time engaged in some form of assessment (Barton & Coley, 1994), researchers have found that teachers lack expertise in sound assessment practices (Marso & Pigge, 1993; Haydel, Oescher, & Banbury, 1995; Plake & Impara, 1997). Aschbacher (1999) found that among sampled middle school teachers, only one quarter to one third had coherent assessments. Aschbacher defined coherence as the extent to which teachers aligned the learning task for students to their stated learning goals and the criteria used to evaluate students' work. Incoherent assessment practices likely signify an issue larger than a lack of understanding of best practices for measuring student learning. Incoherent assessment practices may also indicate a teacher has not crisply defined the learning targets for either themselves or their students.

Although state standards often imply a range of cognitive complexity, and almost all standards are written above the recall level, many teachers are unsure how to interpret the cognitive levels in standards. As a result, teachers oftentimes focus their learning targets on having students recall knowledge (Oescher & Kirby, 1990; Sobolewski, 2002; Brookhart, 2005). Llosa (2005) found that teachers who are

unable to accurately interpret state standards ignore parts of standards they do not understand, develop their own interpretation for standards, or ignore standards entirely. Yap et al. (2007) found 34% of teachers in their study could not accurately interpret a state standard of their *own* choosing. This is likely one reason that researchers have found that many teachers are not providing assessments to match rigorous state standards (Fleming & Chambers, 1983; Carter, 1984; Marso & Pigge, 1993).

Items found on summative large scale assessments undergo standards alignment reviews that provide indicators showing the degree to which they reflect the content and cognitive rigor of the state standards (Webb, Herman, & Webb, 2007). Wolfe, Viger, Jarvinen, and Linksman (2007) noted that teachers should also align their own classroom assessments with the state standards, however, this recommendation is often not specifically addressed in the formative assessment and classroom assessment literature. As may be expected, assessment experts are more consistent in their interpretations of standards than are teachers (Nasstrom, 2009). This is likely because typically teachers are neither trained in how to analyze and interpret state standards nor in how to align items to the standards. The interaction between insufficient teacher training in formative classroom assessment practices and insufficient teacher training in the intended learning targets of standards must certainly play a role in how a teacher interprets and measures state standards in his or her classroom.

Although some researchers have found that using best practices when developing assessments for formative or summative purposes can enhance student achievement (e.g., Andrade, Du, & Wang, 2008; Newmann, Bryk, & Nagaoka, 2001; Ross, Hogaboam-Gray, & Rolheiser, 2002) the studies have not been based in the implementation of a comprehensive professional development program. A handful of researchers have begun using quasi-experimental designs to investigate the efficacy of professional development programs in formative assessment on student achievement (e.g., Bell, Steinberg, Wiliam & Wylie, 2008; Brookhart, Moss & Long, 2007, 2008; Meisels, Atkins-Burnett, Xue, Nicholson, Bickel & Son, 2003; Ragland, Schneider, Yap, & Kaliski, 2008; Wiliam, Lee, Harrison & Black, 2004). Student-achievement related research is sparse and has not supported strong causal conclusions regarding the effect of teacher professional development in formative assessment practices on student achievement (Schneider & Randel, 2010). Studies have typically not had large enough sample sizes, randomization of treatment effects, and many have not accounted for the nesting of student data. The effect of formative assessment professional development for teachers on student achievement continues to be a critical area of study. For this reason, we investigated the effect of a professional development program in formative classroom assessment on teacher assessment knowledge and their students' achievement, and we report three years of findings from the replication study.

## Purpose

In Year 1, the study's purpose was to determine (a) whether teachers who received the professional development in formative classroom assessment with a coach were more knowledgeable about measurement principles, cognitive levels,

and state standards than teachers who did not receive the professional development in formative classroom assessment and (b) whether middle school students of teachers who received professional development in formative classroom assessment practices with a coach had higher achievement than students of teachers who did not receive the professional development in formative classroom assessment.

In Year 2 and 3, the study's purpose was (a) to determine whether teachers who received the professional development in formative classroom assessment with a trained coach were more knowledgeable about measurement principles, cognitive levels, and state standards than teachers who received the same professional development from a relatively untrained facilitator and (b) to compare the achievement of students from three groups:

- Students whose teachers received professional development in formative classroom assessment with a coach.
- Students whose teachers received professional development in formative classroom assessment with a relatively untrained facilitator.
- Students whose teachers received no professional development in classroom assessment.

## Professional Development Program

To address the needs of teachers outlined in the classroom assessment research literature, the South Carolina Department of Education developed a professional development program in formative classroom assessment. The goal of the professional development was to help teachers develop better quality assessment practices to inform their instructional decisions and to provide better information to students about their learning. The professional development was structured as a one year, three-hour recertification course. Teachers who earned an A or B in the professional development earned recertification credit per state guidelines, and in addition, the state compensated the teachers for their time through a federally funded grant.

In Year 1, the professional development was implemented by an assessment coach in the treatment schools. No professional development was provided in the control schools. In Years 2 and 3, levels of treatment were investigated such that the professional development was implemented by an assessment coach (Treatment 1) or a relatively untrained facilitator (Treatment 2). Treatment 1 (coach) and Treatment 2 (untrained facilitator) groups used the same instructional materials. The content of the professional development comprised the following modules: (a) aligning assessments with the cognitive level and content of the curriculum standards; (b) developing and implementing performance tasks; (c) developing and implementing checklists; (d) developing and implementing rubrics; (e) formulating high-quality, multiple choice items; (f) analyzing the quality of multiple choice items to guide the determination regarding what students know; (g) developing portfolios; (h) using valid grading procedures; and (i) interpreting standardized test scores. Teachers were encouraged to use assessment results in both a formative and summative manner with their students, as needed.

A core component of the professional development was based upon the premise that teachers need to understand how the state interprets its own standards from a cognitive and content perspective. The professional development had three broad phases. In phase one, teachers viewed a video-presentation of material that focused on a specific aspect of classroom assessment, and they read a related chapter in a classroom assessment text. Each video presented the assessment principles and practices that teachers applied in phase two and phase three in the form of a performance task.

In phase two, the assessment coach (or untrained facilitator in Year 2 and Year 3) at each school used a companion document, along with the video series, to implement guided practice activities with teachers. Following the video presentation, in most modules, teachers analyzed and critiqued state-developed classroom assessment models that were provided in the companion document. Sample classroom assessment models contained both positive and negative attributes in terms of assessment construction guidelines and interpretations of standards identified by the state as being confusing to teachers. After teachers analyzed each model, the coach used written guidance for each model found in the companion document to guide the teachers in a discussion. The written guidance overviewed the interpretation of the relevant state standard as well as noted assessment construction issues that were posed to the teachers in the exercise. Next, the assessment coach divided the teachers into groups to collaboratively create a stipulated performance task for the module. In addition to applying measurement principles for each performance task, each group also analyzed and explained the cognitive level of the state standard upon which the performance task was based and presented their task to their peers as a whole. Using the rubric for the stipulated performance task, the assessment coach and teachers analyzed and discussed which measurement principles highlighted in the video were incorporated into the group-developed task. Thus, teachers had the intended outcomes modeled for them, they developed models to help them clarify and analyze the learning target, and they engaged in a culture of critique (Andrade, 2010).

In phase three, teachers created a parallel performance task individually as homework, using the same directions for the stipulated performance task used in the group work, for the module and the rubric for that performance task. In the coach group, teachers submitted their performance task to their assessment coach, who was trained by the state in grading the tasks, for a grade. Content-area specialists, who were also test-development specialists for the statewide large scale assessments, separately provided comments to teachers in the treatment group regarding the alignment of their task to the state standards that the teacher was measuring. In Year 2 and Year 3, when different levels of treatment were investigated, those teachers in the Treatment 2 group submitted their performance task only to the untrained facilitator for a grade. Teachers were then asked to give the performance task to students, score students, and develop reflections about what they learned.

Teachers participating in the professional development who were not satisfied with their own grade on a performance task could revise their work and resubmit for a higher grade. In

addition to their resubmission, teachers were also required to develop a reflection discussing their errors, how they resolved the errors, and what they felt they learned through the process. Thus, the state strived to model the formative assessment practice of self-assessment (Andrade & Boulay, 2003) with the expectation that teachers might bring this model into their own classroom. Teachers had multiple opportunities to compare their work with both their cohorts' work and state models while working to improve their own tasks. This enabled teachers to synthesize the state's learning goals, compare their work to the models shown in the video series and companion-document exercises, as well as understand, document, and take ownership of their own learning process through the use of reflections.

### *Investigation Context*

This study is the final component of a larger mixed methods approach that included investigations of treatment fidelity and teacher assessment skill. Fidelity of implementation evaluators (Yap, Whittaker, Liao, & D'Amico, 2006; Yap et al., 2007) found that on average coaches (and Year 2 and 3 facilitators) implemented 68%–73% of the activities that comprised the curriculum for the professional development outlined in the companion document. On average, teachers completed 47%–63% of the required content hours. Because the program was time consuming (requiring approximately 30 hours of professional development contact hours and 24 hours of homework), coaches and facilitators at times, eliminated some of the collaborative group work in phase two, and skipped to phase three, where teachers completed their performance tasks for homework. When this occurred, teachers did not have as many opportunities to clarify and analyze the learning target.

The abbreviated professional development sessions typically occurred because principals did not maintain the sanctity of the professional development time. Coaches and facilitators often were required to cancel, and then reschedule, the professional development sessions due to emergency school meetings or extra duties. Teachers who were prepared to meet one day a week after school for professional development lost that time. The rescheduled sessions required extra commitment from participants and their coaches or facilitators. This oftentimes resulted in the rescheduled sessions being shorter than the original intent of the program.

## Method

The state implemented the professional development program in state-identified, low performing middle schools from November 2005 through April 2008. Year 1 (2005–06), involved a multi-site, cluster randomized trial where schools were randomly assigned to either the treatment or control group. In Year 1, sixth-grade English language arts and mathematics teachers and their students were the study participants. In Year 2, seventh-grade English language arts and mathematics teachers and their students were the study participants, and in Year 3, eighth-grade English language arts and mathematics teachers and their students were the study participants.

In Year 2 (2006–07) and Year 3 (2007–08), the evaluation shifted from comparing a treatment group (professional development with assessment coaches) to a control group

(no professional development), to comparing two levels of treatment: professional development with a trained assessment coach verses professional development with an untrained facilitator. Schools in the Year 1 study that were randomly assigned to the professional development, maintained their trained assessment coach (now the Treatment 1 group). The previous control group schools became the Treatment 2 group, and received the professional development with a relatively untrained facilitator. The main difference between the two levels of treatment was the level of support provided to the assessment coaches and facilitators. Assessment coaches received assessment training and were trained in evaluating the teacher-generated assessments. The untrained facilitators did not receive assessment training and were not trained in evaluating the teacher-generated assessments. In the original program, there was no control group for Year 2 and Year 3. Therefore, we created a Year 2 and Year 3 control group by matching a set of schools to the schools participating in the professional development.

The variable used to match the Year 2 and Year 3 control group schools, to schools participating in the levels of treatment, was the South Carolina Department of Education's (SCDE) poverty index. SCDE creates the poverty index to match "like" schools for adequate yearly progress (AYP) comparisons. SCDE provides an analysis to answer the question "How do our AYP results compare to schools that are most like our school?"

The SCDE school poverty index combines information about schools, based upon free-and-reduced-price lunch data and Title I funding. This information is posted on the State's website: www.sc.ed.gov. We matched each Year 2 and Year 3 school participating in the professional development to a non-participating school from the same district that had the closest poverty index. If no such school existed within the district, the school closest on the school poverty index was selected from another district. Once a matched control group school was identified for each Year 2 and Year 3 school, a random sample of students from each matched school was drawn to equal the number of students in the Year 2 and Year 3 schools participating in the professional development. As a reminder, for Year 1, a control group was a component of the original study design.

Table 1 shows show the number of teachers participating in the project for each year. The sample of Year 1 teachers comprised 89% of those who initially agreed to participate in the study. The sample of Year 2 teachers comprised 75% of those who initially agreed to participate in the study, and the sample of Year 3 teachers comprised 100% of those who initially agreed to participate in the study. Each year's demographics for groups were similar as shown in Table 1.

Table 1
Teacher Demographics for Each Year

| Demographic | Year 1 | Year 2 | Year 3 |
|---|---|---|---|
| Female | 86% | 79% | 82% |
| Male | 14% | 21% | 18% |
| White | 58% | 51% | 54% |
| Minority | 42% | 49% | 46% |
| *N* | 151 | 146 | 71 |

Table 2
Blueprints of Pretest and Posttest for Each Year

| Year | Objective | Common Items | Pretest Items | | Posttest Items | |
|---|---|---|---|---|---|---|
| | | | Unique | Total | Unique | Total |
| 1 | A. Align items to curriculum standards | 7 (37%) | 2 | 9 (30%) | 3 | 10 (33%) |
| | B. Create high-level items | 3 (16%) | 2 | 5 (17%) | 4 | 7 (23%) |
| | C. Apply performance assessment development guidelines | 2 (11%) | 3 | 5 (17%) | 1 | 3 (10%) |
| | D. Understand issues in K-12 assessment | 2 (11%) | 1 | 3 (10%) | 1 | 3 (10%) |
| | E. Writing multiple choice items | 3 (16%) | 2 | 5 (17%) | 1 | 4 (13%) |
| | AC. Combination of objectives A and C | 1 (5%) | 1 | 2 (7%) | 0 | 1 (3%) |
| | BC. Combination of objectives B and C | 1 (5%) | 0 | 1 (3%) | 1 | 2 (7%) |
| | Total | 19 | 11 | 30 | 11 | 30 |
| Year | Objective | Common Items | Pretest Items | | Posttest Items | |
| | | | Unique | Total | Unique | Total |
| 2 | A. Align items to curriculum standards | 6 (32%) | 2 | 8 (26%) | 3 | 9 (32%) |
| | B. Create high-level items | 3 (16%) | 3 | 6 (19%) | 3 | 6 (21%) |
| | C. Apply performance assessment development guidelines | 3 (16%) | 2 | 5 (16%) | 0 | 3 (11%) |
| | D. Understand issues in K-12 assessment | 2 (11%) | 1 | 3 (10%) | 1 | 3 (11%) |
| | E. Writing multiple choice items | 3 (16%) | 3 | 6 (19%) | 1 | 4 (14%) |
| | AC. Combination of objectives A and C | 1 (5%) | 1 | 2 (6%) | 0 | 1 (4%) |
| | BC. Combination of objectives B and C | 1 (5%) | 0 | 1 (3%) | 1 | 2 (7%) |
| | Total | 19 | 12 | 31 | 9 | 28 |

Table 2 Continued
Blueprints of Pretest and Posttest for Each Year

| Year | Objective | Common | Pretest | | Posttest | |
|---|---|---|---|---|---|---|
| | | | Unique | Total | Unique | Total |
| 3 | A. Align items to curriculum standards | 5 (33%) | 2 | 7 (23%) | 4 | 9 (30%) |
| | B. Create high-level items | 2 (13%) | 6 | 8 (27%) | 5 | 7 (23%) |
| | C. Apply performance assessment development guidelines | 2 (13%) | 2 | 4 (13%) | 1 | 3 (10%) |
| | D. Understand issues in K-12 assessment | 2 (13%) | 2 | 4 (13%) | 1 | 3 (10%) |
| | E. Writing multiple choice items | 2 (13%) | 2 | 4 (13%) | 2 | 4 (13%) |
| | AC. Combination of objectives A and C | 1 (7%) | 1 | 2 (7%) | 1 | 2 (7%) |
| | BC. Combination of objectives B and C | 1 (7%) | 0 | 1 (3%) | 1 | 2 (7%) |
| | Total | 15 | 15 | 30 | 15 | 30 |

Each teacher participating in the study took a standardized assessment practices pretest and posttest. Although the teachers were given standardized directions for creating each performance task, the tasks that teachers submitted were in different subject areas (i.e., reading and mathematics) and on different topics. That is, the tasks were not standardized across teachers and content areas thus preventing their use as a single measure of teacher achievement. A multiple-choice test was used as an indirect measure of teacher assessment knowledge. The pretest and posttest were administered four months apart in Year 1, six months apart in Year 2, and seven months apart in Year 3. A split-plot ANOVA was used to compare group results from pretest to posttest each year. Two 40-item multiple choice teacher tests were developed to measure knowledge in developing various types of assessments with 60% of the items being held in common on both forms. The tests measured teacher's skill in concepts such as identifying the cognitive level of sample items, applying performance assessment and multiple-choice assessment writing guidelines, and aligning items to standards. The researchers intentionally focused on indirect measures of assessment development skill rather than broader assessment literacy skills. That is, the researchers designed tests to measure knowledge of assessment development and the ability to use (apply) that knowledge in reasoning contexts that are foundational to the development of quality classroom assessments. Table 2 shows the test blueprints for each pretest and posttest for each of the three years.

Items were reviewed for content by state testing experts from various perspectives. The state testing experts reviewed the items to ensure questions had only one correct answer, were based upon the state's interpretation of cognitive levels, that the alignment of items to the state standards was accurate, and that the

conventions for developing assessments followed those used in statewide assessments as covered in the professional development series. Through these reviews, content validity (Kane, 2006) evidence was collected to support the interpretation that items measured the state's conceptualization of its standards and the best practices the state used in assessment construction.

Classical item analysis was used to determine the quality of the items. As part of the test development process, items with poor statistical fit or distractors with positive point biserial correlations were avoided because such data may indicate that an item is tapping an ability that is not related to the construct being measured. The tests ranged from 28–31 questions in length. The final test forms proportionally measured about the same content across years (see Table 2). The KR20 for the pretest form was generally lower than for the posttest, with the exception of Year 3. This is shown in Table 3.

Table 3
KR-20 Reliability and Stout's *T* for Pretest and Posttest for Each Year

| | Administration | KR20 | Stout's *T* |
|---|---|---|---|
| 1 | Pretest | 0.63 | -0.25 ($p$ = 0.91) |
| | Posttest | 0.78 | -0.88 ($p$ = 0.60) |
| 2 | Pretest | 0.64 | -0.22 ($p$ = 0.59) |
| | Posttest | 0.75 | -0.75 ($p$ = 0.77) |
| 3 | Pretest | 0.71 | — |
| | Posttest | 0.67 | — |

The smaller reliability coefficient for each pretest was expected, given that teachers were more homogenous in terms of true score variance, prior to the professional development program. They were generally equally unfamiliar with the concepts upon which the items were based, and therefore, tended to have similar scores at the pretest. After participation in the professional development program, scores were more variable because the treatment increased teachers' classroom assessment knowledge to different levels creating more variability within the posttest scores.

To determine if construct irrelevant variance was influencing test scores, differential item functioning (DIF) analyses were conducted. DIF occurs when individuals with the same ability on the measured construct have different probabilities of answering an item correctly based upon their subgroup affiliation. Once an item is flagged for a significant DIF, professional judgment is used to determine whether there are aspects of content or an item's format that might bias test scores for a particular subgroup.

Items were placed into one of the Educational Testing Service's DIF categories reflecting the severity of DIF (see Huynh, Meyer, & Barton, 2000, p. 53). No items were classified with significant DIF. Dimensionality was tested using DIMTEST (Stout, 1987) for each pretest and posttest in Year 1 and Year 2. Dimensionality investigations are used to collect construct validity evidence, and unidimensionality is one assumption for the use of item response theory (IRT). As shown in Table 3, Stout's *T* indicated that the pretest and posttest forms were unidimensional. Because of the small Year 3 sample size, Stout's *T* was not calculated, however, we note that the majority of items found on the Year 3 test form were common to Year 1 and Year 2.

The tests were calibrated using IRT, specifically with the Rasch model, using WINSTEPS (Linacre, 2003). Kane (2006) noted IRT models warrant inferences about a person's ability level based upon their performance on a sample of items. All INFIT and OUTFIT mean squares were within the acceptable range of .7 to 1.3 (Linacre, 2003). Common item equating was used to place the two forms on the same scale using items common to both forms. The scale was established by the posttest, and the pretest was equated to the posttest. Ability estimates were transformed to a scale ranging from 100 to 200.

A multilevel analysis, by content area, was conducted for each year of the study to determine whether the professional development intervention improved student achievement. The dependent variables were student English language arts and mathematics scale scores on the Palmetto Achievement Challenge Test (PACT), the state's high-stakes accountability testing program at the time. A two-level model (student and school) was run.

The multilevel model of PACT scores involved multiple grand-mean centered covariates at level-1: Individualized Education Program (IEP), gender (FEMALE), minority status (MINORITY), free-and-reduced price lunch status (FRPLUNCH). In addition to these demographic covariates, a student PACT score from the prior year was also included at level-1. The level-1 model was

$$Y_{ij} = \beta_{0j} + \beta_{1j}(PREPACT) + \beta_{1j}(FEMALE) + \beta_{1j}(MINORITY) + \beta_{1j}(IEP) + \beta_{1j}(FRPLUNCH) + r_{iy}$$

where PREPACT represents the previous year's PACT score. With these covariates, the intercept, $\beta_{0j}$, represents PACT scores adjusted for the covariates. The coefficients for each covariate were fixed, but the coefficient for the intercept was allowed to randomly vary. This allowed professional development treatment status to be included at level-2 for the intercept. The level-2 model was,

$$\beta_{0j} = \gamma_{00} + \gamma_{01}(GROUP) + \mu_{oj}$$

where GROUP was dummy coded as 1 for the professional development schools (termed treatment in the tables hereafter), and 0 for the matched control group schools. GROUP represents the effect of the treatment on adjusted PACT scores. For the Year 2 and Year 3 analysis, facilitator and coach group schools were collapsed into one group for two reasons. First, data supported that there were not differences in teacher assessment knowledge as a function of levels of treatment. Second, previous studies (Ragland, Schneider, Yap, & Kaliski, 2008; Schneider, Meyer, Miller, & Kaliski, 2007) found no differences in investigations regarding differences in student achievement by level of treatment.

## Teacher Achievement Results

Descriptive statistics indicated that in Year 1 the treatment group improved their test scores by 8.51 points from pretest to posttest, whereas, the control group improved their scores by 1.85 points as shown in Table 4.

Table 4
Pretest and Posttest Descriptive Statistics

| Year | Test | Statistic | Control | Treatment |
|---|---|---|---|---|
| 1 | Pretest | *M* | 153.63 | 153.78 |
| | | *SD* | 6.84 | 7.19 |
| | Posttest | *M* | 155.48 | 162.29 |
| | | *SD* | 8.25 | 11.47 |
| Year | Test | Statistic | Treatment 2 | Treatment 1 |
| 2 | Pretest | *M* | 151.01 | 150.63 |
| | | *SD* | 5.79 | 7.45 |
| | Posttest | *M* | 162.00 | 162.40 |
| | | *SD* | 8.68 | 11.60 |
| Year | Test | Statistic | Treatment 2 | Treatment 1 |
| 3 | Pretest | *M* | 154.67 | 154.11 |
| | | *SD* | 7.10 | 9.18 |
| | Posttest | *M* | 162.90 | 162.11 |
| | | *SD* | 6.37 | 8.09 |

The inferential analysis was conducted with a split-plot ANOVA, and an alpha level of .05 was used. The results indicated that an interaction occurred between the factors time and group. The treatment group differed significantly from the control group across time, $F(1, 149) = 19.92$, $p < .001$. The effect size estimate ($\omega^2$) was 0.11 as shown in Table 5. In Year 2, descriptive statistics indicated that the Treatment 2 (facilitator) group improved their test scores by 10.99 points from pretest to posttest, and the Treatment 1 (coach) group improved their scores by 11.77 points as shown in Table 4. The inferential analysis was conducted with a split-plot ANOVA, and the results indicated a statistically significant increase from pretest to posttest for both groups across time [$F(1,138) = 211.50$, $p < .001$]. The effect size estimate ($\omega^2$) was 0.60 as shown in Table 6. There was no significance difference in the amount of gain between the two treatment groups. In Year 3, descriptive statistics indicated that the Treatment 2 (facilitator) group improved their test scores by 8.23 points from pretest to posttest, and the Treatment 1 (coach) group improved their scores by 8.00 points as shown in Table 4. The inferential analysis was conducted with a split-plot ANOVA, and the results indicated a statistically significant increase from pretest to posttest for both groups across time [$F(1,65) = 89.76$, $p < .001$] as shown in Table 7. The effect size estimate ($\omega^2$) was 0.56. There was no significance difference in the amount of gain between the two treatment groups. Tables 5-7 show the full split-plot ANOVA tables for each year along with the effect size for each factor and interaction.

Table 5:
ANOVA Table for Split-Plot Design, Year 1

| Effect | *SS* | *df* | *MS* | *F* | *p*-value | $\omega^2$ |
|---|---|---|---|---|---|---|
| Group | 862.90 | 1 | 862.90 | 8.53 | <0.001 | 0.05 |
| Time | 1919.49 | 1 | 1919.49 | 48.16 | <0.001 | 0.24 |
| Time x Group | 794.08 | 1 | 794.08 | 19.92 | <0.001 | 0.11 |
| Within Error | 5939.19 | 149 | 39.86 | | | |
| Between Error | 15081.71 | 149 | 101.22 | | | |

Table 6
ANOVA Table for Split-Plot Design, Year 2

| Effect | *SS* | *df* | *MS* | *F* | *p*-value | $\omega^2$ |
|---|---|---|---|---|---|---|
| Group | 0.00 | 1 | 0.00 | <0.001 | 1.00 | 0.00 |
| Time | 9042.46 | 1 | 9042.46 | 211.50 | <0.001 | 0.60 |
| Time x Group | 10.68 | 1 | 10.68 | 0.25 | 0.62 | 0.00 |
| Within Error | 5900.01 | 138 | 42.75 | | | |
| Between Error | 14995.47 | 138 | 108.66 | | | |

Table 7
ANOVA Table for Split-Plot Design, Year 3

| Effect | *SS* | *df* | *MS* | *F* | *p*-value | $\omega^2$ |
|---|---|---|---|---|---|---|
| Group | 15.11 | 1 | 15.11 | 0.15 | 0.70 | 0.00 |
| Time | 2182.90 | 1 | 2182.90 | 89.76 | <0.001 | 0.56 |
| Time x Group | 0.45 | 1 | 0.45 | 0.02 | 0.89 | 0.00 |
| Within Error | 1580.68 | 65 | 24.32 | | | |
| Between Error | 6443.82 | 65 | 99.14 | | | |

## Student Achievement Results

Demographic and descriptive statistics are provided separately for each subject area (English language arts and mathematics). This is necessary because the professional development was at the teacher level, and the teacher was either an English language arts (ELA) teacher or a mathematics teacher. School and student sample sizes for each year are presented in Table 8, and this table shows Year 3 saw attrition in school participation.

Tables 9 and 10 show that most students in both school sets (treatment verses control) lived in poverty. Student demographics were similar for treatment and control group schools in Year 1. As seen in Tables 9 and 10, student demographics were similar for the two sets of schools in Years 2 and 3, with the exception of the percentage of minority students and percentage of students receiving free-or-reduced-price lunch. Although both sets of schools had a large portion of minority students and free-or-

reduced-price lunch students, these percentages were larger in the professional development schools, than in the control group schools. Because level-one variables account for student-level characteristics, the demographic differences between school sets do not present a threat to the validity of the findings. Tables 9 and 10 show male and female students and students who had an individual educational program (IEP) were similarly represented in both school sets.

Table 8
Sample Size for Each Year and Subject Area

| Replication | Level | ELA | Mathematics |
|---|---|---|---|
| Year 1 | Students | 2,066 | 2,708 |
| | Schools | 38 | 44 |
| Year 2 | Students | 3,203 | 2,745 |
| | Schools | 44 | 36 |
| Year 3 | Students | 1,403 | 1,297 |
| | Schools | 20 | 14 |

Intraclass correlations (ICCs) were 0.17 for English language arts and mathematics in Year 1 indicating that about 17% of the variance in student scale scores is attributable to schools. This correlation decreased in Years 2 and 3. English language arts and mathematics ICCs were 0.09 and 0.14 in Year 2, respectively. In Year 3, these values were 0.04 and 0.09 for English language arts and mathematics, respectively.

Control group schools typically had higher average (see Table 11) beginning English language arts scores than the professional development schools. Given the differences in the percentages of students receiving free-or-reduced-price lunch, this initial difference was not unexpected. However, as shown in Tables 12, 13, and 14 once student-level characteristics were accounted for, the treatment effect was found to be a predictor of adjusted English language arts PACT scores each of the three years. For English language arts, adjusted PACT scores were 2.39, 1.15, and 3.46 points lower (by year) for professional development schools than the control group schools. Cohen's *d* like standardized effect sizes (see McCoach, 2010) for these three years of ELA were -0.16, -0.08, and -0.26, respectively. These differences were statistically significant at the 0.05 level. A significant amount of level-1 variation still existed for the Year 1 and Year 2 data suggesting that additional factors were also influencing the variation of student adjusted PACT scores; however, the model explained the variation in test scores for Year 3.

Control group schools had higher average (see Table 15) beginning mathematics scores than the professional development schools. As noted previously, this initial difference was not unexpected. However, as shown in Table 18 once student-level characteristics were accounted for in Year 3, the treatment effect was found to be a predictor of adjusted Mathematics PACT scores. The adjusted PACT scores were 3.76 points lower for professional development schools than the control group schools. This difference was statistically significant at the 0.05 level (no significant differences were found in Year 1 and Year 2). The standardized effect size for Year 3 was -0.33, but it was only 0.03 for Year 1 and -0.01 for Year 2. As Tables 16, 17, and 18 show for each year a significant amount of level-1 variation still existed in the data suggesting that additional factors were also influencing the variation of student adjusted PACT scores.

Table 9
Student Demographics for English Language Arts

| Replication | Characteristic | Treatment | Control |
|---|---|---|---|
| Year 1 | Female | 43.90 | 45.60 |
| | Minority | 60.10 | 58.10 |
| | IEP | 19.40 | 20.20 |
| | FRP Lunch | 78.50 | 72.50 |
| Replication | Characteristic | Treatment | Control |
| Year 2 | Female | 52.50 | 51.10 |
| | Minority | 77.80 | 64.40 |
| | IEP | 8.90 | 10.90 |
| | FRP Lunch | 77.30 | 68.20 |
| Year 3 | Female | 50.50 | 45.20 |
| | Minority | 66.70 | 53.50 |
| | IEP | 10.90 | 13.90 |
| | FRP Lunch | 72.40 | 62.90 |

Table 10
Student Demographics for Mathematics

| Replication | Characteristic | Treatment | Control |
|---|---|---|---|
| Year 1 | Female | 42.90 | 39.30 |
| | Minority | 63.50 | 48.10 |
| | IEP | 21.60 | 30.90 |
| | FRP Lunch | 79.40 | 75.20 |
| Replication | Characteristic | Treatment | Control |
| Year 2 | Female | 49.00 | 50.50 |
| | Minority | 66.30 | 52.60 |
| | IEP | 8.70 | 11.60 |
| | FRP Lunch | 71.70 | 63.60 |
| Year 3 | Female | 51.00 | 45.50 |
| | Minority | 66.70 | 48.20 |
| | IEP | 7.90 | 12.50 |
| | FRP Lunch | 74.10 | 55.90 |

Table 11
Student-level Descriptive Statistics for the PACT English Language Arts

| Replication | Year | Statistic | Treatment | Control |
|---|---|---|---|---|
| Year 1 | 2005 | *M* | 498.99 | 498.73 |
| | | *SD* | 12.30 | 11.88 |
| | 2006 | *M* | 597.20 | 598.62 |
| | | *SD* | 14.35 | 15.01 |
| Replication | Year | Statistic | Treatment | Control |
| Year 2 | 2006 | *M* | 596.96 | 600.02 |
| | | *SD* | 14.42 | 15.48 |
| | 2007 | *M* | 696.34 | 699.65 |
| | | *SD* | 12.86 | 17.08 |
| Year 3 | 2007 | *M* | 699.06 | 700.08 |
| | | *SD* | 23.39 | 18.83 |
| | 2008 | *M* | 796.62 | 800.68 |
| | | *SD* | 11.73 | 14.05 |

Table 12
Fixed and Random Effects for 2005–06 English Language Arts

| Intercept | | | | | |
|---|---|---|---|---|---|
| Fixed Effect | Coefficient | *SE* | *t* | *df* | *p*-value |
| Intercept | 599.00 | 0.64 | 940.12 | 36.00 | 0.00 |
| Treatment | -2.39 | 1.00 | -2.39 | 36.00 | 0.02 |
| 2005 PACT | 0.90 | 0.02 | 44.32 | 2059.00 | 0.00 |
| IEP | -0.65 | 0.75 | -0.87 | 2059.00 | 0.39 |
| Female | 2.29 | 0.43 | 5.38 | 2059.00 | 0.00 |
| Minority | -1.66 | 0.51 | -3.23 | 2059.00 | 0.00 |
| FRP Lunch | -1.88 | 0.50 | -3.75 | 2059.00 | 0.00 |
| Variance Components | | | | | |
| Random Effect | Variance | $X^2$ | *df* | *p*-value | |
| Intercept | 7.27 | 222.26 | 36.00 | 0.00 | |
| Level-1 | 78.17 | | | | |

Table 13
Fixed and Random Effects for 2006–07 English Language Arts

| Intercept | | | | | |
|---|---|---|---|---|---|
| Fixed Effect | Coefficient | *SE* | *t* | *df* | *p*-value |
| Intercept | 698.63 | 0.38 | 1861.95 | 42.00 | 0.00 |
| Treatment | -1.15 | 0.54 | -2.14 | 42.00 | 0.04 |
| 2006 PACT | 0.66 | 0.01 | 45.98 | 3196.00 | 0.00 |
| IEP | -3.95 | 0.76 | -5.19 | 3196.00 | 0.00 |
| Female | 2.23 | 0.23 | 9.81 | 3196.00 | 0.00 |
| Minority | -1.34 | 0.43 | -3.13 | 3196.00 | 0.00 |
| FRP Lunch | -1.98 | 0.32 | -6.22 | 3196.00 | 0.00 |
| Variance Components | | | | | |
| Random Effect | Variance | $X^2$ | *df* | *p*-value | |
| Intercept | 2.42 | 169.45 | 42.00 | 0.00 | |
| Level-1 | 58.16 | | | | |

Table 14
Fixed and Random Effects for 2007–08 English Language Arts

| Intercept | | | | | |
|---|---|---|---|---|---|
| Fixed Effect | Coefficient | *SE* | *t* | *df* | *p*-value |
| Intercept | 800.41 | 0.45 | 1768.19 | 18.00 | 0.00 |
| Treatment | -3.46 | 0.66 | -5.21 | 18.00 | 0.00 |
| 2007 PACT | 0.24 | 0.01 | 17.64 | 1396.00 | 0.00 |
| IEP | -10.69 | 0.85 | -12.61 | 1396.00 | 0.00 |
| Female | 3.59 | 0.54 | 6.67 | 1396.00 | 0.00 |
| Minority | -3.61 | 0.59 | -6.10 | 1396.00 | 0.00 |
| FRP Lunch | -3.84 | 0.62 | -6.24 | 1396.00 | 0.00 |
| Variance Components | | | | | |
| Random Effect | Variance | $X^2$ | *df* | *p*-value | |
| Intercept | 0.54 | 26.07 | 18.00 | 0.10 | |
| Level-1 | 96.95 | | | | |

Table 15
Student-level Descriptive Statistics for the PACT Mathematics

| Replication | Year | Statistic | Treatment | Control |
|---|---|---|---|---|
| Year 1 | 2005 | *M* | 504.49 | 507.23 |
| | | *SD* | 13.11 | 12.43 |
| | 2006 | *M* | 605.89 | 607.45 |
| | | *SD* | 13.03 | 12.61 |
| Replication | Year | Statistic | Treatment | Control |
| Year 2 | 2006 | *M* | 608.22 | 608.36 |
| | | *SD* | 14.25 | 15.37 |
| | 2007 | *M* | 706.48 | 707.29 |
| | | *SD* | 13.44 | 14.52 |
| Year 3 | 2007 | *M* | 706.47 | 710.05 |
| | | *SD* | 20.92 | 19.05 |
| | 2008 | *M* | 800.53 | 805.53 |
| | | *SD* | 9.42 | 12.30 |

Table 16
Fixed and Random Effects for 2005–06 Mathematics

| Intercept | | | | | |
|---|---|---|---|---|---|
| Fixed Effect | Coefficient | *SE* | *t* | *df* | *p*-value |
| Intercept | 606.66 | 0.45 | 1339.91 | 42.00 | 0.00 |
| Treatment | 0.44 | 0.72 | 0.61 | 42.00 | 0.55 |
| 2005 PACT | 0.72 | 0.01 | 56.25 | 2701.00 | 0.00 |
| IEP | -3.85 | 0.52 | -7.44 | 2701.00 | 0.00 |
| Female | 1.49 | 0.33 | 4.49 | 2701.00 | 0.00 |
| Minority | -1.86 | 0.41 | -4.58 | 2701.00 | 0.00 |
| FRP Lunch | -1.60 | 0.39 | -4.15 | 2701.00 | 0.00 |
| Variance Components | | | | | |
| Random Effect | Variance | $X^2$ | *df* | *p*-value | |
| Intercept | 4.23 | 208.88 | 42.00 | 0.00 | |
| Level-1 | 58.43 | | | | |

Table 17
Fixed and Random Effects for 2006–07 Mathematics

| Intercept | | | | | |
|---|---|---|---|---|---|
| Fixed Effect | Coefficient | *SE* | *t* | *df* | *p*-value |
| Intercept | 706.98 | 0.56 | 1271.14 | 34.00 | 0.00 |
| Treatment | -0.16 | 0.78 | -0.21 | 34.00 | 0.84 |
| 2006 PACT | 0.66 | 0.01 | 52.15 | 2738.00 | 0.00 |
| IEP | -2.66 | 0.57 | -4.65 | 2738.00 | 0.00 |
| Female | -1.00 | 0.33 | -3.01 | 2738.00 | 0.00 |
| Minority | -2.20 | 0.40 | -5.50 | 2738.00 | 0.00 |
| FRP Lunch | -2.20 | 0.40 | -5.55 | 2738.00 | 0.00 |
| Variance Components | | | | | |
| Random Effect | Variance | $X^2$ | *df* | *p*-value | |
| Intercept | 4.24 | 162.41 | 34.00 | 0.00 | |
| Level-1 | 73.04 | | | | |

Table 18
Fixed and Random Effects for 2007–08 Mathematics

| Intercept | | | | | |
|---|---|---|---|---|---|
| Fixed Effect | Coefficient | *SE* | *t* | *df* | *p*-value |
| Intercept | 805.06 | 0.87 | 925.16 | 12.00 | 0.00 |
| Treatment | -3.76 | 1.23 | -3.06 | 12.00 | 0.01 |
| 2007 PACT | 0.23 | 0.01 | 16.96 | 1290.00 | 0.00 |
| IEP | -6.38 | 0.87 | -7.34 | 1290.00 | 0.00 |
| Female | -0.63 | 0.51 | -1.25 | 1290.00 | 0.21 |
| Minority | -1.75 | 0.59 | -2.94 | 1290.00 | 0.00 |
| FRP Lunch | -2.64 | 0.60 | -4.41 | 1290.00 | 0.00 |
| Variance Components | | | | | |
| Random Effect | Variance | $X^2$ | *df* | *p*-value | |
| Intercept | 4.18 | 64.04 | 12.00 | 0.00 | |
| Level-1 | 81.27 | | | | |

## Discussion

The findings of this three-year study indicate that the professional development curriculum improved teacher knowledge of formative classroom assessment skills regardless of whether an assessment coach or a relatively untrained facilitator implemented the professional development curriculum. The finding that an increase in teacher achievement in assessment knowledge results in a

decrease in student achievement was unexpected. Although strong causal conclusions may not be warranted from a quasi-experimental design, we sought to understand why the decrease occurred in this study.

Based upon fidelity of implementation reports, teacher interviews, and survey information (Yap, Whittaker, Liao, & D'Amico, 2006; Yap et al., 2007) teachers who participated in the professional development program moved away from primarily assessing students with multiple choice items to using a more diverse set of assessment practices. Teachers described aligning their instruction and assessments to the standards, sharing their grading criteria, and using more ethical classroom assessment grading practices.

In interviews one to two years after the professional development, teachers reported being more proficient and selective in the types of tools they used to gather information about students (D'Amico, Hardee, Morgan, Yap, & Ishikawa, 2008). In portfolios developed as the culmination of the professional development, teachers described moving toward collaborative assessment techniques (e.g., building rubrics with students). One teacher realized after reviewing a distractor analyses on a multiple-choice test she developed that superficial test taking strategies she had taught her students had been transferred to their approach to reading texts in general. She wrote. "I am… choosing to analyze my teaching strategies and methods because my students did poorly…When students had to hunt for the information, no matter what the standard, there were breaks in their answers. I noticed that they looked for information in the beginning of the text and at the end, but not in the middle." She later concluded she had mistaught students and now needed to go back and reteach. Although evaluators collected evidence that suggested a change in teacher abilities and practices occurred in regard to classroom assessment, information related to how teachers used the information to inform instructional decisions was not collected.

Heritage, Kim, Vendlinski, and Herman (2009) wrote "A review of recent literature suggests there is little or no extant research that has investigated teachers' abilities to adapt instruction based on assessment of student knowledge and understanding (p. 24)." In the Heritage et al. study, teachers were asked to review student responses to assessment tasks in three areas. They found interesting results. Teachers' abilities to (a) identify what mathematics principal was being measured, (b) infer what the student knew and could do, and (c) determine next instructional steps based upon the student's response, differed by assessment task. That is, some tasks presented to students were easier for teachers to analyze than others. Although Heritage et al. concluded that determining instructional next steps based upon student data was the most difficult task for teachers, we derive a slightly different but equally disturbing conclusion. By transforming the average score for each teacher skill measured (based upon data presented in the article) to a percent of the possible rubric points obtained, we concluded that teachers had roughly equal difficulty with each task. If teachers are not sure what is being measured or how to break apart student responses to determine where a student misconception occurs, they surely are unable to determine what to do next. Moreover, these findings also call into question the quality of feedback that teachers are able to provide students in such situations.

The findings of Heritage et al. (2009) help us conjecture what occurred in our study. Perhaps, teachers repeatedly struggled with breaking apart student responses to identify where the student was in the learning progression for a learning target. If true, although the student was retaught his or her specific needs may not have been addressed. Second, if the teacher could identify the student misconception, he or she may not have *changed* the instructional approach. That is, knowing that reteaching was necessary, but not knowing what to do differently, he or she simply repeated the previous instructional approach. Perhaps, reteaching does not improve student achievement of the intended learning target when it is not changed or directed specifically to where the student is in the learning progression. Moreover, reteaching may then detract from what the student would have learned if the teacher would have just moved to the next area of instruction. That is, the breadth of the content coverage across the year may have been unintentionally restricted with no additional value added from the time spent reteaching. As noted by Heritage et al. (2009) if the teacher is not able to move learning forward, the value of formative assessment is called into question.

Schneider and Randel (2010) contended that professional development in formative classroom assessment does nothing to directly increase knowledge of how students learn in the content area. Strong content knowledge and understanding of learning progressions are likely precursor skills to using assessment information accurately. Professional development providers in formative classroom assessment practices may need to model how to analyze what student responses can tell us about what students know, and in addition, they may need to model what to do next. On the other hand, Andrade (2010) posited that students should be the main producer and user of formative classroom assessment information. She noted that teachers are responsible for creating a culture of critique, providing students multiple exemplars of the intended learning target, having students compare their work to the intended learning target, and expecting that students should revise and improve their work. Given this framework, the teacher becomes a facilitator of learning, and the student owns his or her own learning growth.

There is much to learn about formative classroom assessment practices and how they can be used to change student learning. In this study we found that implementation of a professional development program in formative classroom assessment practices can lead to diminished student learning. To our knowledge, this is the first study with such a finding. Based upon our findings we agree with Nichols, Meyers, and Burling (2009). For assessment practices to be considered formative they must "causally link information from performance on a particular assessment to the selection of instructional actions whose implementation leads to gains in student learning (p.15)." It may be that we only know if assessment practices are truly formative in hindsight.

## References

Andrade, H., & Boulay, B. (2003). The role of rubric-referenced self-assessment in learning to write. *Journal of Educational Research, 97*(1), 21–34.

Andrade, H. L., Du, Y., & Wang, X. (2008). Putting rubrics to the test: The effect of model, criteria generation, and rubric-referenced self-assessment on elementary school students' writing. *Educational Measurement: Issues and Practice, 27*(2), 3–13.

Andrade, H. (2010). Students as the definitive source of formative assessment: Academic self-assessment and the self-regulation of learning. In H. Andrade & G. Cizek (Eds.), *Handbook of formative assessment.* (pp. 90–105). New York: Routledge.

Aschbacher, P. R. (1999). *Developing indicators of classroom practice to monitor and support school reform. University of California, Los Angeles: CRESST Technical Report 513. University of California, Los Angeles: Center for Research on Educational Standards and Student Testing.* Retrieved January 30, 2009, from http://research.cse.ucla.edu/Reports/TECH513.pdf.

Barton, P., & Coley, R. (1994). *Testing in America's schools.* Princeton, NJ: Educational Testing Services.

Bell, C., Steinberg, J., Wiliam, D., & Wylie, C. (2008, March). *Formative assessment and teacher achievement: Two years of implementation of the Keeping Learning on Track Program.* Paper presented at the annual meeting of the National Council on Measurement in Education, New York.

Brookhart, S. M. (2005, April). *Research on formative classroom assessment.* Paper presented at the annual meeting of the American Educational Research Association, Montreal, Canada.

Brookhart, S. M., Moss, C. M., & Long, B. A. (2007). *A cross-case analysis of teacher inquiry into formative assessment practices in six Title I reading classrooms. CASTL Technical Report Series No. 1-07.* Retrieved August 18, 2008, from http://www.castl.duq.edu/Castl_Tech Reports.htm

Brookhart, S. M., Moss, C. M., & Long, B. A. (2008, March). *Professional development in formative assessment: Effects on teacher and student learning.* Paper presented at the annual meeting of the National Council on Measurement in Education, New York.

Brookhart, S.M. (2009). Editorial. *Educational Measurement: Issues and Practice, (28)*3, 1–3.

Brookhart, S. M. (2010). Mixing it up: Combining sources of classroom achievement information for formative and summative purposes. In H. Andrade & G. Cizek (Eds.), *Handbook of formative assessment* (pp. 279–296). New York: Routledge.

Carter, K. (1984). Do teachers understand the principles for writing tests? *Journal of Teacher Education, 35,* 57–60.

D'Amico, L., Hardee, S., Morgan, G., Yap, C. C., & Ishikawa, T. (2008). *An evaluation of teacher quality research: Investment, implementation, and long-term student achievement effects.* Columbia, SC: University of South Carolina, Office of Program Evaluation.

Fleming, M., & Chambers, B. (1983). Teacher-made tests: Windows on the classroom. In W. E. Hathaway, (Ed.), *Testing in the schools. New directions for testing and measurement* (pp. 29–38). San Francisco: Jossey-Bass.

Haydel, J. B., Oescher, J., & Banbury, M. (1995, April). *Assessing classroom teachers' performance assessments.* Paper presented at the annual meeting

of the American Educational Research Association, San Francisco.

Heritage, M., Kim, J., Vendlinski, T., & Herman, J. (2009). From evidence to action: A seamless process in formative assessment? *Educational Measurement: Issues and Practice, (28)*3, 24–31.

Huynh, H., Meyer, J. P., & Barton, K. (2000). *Technical documentation for the 1999 Palmetto Achievement Challenge Tests of English Language Arts and Mathematics, grades 3 – 8.* Columbia, South Carolina: University of South Carolina, South Carolina Department of Education.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement: Fourth edition.* (pp. 17–64). Westport, CT: Praeger Publishers.

Linacre, J. M. (2003). *A user's guide to WINSTEPS MINISTEP Rasch-model computer program*. Author.

Llosa, L. (2005). Assessing English learners' language proficiency: A Qualitative Investigation of Teachers' Interpretations of the California ELD standards. *The CATSOEL Journal, 17*(1), 7–18.

Marso, R. N., & Pigge, F. L. (1993). Teachers' testing knowledge, skills, and practices. In S. L. Wise (Ed.), *Teacher training in measurement and assessment skills* (pp. 129-185). Lincoln, NE: Buros Institute of Mental Measurements, University of Nebraska-Lincoln.

McCoach, D. B. (2010). Hierarchical linear modeling. In G. R. Hancock & R. O. Mueller (Eds.), *The reviewer's guide to quantitative methods in the social sciences* (pp.123-140). New York: Routledge.

Meisels, S., Atkins-Burnett, S., Xue, Y., Nicholson, J., Bickel, D. D., & Son, S. (2003). Creating a system of accountability: The impact of instructional assessment on elementary children's achievement scores. *Educational Policy Analysis Archives, 11*(9), Retrieved March 23, 2007, from http://epaa.asu.edu/epaa/v11n9/

Nasstrom, G. (2009). Interpretations of standards with Bloom's revised taxonomy: A comparison of teachers and assessment experts. International *Journal of Research & Method in Education, 32*(1), 39–51.

Newmann, F., Bryk, A. S., & Nagaoka, J. K. (2001). *Authentic intellectual work and standardized tests: Conflict or coexistence?* Retrieved June 27, 2007, from http://ccsr.uchicago.edu/publications/p0a02.pdf

Nichols, P. D., Meyers, J. L., & Burling, K. S. (2009). A framework for evaluating and planning assessments intended to improve student achievement. *Educational Measurement: Issues and Practice (28)*3, 14–33.

Oescher, J., & Kirby, P. C. (1990). *Assessing teacher-made tests in secondary math and science classrooms.* Paper presented at the annual meeting of the National Council on Measurement in Education, Boston. (ERIC Document Number ED 322 169).

Plake, B. S., & Impara, J. C. (1997). Teacher assessment literacy: What do teachers know about assessment? In G. D. Phye (Ed.), *Handbook of classroom assessment, learning, adjustment, and achievement.* San Diego, CA: Academic Press.

Ragland, S., Schneider, M. C., Yap, C. C., & Kaliski, P. K. (2008, April). *The effect of classroom assessment professional development on English*

*language arts and mathematics student achievement: Year 2 results.* Paper presented at the 2008 meeting of the National Council on Measurement in Education, New York, NY.

Ross, J. A., Hogaboam-Gray, A., & Rolheiser, C. (2002). Student self-evaluation in grade 5–6 mathematics effects on problem solving achievement. *Educational Assessment, 8*(1), 43–59.

Schneider, M. C., Meyer, J. P., Miller, B. J., & Kaliski, P. K. (2007, April). *The effect of classroom assessment professional development on English language arts and mathematics achievement.* Paper presented at the 2007 meeting of the National Council on Measurement in Education, Chicago, IL.

Schneider, M. C., & Randel, B. (2010). Research on characteristics of effective professional development programs for enhancing educators' skills in formative assessment. In H. Andrade & G. Cizek (Eds.), *Handbook of formative assessment* (pp. 251–276). New York: Routledge.

Shepard, L. A. (2009). Commentary: Evaluating the validity of formative and interim assessment. *Educational Measurement: Issues and Practice (28)*3, 32–37.

Sobolewski, K. B. (2002). *Gender equity in classroom questioning.* Unpublished doctoral dissertation, South Carolina State University, Orangeburg.

Stout, W. F. (1987). A nonparametric approach for assessing latent trait dimensionality. *Psychometrika, 52,* 589-617.

Webb, N. M., Herman, J. L., & Webb, N. L. (2007). Alignment of mathematics' state-level standards and assessments: The role of reviewer agreement. *Educational Measurement: Issues and Practice, 26,* 17–29.

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

Wolfe, E. W., Viger, S. G., Jarvinen, D. E., & Linksman, J. (2007). Validation of scores from a measure of teachers' efficacy toward standards-aligned classroom assessment. *Educational and Psychological Measurement, 67*(3), 460–474.

Yap, C. C., Whittaker, L., Liao, C., & D'Amico, L. (2006). *Evaluation of a professional development program in classroom assessment: 2005–06.* Columbia, SC: University of South Carolina, Office of Program Evaluation.

Yap, C. C., Pearsall, T., Morgan, G., Wu, M., Maganda, F., Gilmore, J., Lewis, A., Halladay, K., & D'Amico, L. (2007). *Evaluation of a professional development program in classroom assessment: 2006–07.* Columbia, SC: University of South Carolina, Office of Program Evaluation.