Educational Measurement
ISSUES AND PRACTICE

NCME national council on measurement in education

Educational Measurement: Issues and Practice
Summer 2013, Vol. 32, No. 2, pp. 26–36

Grading as a Reform Effort: Do Standards-Based Grades Converge With Test Scores?

Megan E. Welsh, *University of Connecticut*, Jerome V. D'Agostino, *The Ohio State University*, and Burcu Kaniskan, *Pearson*

Standards-based progress reports (SBPRs) require teachers to grade students using the performance levels reported by state tests and are an increasingly popular report card format. They may help to increase teacher familiarity with state standards, encourage teachers to exclude nonacademic factors from grades, and/or improve communication with parents. The current study examines the SBPR grade–state test score correspondence observed across 2 years in 125 third and fifth grade classrooms located in one school district to examine the degree of consistency between grades and state test results. It also examines the grading practices of a subset of 37 teachers to determine whether there is an association between teacher appraisal style and convergence rates. A moderate degree of grade–test score convergence was observed using three agreement estimates (coefficient kappa, tau-b correlations, and classroom-level mean differences between grades and test scores). In addition, only small amounts of grade–test score convergence were observed between teachers; a much greater proportion of variance lay within classrooms and subjects. Appraisal style correlated weakly with convergence rates, but was most strongly related to assigning students to the same performance level as the test. Therefore using recommended grading practices may improve the quality of SBPR grades to some extent.

Keywords: assessment, grading, multiple measures

Standards-based progress reports (SBPRs) differ from traditional letter grade, percentage, narrative, or pass/fail report cards by requiring teachers to report student performance levels on specific educational goals instead of broad content areas. It is believed that if teachers must assess student progress on precise goals or objectives, they will be more likely to focus their instruction on them as well. Therefore, SBPRs have emerged as a standards-based reform lever. Districts often implement SBPRs to provide a measure of standards attainment that supplements state assessment scores in helping parents to understand student achievement. Taken together, SBPRs and state assessment scores can provide a richer description of academic progress than is provided with traditional report cards.

SBPRs are increasing in popularity in part because they are believed to improve communication with parents (Guskey & Jung, 2009). Currently used in a wide array of school districts across the United States, SBPRs take a variety of forms. While they usually report performance on specific skills rather than broad content areas (e.g., rating grasp of number sense instead of overall mathematics achievement) and abandon traditional letter grades in favor of other descriptors, SBPRs vary in their choice of skills and descriptors. Some SBPRs include many quite specific objectives (e.g., "multiplies two-digit numbers," "identifies author purpose," etc.) while others use a smaller number of more general terms (e.g., "number sense," "comprehension," etc.) and districts might choose to adopt the performance level descriptors used on the state assessment (e.g., "advanced," "proficient," "basic," and "below basic") while others use district-developed terms.

SBPRs are an established way of reporting student learning, and have been addressed in the literature for the past 20 years (Clarridge & Whitaker, 1994; Guskey, 2001; Guskey & Bailey, 2009; Scriffiny, 2008). One established advantage of SBPRs is that they improve communication about student achievement, in particular helping teachers to differentiate between process, progress, and the quality of student work products (Guskey, 2001). They have emerged more recently as a standards-based reform lever in that they require teachers to become intimately familiar both with state standards and with the performance level descriptors used on high-stakes assessments; one way to get teachers to focus their instruction on state standards is to mandate that they grade students on them. However, as far as we know, the linkage between SBPR grades and standards-based assessment scores has not been explored in the academic literature.

In addition, districts might use SBPRs to provide an alternative measure of standards attainment that they can couple with state test scores within a multiple measures framework. As such, we would expect convergence between grades and test scores. The value in SBPR grades, however, lies in the

Megan E. Welsh, Department of Educational Psychology, Gentry Building Room 335, 249 Glenbrook Road, Unit 3064, Storrs, CT 06269-3064; megan.welsh@uconn.edu. Jerome D'Agostino, College of Education and Human Ecology, 210A Ramseyer Hall, 29 W. Woodruff Avenue, Columbus, OH 43210. Burcu Kaniskan, Pearson, 19500 Bulverde Road, San Antonio, TX 78259-3701.

information provided about student performance that state test scores do not capture: attainment of skills at different points in the school year, performance on tasks that require students to more deeply explore a skill, and/or the ability to demonstrate knowledge in ways that paper and pencil tests cannot address (Baker, 2003; Guskey, 2007). We first address the literature on grade–test score convergence and then further discuss the importance of including classroom assessment results as one of multiple measures of student performance.

Literature on Grade–Test Score Convergence

Several researchers have examined the correspondence between grades and test scores, but have not examined the correspondence between SBPR grades and standards-based assessment scores. Studies have generally found a moderate association between grades and test scores and often attribute discrepancies to incorporation of nonacademic factors into grades (Brennan, Kim, Wenz-Gross, & Siperstein, 2001; Martinez, Stecher, & Borko, 2009; Willingham, Pollack, & Lewis, 2002).

We believe that challenges to convergence extend beyond teachers' use of nonacademic factors in grading. First, large-scale assessments may not adequately capture student attainment of the standards. Polikoff, Porter, and Smithson (2011) examined the alignment between standards and assessments in 19 states and found that only about half of test items addressed state standards. In addition, only half of the standards were included on the test. Teachers seem aware of this disconnect. McCombs (2005) surveyed teachers and principals about their opinions of state standards and assessments in three states and found that teachers felt that state assessments were not good measures of standards attainment.

Second, teachers may have difficulty interpreting the intent of state standards and therefore operationalize them incorrectly in their classrooms. During her observations of district curriculum committees working to align state standards and district curriculum materials, Hill (2001) concluded that teachers interpreted the same objective quite differently and experienced difficulty in coming to consensus about what is intended by standards documents. Similarly, Conley (2000, April) theorized that the grade–test score relationship may be weakened due to variations in the way that teachers operationalize learning goals and in their requirements for a "proficient" grade. D'Agostino, Welsh, & Corson (2007) confirmed this theory by examining the degree of alignment between operationalization of state mathematics standards on a state test and in classrooms. They then used degree of alignment to predict mathematics performance and found both that teachers varied in the ways that they implemented the standards and that the degree of alignment predicted student achievement.

Third, the grading practices teachers use may jeopardize the reliability of grades and therefore weaken the link between grades and academic achievement. Teachers may inflate grades with nonacademic extra credit assignments, base grades on improvement instead of mastery, or incorporate formative assessments into summative scores, all of which are unrelated to how much a student knows and can do at the end of a grading period (Brookhart, 1994; McMillan, 2001). In addition, grading practices may differ across content areas, by perceived rigor of the course, and by the policies of a particular school or district.

Finally, because classroom assessments and large-scale tests are used differently, and the characteristics required for the assessments to be of high quality vary (Airasian & Jones, 1993; Brookhart, 2003; Cizek, 2009), scores on classroom assessments and state tests might yield different but equally valid information. For example, a student may receive a grade of "exceeds" on extended projects that require deep thinking about a topic, but only receive a score of "meets" on the state assessment because it requires students to work quickly and to think through problems on a wide array of content—a very different (but also important) skill.

Despite these concerns, the degree of SBPR grade–standards-based assessment score convergence should be examined. Both measures purport to address the same construct and are the main sources of information parents get about their child's school performance. If the two measures provide drastically different information about student performance, parents are likely to be confused and concerned. Therefore, it is important to examine both the degree of convergence between the two measures and factors that contribute to differences between the two scores.

Conceptualizing SBPR and Standards-Based Test Score Convergence

SBPRs offer a unique opportunity to examine the connection between teacher appraisals and state test scores because SBPRs require teachers to report student progress on the educational objectives which the state test was designed to measure. Although the studies we have discussed thus far are informative, they suffer from a common limitation—teacher ratings did not necessarily address the same skills as those measured by the test. Because SBPR scores directly reflect the attainment of state standards using the same performance level descriptors employed by the state test, we can examine the convergent validity of SBPRs and state test scores (Campbell & Fiske, 1959).

Convergent information is a fundamental source of validity evidence as stipulated in the *Standards for Educational and Psychological Testing* (*Standards*; American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999). However besides the technical reasons, convergent evidence is necessary given the mistrust of grades and state tests by key stakeholders. Parents should be ensured that both indicators provide supporting evidence of student progress, and teachers must believe the state tests yield accurate results before they will embrace the standards-based reform effort.

In addition, it is important to identify grading practices likely to yield SBPR appraisals consistent with state test results. Little is known about how teachers convert student performance on classroom assessments into SBPR grades. And, while parents might expect that the same degree of knowledge and skill required to meet the standard on a state assessment is required for a "meet" the standard grade, it is unclear whether teachers picture grades in the same way (Waltman & Frisbie, 1994). Even when teachers and parents have a shared interpretation, it may be difficult for teachers to generate grades that set cuts between performance levels consistently with the state test. This difficulty is likely attributable both to a mismatch between teachers and the test in the degrees of competency required for different performance levels and with measurement error surrounding cut points on both measures. Regardless of the cause, steps should be taken to strengthen the consistency in how performance level cuts are envisioned and implemented.

Guskey and Bailey (2001) discuss the four main steps that must be taken to produce accurate SBPRs. First, the learning goals that define what students will know and do must be articulated. Oftentimes those goals are prespecified within a set of district or state academic standards. Second, the student performance indicators of each goal must be stated. Hence, teachers must figure out the tasks and activities that will reveal each student's progress in meeting the goals. The third step requires teachers to define graduated steps of performance that indicate a student's development on multiple performance levels. The teacher must consider the different degrees of student performance on the indicators and define the thresholds between each level. Finally, the actual reporting devices and SBPR format must be created to communicate the results most effectively to parents and students.

This is obviously a very challenging process, one that requires teachers to interpret and operationalize often vague goals in a way that is consistent with the intentions of standards developers and assessment publishers. For example, one mathematics standard requires students to "solve grade-level appropriate problems using estimation" without guidance about what constitutes a "grade-level appropriate problem." It is likely that instruction and assessment on this skill looks very different from classroom to classroom and may be inconsistent with the intentions of standards writers. Determining student performance levels also requires a common understanding of behaviors associated with each level, a task made more difficult because state performance level definitions are also often ambiguous. Therefore, it is unlikely that teachers similarly conceptualize attainment of state standards or that they use a completely consistent approach in assigning SBPR grades.

Some assessment programs address this issue by providing detailed performance level descriptors for each objective. For example, the Namibian National Standardized Achievement Test (NNSAT) provides descriptions of performance level categories by competency that describe the varying levels of achievement related to "identify and place numbers on a number chart and number line" such that a learner who is below basic can "identify and place up to 2-digit numbers"; a basic learner can "identify and place up to 3-digit numbers;" an above-basic learner can "identify and place up to 4-digit numbers;" and an excellent learner can "identify and place numbers up to 10,000" (NNSAT, 2011). This practice is likely to yield a greater degree of consistency between grades and test scores because it helps to clarify what different levels of performance looks like on an objective-by-objective basis.

In our opinion, however, the promise of SBPRs to offer rich descriptions of student achievement outweighs the challenges. In particular, SBPRs present an opportunity to combine with test scores in generating multiple measures of student performance and are especially appealing because they are expressed on the same scale as the state test. The contribution of SBPRs to multiple measures is discussed next.

Benefits of Including SBPR Grades in Multiple Measures of Student Achievement

Multiple measures are intended to improve the quality of information about students and decisions related to their education. As measurement professionals, we recognize that all measures are flawed, necessitating the integration of many sources of evidence in decision making. This is captured in the *Standards* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999), which state: "In educational settings, a decision or characterization that will have major impact on a student should not be made on a simple test score. Other relevant information should be taken into account if it will enhance the overall validity of the decision" (pp. 147–148). However, little consensus exists about what constitutes a multiple measure, how measures should be evaluated for inclusion in a decision-making system, and how they should be combined (Henderson-Montero, Julian, & Yen, 2003).

Baker (2003) argues that including classroom-based assessments as a multiple measure broadens inferences about learning to reflect deeper and more intensive aspects of student achievement, allowing students a wider array of methods they could use to demonstrate their knowledge, an opinion shared by Guskey (2007), who asserts that multiple measures that incorporate performance on classroom assessments, teacher observations of students, and other teacher-generated measures are needed to fully capture the array of skills students are expected to learn.

A key challenge to combining multiple measures to make inferences about student performance is finding measures that address attainment of state performance standards, conceptualize performance in similar ways, and that are scaled similarly to allow for meaningful aggregation of results (Schafer, 2003). If the convergence between SBPR grades and standards-based assessments is established and if performance level cuts are roughly equivalent, then SBPRs offer a unique opportunity to easily construct a multiple measure of student achievement that combines grades and test scores because the results of each measure are presented on the same scale and share the same interpretation. Therefore, research is needed to determine whether SBPR grades and test scores share similar conceptualizations of proficiency and to identify promising strategies for transforming classroom assessment results into SBPR grades. This study examines grading practices in one district to address the following questions:

1. What degree of correspondence is observed in SBPR grades and state test scores in reading, writing and mathematics?
2. How much do convergence rates vary across teachers, years, and content areas?
3. Do teacher appraisal styles correlate with degree of grade–test score convergence?

Method

Participants

Teachers from 11 elementary schools in one suburban school district located in the southwestern United States participated in the study. Like other suburban districts, the district serves a predominantly white (68%), middle-class (33% eligible for free or reduced-price lunch) student body, and is moderately sized, serving approximately 13,000 students in 17 schools. Students in the district are also relatively high-performing. They outperformed the state as a whole on the state assessment; 84% of district third-graders met or exceeded mathematics standards, compared with 76% of third-graders statewide.

The district implemented SBPRs for 3 years at the time of data collection. In the first year, teachers were instructed to grade student performance on specific objectives using the same performance level descriptors as appear on the state test. A brief definition of each performance level consistent with that provided for the state assessment program was included on the SBPR form and also on informational materials distributed to teachers and parents. It is unclear how familiar teachers were with the more detailed performance level descriptors generated by the state department of education. When asked how they define "meeting" state standards, teachers tended to say either that "meets" is analogous to receiving a "B" grade or that "meets" represents grade-level mastery of all or most objectives, but falls short of exceptional performance. No teacher mentioned the state-generated performance level descriptors when questioned about their definition of "meets the standard."

To help teachers adjust to the new reporting system, the district provided professional development on standards-based assessment and grading. These sessions emphasized the importance of keeping effort separate from standards-based grades (effort is graded in its own section of the standards reports) and suggested approaches for generating objective-level scores. However, SBPR forms and instructions for generating grades were changed after the first year.

Teachers were initially asked to grade based on "patterns of progress" over the school year by grading according to both overall level of achievement and degree of improvement achieved. That is, they were instructed to consider whether student performance consistently improved (or declined) on specific objectives over the course of the semester rather than taking an average across all assessments used. District officials believed this method more accurately reflected competency because those who started out having mastered the concept and those who did not initially understand but eventually attained the skill would be graded differently.

Teachers were also asked to provide performance level grades consistent with the state definitions. For example, the definition of a score of "meets" the standard involved "demonstrate(ing) solid academic performance on subject matter as reflected by the reading, math, and writing standards. Students who perform at this level are prepared to begin work on materials that may be required for the next grade level. Attainment of at least this level is the goal for all students" (Arizona Department of Education, 2005). Applying this definition in grading proved especially challenging and was changed in the second year. Instead of using the state's definitions, teachers took averages across assessments and converted them to performance levels using the standard method applied to letter grades. That is, students who average 90% correct and above were graded "exceeds," those who average between 80 and 90% correct were graded "meets," and so on. These adaptations to the grading system eased the record-keeping requirements placed on teachers, made grades more interpretable for parents, and generally resulted in a more straightforward grading system. They also introduced inconsistencies in the way that SBPRs and the state test conceptualized performance levels.

The study used data collected from all third- and fifth-grade teachers in the 11 schools for some analyses and a smaller subset of teachers for others. Third- and fifth-grade teachers were selected because state test scores were only available at Grades 3 and 5 in the years in which test score–SBPR convergence was examined. Participants can be separated into two categories: teachers who participated only in administrative record review and teachers who were interviewed. Administrative data (SBPR grades and state test scores) were provided for all third- and fifth-grade teachers working in the district over a 2-year period (39 third grade classrooms in Year 1, 40 third grade classrooms in Year 2, and 43 fifth grade classrooms in both Years 1 and 2). Fewer teachers are included in most analyses because only a small number of students had both valid test scores and SBPR grades in some classrooms. Ten third grade teachers (9 interviewees) and 15 fifth grade teachers (7 interviewees) had 2 years of SBPR grades and tests scores in all three content areas (reading, writing, and mathematics). Partial results were collected for many more teachers (and interviewees), they were missing data in a particular content area or for an entire year.

Thirty-seven teachers (17 third-grade teachers and 20 fifth-grade teachers) participated in interviews about their mathematics assessment and grading practices at the end of the third year of SBPR implementation. Interviews were conducted in the third year of SBPR implementation with teachers who had participated in at least one full year of SBPR grading at third or fifth grade and who were currently responsible for mathematics instruction (the focus of the larger study from which these data were drawn). We identified 67 teachers who met these criteria (36 third grade, 31 fifth grade) and attempted to recruit every teacher.

The 30 teachers who were not interviewed did not participate for the following reasons: they simply declined to participate (23 teachers), they were out on disability leave during the data collection period or had a major medical issue (3 teachers), they did not teach math (the focus of the larger study, 3 teachers), and they went on maternity leave (1 teacher). Participants and nonparticipants had similar levels of experience teaching at third or fifth grade: 60% of participants had taught at their assigned grade level for more than 2 years (and should therefore be familiar with the standards) compared with 53% of nonparticipants.

The interviewed teachers varied a great deal in their education level, overall teaching experience, and number of years with the district. Although participants taught for 13 years on average, we spoke with two first-year teachers who were simultaneously refining their instructional skills, learning the curriculum, and developing their approach to standards-based grading. Others had been teaching for as long as 25 years and were either creating new strategies to support the reform or were using the parts of the grading system they found valuable and ignoring less helpful components. Approximately one-third of teachers had master's degrees in education, but only one had content area expertise in mathematics (that teacher was working towards a mathematics endorsement at the time of interviews).

Finally, at the time of interviews, all teachers were required to use a district-adopted mathematics text and the adoption of a district reading text was in progress, with implementation slated for the following year. The district emphasized that teachers should not supplement material or deviate from the text pacing. The text provides a range of assessment options to teachers, which serves as additional encouragement to use only the provided materials. Many third-grade teachers said that the text did not address most state standards and believed that their students would not do well on the test if they followed the district directive. While some teachers opted to stick with the text, many deviated from it because of this concern. In contrast, fifth grade teachers reported a high degree of alignment between the mathematics text and state standards. Therefore, we anticipated significant variation in the degree to which teachers implement standards-based instruction, especially at third grade.

Materials

Three data sources contribute to this study: standards-based report card grades, state test scores, and teacher interviews. Standards-based progress report grades from two school years were collected for all third- and fifth-grade students in the district. SBPRs varied a great deal from grade level to grade level and from year to year, as described in the participant section, necessitating analyses by grade level/school year cohort.

The SBPR form itself changed substantially from Year 1 to Year 2 of the study; the number of grades generated and the level of detail in graded skills changed. Teachers were initially asked to grade on a variety of performance objectives from the state standards in math, reading, and writing (e.g., "multiplies whole numbers"), while content area grades were not required. In the second year, objective-level grades were abandoned in favor of grading by subject (fifth grade) and strand (third grade). Strands require teachers to group objectives into broad categories within a content area (e.g., number sense or geometry) while subject grades refer to the content area involved (e.g., reading, math, writing). To further alleviate the challenges associated with standards-based grading, strand-level grades were no longer reported using performance level descriptors at fifth grade. Teachers used one of four ratings for each strand: "demonstrates consistently," "developing," "needs support," and "not evaluated."

State test scores were also collected for the same 2-year period. The state standards-based test gauges student content knowledge in reading, writing, and mathematics. The reading and mathematics tests are comprised of between 76 and 84 multiple choice items. The writing test is a constructed-response test which requires students to respond to a writing prompt. In all subjects, performance level scores are generated on a four point scale: "falls far below," "approaches," "meets," and "exceeds." The test is used to meet the requirements of the No Child Left Behind Act of 2001 (NCLB, 2002) and to determine state accountability ratings. Because of its strong association with state and federal accountability systems, teachers are keenly aware that the test is used to evaluate their instruction.

It is important to note that the approaches used to determine performance levels are different for each measure. On the fifth grade state mathematics test, students were required to correctly answer at least 86% of items to receive a score of "exceeds," between 80 and 85% of items to receive a score of "meets," between 52 and 79% of items to receive a score of "approaches," and 51% or fewer items to receive a score of "falls far below." In contrast, SBPR grades were conceptualized similarly to letter grades ("exceeds" is an A, "meets" is a B, etc.) or were based on each teachers' perception of the kinds of performance associated with different scores.

Test scores were matched to spring progress report grades to gauge test score–SBPR convergence. Analyses were restricted to students both with valid state test scores (e.g., those students tested off grade level or with nonstandard accommodations were omitted) and with progress report grades that focused on grade-level standards (some students grades were adjusted per their individualized educational plans). The scores for approximately 750 students could be merged per year at each grade level, for a total sample size of 3,026 students nested in roughly 80 classrooms and 11 elementary schools.

Finally, teachers were interviewed about their assessment and grading practices in mathematics. Interviews were part of a larger study of standards-based instruction and assessment in elementary mathematics. As such, interview topics ranged from instructional practices used with specific mathematics objectives, to assessment and grading methods, to teacher reviews of test items. Interviews were conducted after the study team analyzed the degree of convergence between standards-based report card grades and state test scores and had concluded that teachers generally assigned grades below state test scores in mathematics and above state test scores in reading and writing. Teachers were also asked if they could provide any insight into this discrepancy. Interviews lasted between 90 minutes and 2 hours, with approximately 30 minutes devoted to assessment and grading. The assessment and grading portion of the interview addressed the following topics:

- How teachers assess student learning, both formally and informally,
- the purposes of different assessments,
- how teachers decide which skills to assess (e.g., do they base assessments solely on the curriculum or do they take state standards into account?),
- what types of information teachers use in assigning grades,
- the role of overall achievement level, amount of improvement made over the school year, and effort in assigning grades,
- the weight assigned to different types of information in assigning grades,
- what a grade of "meets" represents to teachers,
- the frequency of assessment (especially those that contribute to grades), and
- the methods used to convert assessment scores to progress report grades.

Procedure

Spring SBPR grades were entered into a database along with student identification numbers provided by the school district. Fifth-grade scores were provided for each content area (i.e., math, reading, and writing) and were entered as they appeared on progress reports. All other grades were presented by strand or objective and were averaged and rounded to the nearest whole number to arrive at overall content area scores. SBPR grades were then merged with student-level test scores using the student identifiers and convergence was examined according to a variety of indices intended to gauge the degree of consensus and consistency between scores and also teacher rigor in grading.

We define consensus as the degree to which ratings match exactly and examine consistency according to the similarity in student rank orders on grades and on test scores, a distinction elucidated by Kozlowski & Hattrup's (1992) paper on interrater agreement. It is important to consider both consensus and consistency because it is possible for convergence estimates to be rather low, but correlation indices to be strong if one indicator yielded scores that were more stringent than the other. Consensus is gauged using Cohen's (1960) coefficient kappa, while the degree of consistency is measured using Kendall's (1955) tau-b rank order correlation coefficient. Cohen's kappa estimates the degree of consensus while correcting for chance agreement rates by adjusting agreement rates with the marginal distribution of scores (Cohen, 1960) while Kendall's tau-b values estimate the difference between the probability that SBPR grades and state test scores are in the same order and the probability that they are in different orders (Kendall, 1955). Both kappa and tau-b range in value from zero to one, with values close to one indicating a strong relationship and, like correlation coefficients, can have positive or negative values with polarity indicating the direction of the relationship (Agresti, 1996).

Finally, we assessed teacher rigor in grading by taking the difference between state test scores and SBPR grades where positive difference scores indicate that teachers assigned grades lower than those observed on the state test. All agreement indices were calculated at the classroom level where rigor is determined by taking the mean difference within each classroom.

Teacher interviews were audiotaped and transcribed. All transcripts were initially coded by the first author who is a measurement faculty member with elementary school teaching experience and who teaches classroom assessment methods classes to preservice teachers. Interviews were transcribed and coded iteratively. Initial codes were generated based on methods widely believed to yield reliable grades (Linn & Miller, 2005; Oosterhof, 2003). As transcripts were reviewed, the coder took notes on additional themes that emerged from the data. Based on these notes, transcripts were reread and additional codes were generated.

The following assessment characteristics were coded using a three-point scale ("clearly evident," "somewhat evident," and "not evident") for each teacher. Coding schemes are presented below:

- Performance-focused. Whether the teacher focused on measuring standards achievement instead of effort. Teachers who reported taking class participation and effort into account in assigning grades were scored "0," those who more subtly included effort in grading (such as increasing borderline grades because of good effort) were coded "1," and those who reported not taking effort into account in any way were coded "2."
- Overall achievement. Extent of focus on overall achievement rather than student progress. Teachers who reported grading students based on the progress made over the course of the semester, rather than on skill attainment were coded "0," teachers who calculated grades and then made adjustments based on progress were coded "1," and teachers who graded students based solely on overall achievement level were coded "2."
- Frequently assessed. Teachers who collected assessment data for grading purposes less than once a week were coded "0", those who assessed weekly were coded "1", and those who assessed at least twice a week were coded "2."
- Multiple approaches. Use of approaches that allow teachers to gauge different aspects of a skill. Teachers who used only the assessments provided by the district mathematics text were coded "0," those who occasionally supplemented text-based assessments with other assessments were coded "1," and teachers who regularly assessed in ways that required students to show understanding using a variety of modalities (presentations, performance tasks, paper and pencil tests, etc.) were coded "2."
- Linked assessments to objectives. Whether teachers maintained objective-based records. Teachers who did not maintain records on the objectives assessed were coded "0," teachers who only identified the general area (e.g., algebra, geometry, etc.) assessed or only linked assessments and objectives in a limited number of cases were coded "1," and those who consistently linked assessments to objectives were coded "2."
- Clear grading method. Whether teachers had a clear method of converting students' scores on assessments to progress report grades. Teachers who could not explain a set grading method were coded "0," those who described a method but also said that they use their general knowledge of students in grading were coded "1," and those who described their method and reported consistently employing it were coded "2."
- Created assessments. Teachers who did not create assessments were coded "0," teachers who created their own assessments to measure skills also assessed by text-based assessments were coded "1," and teachers who created assessments to address objectives not covered by the text were coded "2."
- Assessed most objectives. Teachers who did not make an effort to assess the objectives in the state standards were coded "0," teachers who assessed some state standards but did not attempt to cover the full range of objectives were coded "1," and teachers who reported assessing the full range of performance objectives were coded "2."
- Standards-focused. The degree to which teachers focused on assessing standards more than curriculum attainment. Teachers who limited their assessment to the district-adopted text were coded "0," teachers who focused both on curriculum attainment and on the standards were coded "1," and teachers who concentrated on state standards were coded "2."

A doctoral student in educational measurement coded 10 randomly selected transcripts to establish inter-rater reliability after being trained on the coding scheme and completing three practice transcripts. The scores were compared using a weighted kappa statistic, which examines the degree of agreement after correcting for chance agreement levels and assigns partial credit for ratings that are similar but do not match exactly (Cohen, 1968). Weightings gave full credit to exact matches, half a point to ratings only one category apart, and scores two categories apart were not counted as matching. We then calculated the maximum weighted kappa possible based on the marginal distributions of scores and examined agreement rates relative to the maximum possible level of agreement. We found strong levels of agreement for all indicators, the smallest $\kappa_{\text{weighted}} / \kappa_{\text{mx}} = .71$ (Table 1).

We created a composite mathematics appraisal style measure using exploratory factor analysis; we generated a factor

Table 1. Interrater Reliability Estimates for Ratings of Teacher Grading Practice

Grading Practice	$\kappa_{weighted}$	κ_{max}	$\kappa_{weighted}/\kappa_{max}$
Performance-focused	0.52	0.52	1.00
Overall achievement	0.67	0.67	1.00
Frequently assessed	0.04	0.04	1.00
Multiple approaches	0.65	0.88	0.74
Linked assessments to objectives	0.65	0.88	0.74
Clear grading method	0.70	0.90	0.78
Created assessments	0.55	0.78	0.71
Assessed most objectives	0.56	0.56	1.00
Standards-focused	0.79	0.79	1.00

Table 2. SBPR–test Score Agreement, by Content Area, Grade Level, and Year

Subject	Cohort	*N*	κ	Tau-b	Mean Difference (*SD*) Test Score–SBPR Grade
Reading	Grade 3, Year1	706	0.321[a]	0.476[a]	– 0.068 (0.679)[a]
	Grade 3, Year2	719	0.305[a]	0.498[a]	– 0.038 (0.727)
	Grade 5, Year 1	705	0.194[a]	0.429[a]	– 0.260 (0.785)[a]
	Grade 5, Year2	753	0.203[a]	0.456[a]	– 0.185 (0.953)[a]
Writing	Grade 3, Year1	716	0.190[a]	0.378[a]	– 0.042 (0.642)
	Grade 3, Year2	661	0.118[a]	0.355[a]	– 0.309 (0.780)[a]
	Grade 5, Year 1	731	0.129[a]	0.319[a]	– 0.186 (0.777)[a]
	Grade 5, Year2	750	0.175[a]	0.416[a]	– 0.305 (0.861)[a]
Mathematics	Grade 3, Year1	724	0.268[a]	0.469[a]	0.164 (0.687)[a]
	Grade 3, Year2	702	0.229[a]	0.472[a]	0.123 (0.804)[a]
	Grade 5, Year 1	711	0.142[a]	0.443[a]	0.332 (0.865)[a]
	Grade 5, Year2	767	0.239[a]	0.496[a]	0.309 (0.921)[a]

[a] Value is different from zero at a statistically significant level $p < .05$.

score by forcing a one-factor solution using the generalized least squares extraction method. The appraisal style composite measure is moderately reliable (coefficient alpha = .66).

Results

Correspondence Between SBPR Grades and Test Scores

We observed a moderate to weak correspondence between SBPR grades and test scores, depending on the measure used. Kappa, tau-b, and mean difference values are presented in Table 2 for each content area and are disaggregated by grade level and year of SBPR implementation. Kappa values ranged from $\kappa = .118$ to $\kappa = .321$, which are only "slight" to "fair" levels of agreement under Landis & Koch's (1977) guidelines. Tau-b correlations are somewhat stronger, indicating that tests and teachers were more likely to rank-order students consistently than they were to assign the same score, but the relationship was still only moderate, ranging from $\tau = .319$ to $\tau = .496$. Even so, the degree of association between grades and test scores was significantly different from zero for all analyses ($p < .05$).

SBPR grades and test scores are most weakly related in writing, the subject gauged solely with constructed-response items (writing prompts). This finding might be explained by differences in the way that writing is assessed on the test and in the classroom. In interviews, teachers indicated that their assessments include both examples of student writing and exercises intended to build grammatical skills. In addition, while both the state test and classroom assessments focused on the Six Traits of Writing Rubric, it is unclear how either the state assessment or classroom teachers converted this six-part analytic rubric into one writing score. Differences in approach, teachers' tendency to grade on one trait at a time instead of applying the entire rubric to one piece of writing, or the fact that state assessment results are based on an observation made at only one point in time, may account for the lack of convergence.

Teachers also appear to grade less rigorously than the test in reading and writing and more rigorously in mathematics. Mean difference scores were universally negative across grades and years in reading and writing, indicating that grades were higher than test scores. However, the mean difference between test scores and SBPR grades was not statistically significant for third-grade writing in Year 1 and third-grade reading in Year 2. In contrast, mean differences were universally positive in mathematics; teachers were more rigorous than the test in their evaluations of students. The magnitude of these differences varied considerably across years, grade levels, and subjects, from as little as 0. 038 of a performance level in third-grade reading in Year 2 to 10 times that amount—.332 of a performance level—in fifth-grade mathematics in Year 1. We speculate that elementary teachers may be less confident in the field of mathematics and therefore reticent to assert that students have mastered the material. The teachers we interviewed seemed surprised that grades were more rigorous in mathematics and said they could not offer an explanation for this difference.

Teachers were also most consistent with state test scores for those students who scored at the "meets" performance level on the state test and were least accurate in grading students in the "falls far below" category. Because the teachers studied worked with relatively high-performing students, it is predictable that they would be better versed on how to distinguish between students at the upper end of performance. Figures 1–3 present the percent of students graded the same as, less rigorously than, and more rigorously than they were scored on the state test.

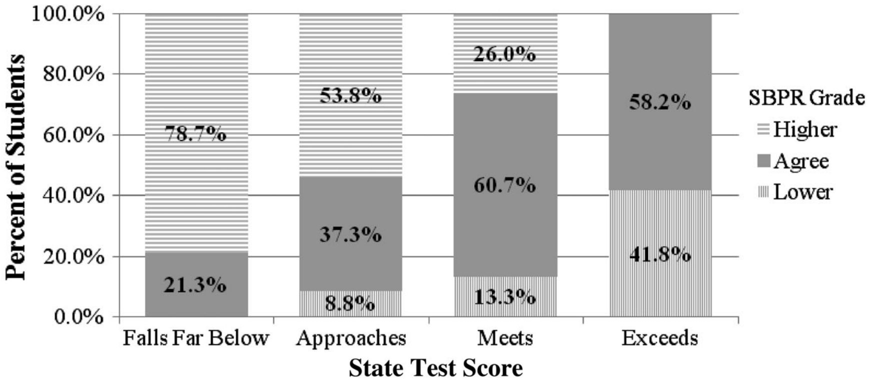

FIGURE 1. Percent of students graded higher, lower, and equal to their state test score in reading across grade levels and years.

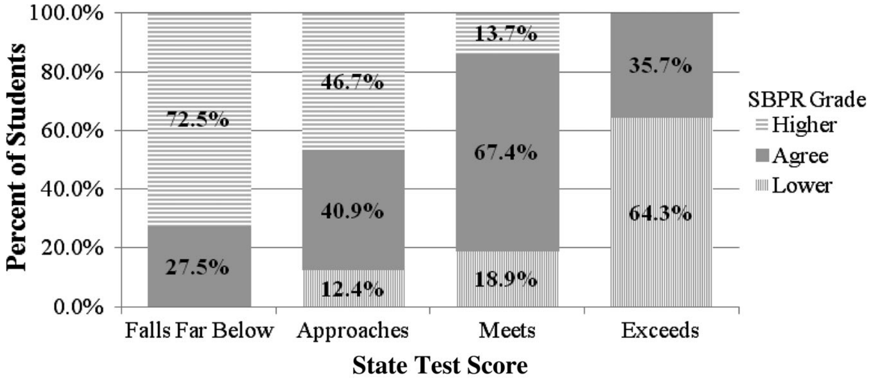

FIGURE 2. Percent of students graded higher, lower, and equal to their state test score in mathematics across grade levels and years.

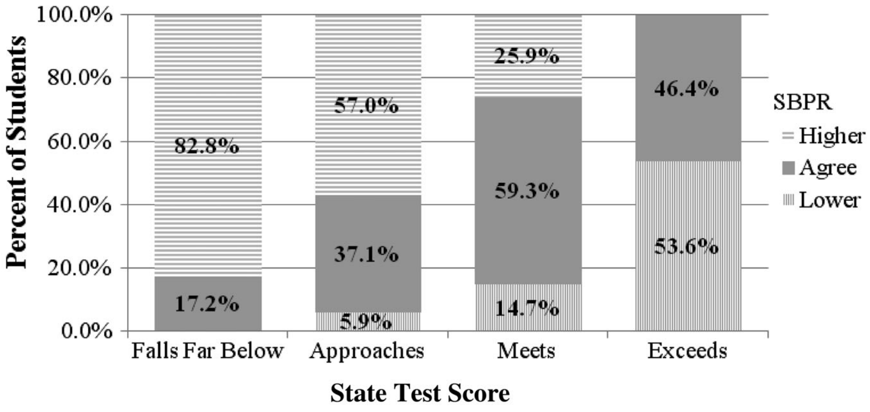

FIGURE 3. Percent of students graded higher, lower, and equal to their state test score in writing across grade levels and years.

Variability in Convergence Rates

We calculated the proportion of variability in classroom-level agreement rates found between teachers, subjects, and years and in the interactions among them, conducting separate analyses with tau-b, kappa, and mean difference scores as the dependent variables. To do so, we conducted a generalizability study (Brennan, 2001), treating classroom-level convergence estimates as teacher scores. Our study followed a p × S × Y design, with teachers (p) fully crossed with subjects (S), and years (Y). We used the variance components procedure found in SPSS Statistics 17, Release Version 17. 0. 0 (SPSS, Inc. 2008) using restricted maximum likelihood estimation.

This approach is preferred because it allows us to decompose the variance between the three elements while allowing for only one score within each cell of the design. In addition it focuses on estimation of variance components which supports the goal of this analysis—to describe the amount of variation in convergence rates attributable to differences between teachers, subject areas, and years which we treat as a proxy for different SBPR forms and policies given the changes in SBPR implementation between Years 1 and 2 discussed earlier. Although the main goal of most generalizability studies is to generate variance components that can be used to estimate the reliability of scores, estimating the

Table 3. Estimated Variance Components and Their Percentages for SBPR–Test Score Comparisons

		Kappa			Tau-b			Mean Test–grade Difference	
Source of Variability	*n*	Estimated Variance Component	Percentage of Total Variance	*N*	Estimated Variance Component	Percentage of Total variance	*n*	Estimated Variance Component	Percentage of Total Variance
Teacher	38	.0012	3.5%	38	.0031	7.2%	47	.0028	1.6%
Subject	3	.0044	12.7%	3	.0090	20.7%	3	.0600	34.1%
Year	2	.0000	.0%	2	.0000	.0%	2	.0000	.0%
Teacher* Subject	114	.0013	3.7%	114	.0000	.0%	141	.0146	8.3%
Teacher*Year	76	.0000	.0%	76	.0053	12.2%	94	.0440	25.0%
Subject*Year	6	.0000	.0%	6	.0002	.6%	6	.0060	3.4%
Teacher*Subject*Year, error	228	.0277	80.1%	228	.0257	59.4%	282	.0487	27.7%

reliability of kappa, tau-b, and mean difference scores is not a goal of this study and reliability estimates are not provided.

Results for all three convergence estimates are presented in Table 3, which shows that teachers did not vary much in their SBPR grade–state test score convergence rates, while the proportion of unexplained variation within teachers remained quite large. This suggests that convergence rates are not affected much by the particular teacher who did the grading, but may be greatly affected by other factors within the classroom, such as student characteristics, a factor that was not accounted for in this analysis. Interestingly, the proportion of unexplained variance was largest for the kappa estimates and much smaller for our measure of rigor. This seems to suggest that teacher rigor in grading is affected considerably less by within-classroom factors than is teacher skill in assigning grades that match state test scores.

In addition, the subject graded greatly impacts convergence. Thirty-four percent of the variation in rigor (mean difference) estimates is attributable to the subject studied, and 20.7% of the variance in tau-b correlations is due to subject, as is 12.7% of the variation in kappa. In contrast, the year of SBPR implementation had no impact on convergence; the particular report card form or approach to generating SBPR grades did not affect the correspondence between grades and test scores. This finding is somewhat mitigated by interview data, which indicate that many teachers continued to grade on a pattern of progress instead of calculating SBPR grades by taking the mean of classroom assessment scores.

Interactions between year and subject also failed to explain variation in convergence. However, the teacher by year interaction explained substantial amounts of variation in rigor (25.0%) and tau-b correlations (12.2%); certain teachers did a better job of rank-ordering students and/or were differentially rigorous graders relative to state test performance depending on the SBPR form and associated grading method implemented. Teachers were also differentially rigorous and/or differentially adept at assigning grades that matched test performance depending on the subject graded. However, the proportions of variance explained by the teacher by subject interactions were smaller than for the other interactions.

Relationship Between Appraisal Style and SBPR–Test Score Convergence

The frequency with which teachers adopted each grading practice is presented in Table 4. The teachers in our sample tended to assess students frequently and to focus on overall achievement instead of taking progress into account. Many (but fewer) teachers also linked assessment items to specific objectives, could describe a clear method used to assign grades, focused on attainment of state standards instead of the district curriculum, or regularly created their own assessments. Clear evidence of grading on performance instead of effort, of assessing most objectives, or of using multiple approaches to gauge student performance existed for a third or fewer of the teachers interviewed.

Correlations between the mathematics appraisal style composite and convergence estimates are presented in Table 5. We observed small correlations between Appraisal Style and kappa coefficients, which reflect the consistency in grades and test scores, in all three subject areas and no correlation with tau-b correlations, which gauge the degree to which rank order based on grades is consistent with rank order based on test scores. Interestingly, appraisal style was weakly and negatively associated with the mean difference between state test scores and grades in mathematics, not correlated in reading, and weakly and positively correlated in writing. This corresponds with the positive mean difference scores in mathematics, negative (and near zero) mean differences in reading, and negative (and slightly larger) mean differences in writing. That is, higher appraisal style scores seem to counteract teachers' tendency to underestimate performance in mathematics and to overestimate performance in writing, relative to state test performance. Strong positive correlations were observed between coefficient kappa and tau-b correlations in the same subject and other small correlations (both positive and negative) were observed between many of the SBPR grade–test score convergence measures.

Discussion

SBPR grades were moderately associated with state test scores, indicating that grades and test scores converged. The moderate degree of association may also suggest that grades capture different aspects of student performance than the test. However, it is also probable that measurement error associated both with test scores and with grades affects this finding, as they likely attenuate the magnitude of association. This finding is consistent with other studies that have shown grades and test scores to be moderately related (Brennan et al., 2001; Conley 2000, April), and illustrates the importance of gathering multiple sources of information when making judgments about students. However, the usefulness

Table 4. Percentage of Teachers Adopting Each Grading Practice

Grading Practice	*N*	Clearly Evident	Somewhat Evident	Not Evident
Performance-focused	37	35.1	16.2	48.6
Overall achievement	37	70.3	18.9	10.8
Frequently assessed	37	64.9	29.7	5.4
Multiple approaches	37	21.6	37.8	40.5
Linked assessments to objectives	37	54.1	10.8	35.1
Clear grading method	37	48.6	27.0	24.3
Created assessments	37	43.2	21.6	35.1
Assessed most objectives	37	32.4	16.2	51.4
Standards-focused	37	45.9	16.2	37.8

Table 5. Correlation Between Appraisal Style and Convergence Estimates, Pooled Across Years and Grade Levels

	1	2	3	4	5	6	7	8	9	10
1. Appraisal style	–									
2. Kappa math	.214	–								
3. Kappa reading	.240	.008	–							
4. Kappa writing	.229	.036	.298[a]	–						
5. Tau-b math	.129	.446[a]	−.184	−.110	–					
6. Tau-b reading	.037	−.177	.656[a]	.207	−.215	–				
7. Tau-b writing	.036	.312[a]	.137	.630[a]	.214	.287	–			
8. Mean difference math	−.237	−.096	−.103	−.104	.087	−.175	−.056	–		
9. Mean difference reading	.057	.026	.289	−.221	−.178	.043	−.117	.088	–	
10. Mean difference writing	.331[a]	.235	.204	.063	−.097	.098	.011	−.263	.129	–

[a] Correlation is significantly different from zero, $p < .05$ ($N = 48$).

of SBPRs as a communication tool is limited if parents receive contradictory information about student performance. Therefore, it is important to identify and reduce sources of inconsistency.

One reason that grades and test scores may differ is that teachers set different internal cut points between performance levels than the test. "Meeting" or "approaching" a standard are somewhat ambiguous concepts, tricky to operationalize; teachers must determine how to implement the objective and what kinds of behaviors constitute different levels of proficiency. The teachers studied grade consistently with the test when students "meet" state standards. This result is promising; understanding what proficiency looks like is the first step in determining achievement levels associated with higher and lower degrees of performance. However, it is also important to make sure that teachers also understand how to identify where students fit across all gradations of performance.

Because the teachers in this study worked with relatively high-performing students, it is predictable that they would be better versed on how to distinguish between students at the upper end. Further research might also examine whether teachers in low-performing schools are most consistent with the test in distinguishing between performance levels at the lower end of the spectrum. If so, then teachers may require additional training to set cuts at performance levels they encounter less often. On the other hand, if teachers in low performing schools are also more consistent at the "meets" level, SBPR implementation may indeed help to elucidate what proficiency looks like. State departments of education could also support this effort by providing concrete information about the skills that delineate between performance levels on an objective by objective basis, as is provided for the NNSAT (2011).

In addition to working towards a common understanding of each performance level, consistency is improved when the methods used to generate grades are standardized. For example, some teachers in our study graded both on achievement and on improvement made during the year while others graded solely on achievement. And half of the teachers either could not explain the process they used to grade or could describe the process they used but admitted that they did not strictly adhere to it. For standards-based grading to work as a multiple measure, teachers need training on the expected grading method and on the importance of its faithful implementation.

While past research has speculated that teacher grading practices contribute to moderate convergence rates (Willingham et al., 2002), we found only small correlations between teacher appraisal style and test-grade convergence, indicating that other factors are at play. However, the fact that the appraisal style measure correlated more strongly with kappa values than with tau-b correlations may suggest that using appropriate grading practices squarely focused on attainment of state standards might have some impact on improving the match between SBPR grades and state test scores. Further study, with a larger and more diverse sample of teachers, is needed to identify those practices that yield high degrees of convergence.

Finally, the limited variation in convergence rates across teachers may indicate that SBPR implementation helps improve consistency in grading practices. Even so, some teachers varied in convergence rates according to the subject graded and the report card form. Therefore, districts should carefully attend to the design of report cards and on the methods teachers might use to distinguish between performance levels in grading, to training teachers on operationalization of the standards, on setting proficiency levels for each objective, on good grading practices, and on the usefulness of SBPRs as a multiple measure that provides information about student attainment of the standards throughout the school year.

For standards-based reform to work, it is important that teachers be well versed not only in the content of state standards, but also in what it means to assign students to specific performance levels in terms of the skills that must be attained or that are yet to be mastered. SBPRs are one promising approach to help achieve that goal in that they can help both parents and teachers think more deeply about student attainment of the standards at multiple points in the school year with results provided proximally to instruction. Struggling with these concepts is the real work of standards-based reform, one in which SBPRs may assist. It is likely that SBPR grades and state test scores will always differ to some extent, the very reason why multiple measures are needed. Careful attention to the grading methods used could limit the degree of discordance and improve the quality of information parents receive.

References

Agresti, A. (1996). *Introduction to categorical data analysis*. New York, NY: John Wiley and Sons.

Airasian, P. W., & Jones, A. M. (1993). The teacher as applied measurer: Realities of classroom management and assessment. *Applied Measurement in Education*, *6*, 241–254.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Arizona Department of Education. (2005). *Arizona's instrument to measure standards: State board approved AIMS performance level descriptors*. Phoenix, AZ: Author. Retrieved on May 29, 2007 from http://www. azed. gov/standards/aims/PerformanceStandards/GeneralAIMSPerformanceLevelDescriptors. pdf

Baker, E. L. (2003). Multiple measures: Toward tiered systems. *Educational Measurement: Issues and Practice*, *22*(2), 13–17.

Brennan, R. L. (2001). *Generalizability theory*. New York, NY: Springer.

Brennan, R. T., Kim, J., Wenz-Gross, M., & Siperstein, G. N. (2001). The relative equitability of high-stakes testing versus teacher-assigned grades: An analysis of the Massachusetts Comprehensive Assessment System (MCAS). *Harvard Educational Review*, *71*, 173–216.

Brookhart, S. M. (1994). Teacher's grading: Practice and theory. *Applied Measurement in Education*, *7*, 279–301.

Brookhart, S. M. (2003). Development measurement theory for classroom assessment purposes and uses. *Educational Measurement: Issues and Practice*, *22*(4), 5–12.

Campbell, D., & Fiske, D. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*, 81–105.

Cizek, G. J. (2009). Reliability and validity of information about student achievement: Comparing large-scale and classroom testing contexts. *Theory into Practice, 48*, 63–71.

Clarridge, P. B., & Whitaker, E. M. (1994). Implementing a new elementary progress report. *Educational Leadership*, *52*(2), 7–9.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*, 37–46.

Cohen, J. (1968). Weighed kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, *70*, 213–220.

Conley, D. T. (2000, April). *Who is proficient?: The relationship between proficiency scores and grades*. Paper presented at the meeting of the American Educational Research Association. New Orleans, LA.

D'Agostino, J. V., Welsh, M. E., & Corson, N. M. (2007). Instructional validity of a state's standards-based assessment. *Educational Assessment*, *12*(1), 1–22.

Guskey, T. R. (2001). Helping standards make the grade. *Educational Leadership*, *59*(1), 20–27.

Guskey, T. R. (2007). Multiple sources of evidence: An analysis of stakeholders' perceptions of various indicators of student learning. *Educational Measurement: Issues and Practice*, *26*(1), 19–27.

Guskey, T. R., & Bailey, J. M. (2001). *Developing grading and reporting systems for student learning*. Thousand Oaks, CA: Corwin.

Guskey, T. R., & Bailey, J. M. (2009). *Developing standards based report cards*. Thousand Oaks, CA: Corwin.

Guskey, T. R., & Jung, L. A. (2009). Grading and reporting in a standards-based environment: Implications for students with special needs. *Theory into Practice*, *48*, 53–62.

Henderson-Montero, D., Julian, M. W., & Yen, W. M. (2003). Multiple perspectives on multiple measures: An introduction. *Educational Measurement: Issues and Practice*, *22*(2), 7–12.

Hill, H. C. (2001). Policy is not enough: Language and the interpretation of state standards. *American Educational Research Journal*, *38*, 298–318.

Kendall, M. G. (1955). *Rank correlation methods*. New York, NY: Hafner.

Kozlowski, S. W. J. & Hattrup, K. (1992). A disagreement about within group agreement: Disentangling issues of consistency versus consensus. *Journal of Applied Psychology*, *77*, 161–167.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*, 159–174.

Linn, R. L., & Miller, M. D. (2005). *Measurement and assessment in teaching* (9th ed.). Upper Saddle River, NJ: Pearson Education.

Martinez, J. F., Stecher, B., & Borko, H. (2009). Classroom assessment practices, teacher judgments, and student achievement in mathematics: Evidence from the ECLS. *Educational Assessment*, *14*, 78–102.

McCombs, J. S. (2005). *Progress in implementing standards, assessments, and the highly qualified teacher provisions of NCLB: Perspectives from California, Georgia, and Pennsylvania (Rand Technical Report No. WR-256-EDU)*. Santa Monica, CA: RAND. Retrieved July 19, 2005, from http://www.rand.org/publications/WR/WR256/

McMillan, J. H. (2001). Secondary teachers' classroom assessment and grading practices. *Educational Measurement: Issues and Practice*, *20*(1), 20–32.

Namibian National Standardized Achievement Test. (2011). *Description of performance level categories by competencies Grade 5 mathematics standard setting workshop*. Windhoek, Namibia: Namibian Ministry of Education.

No Child Left Behind Act of 2001. (2002) Pub. L. No. 107–110, 115 Stat. 1425.

Oosterhof, A. (2003). *Developing and using classroom assessments* (3rd ed.). Upper Saddle River, NJ: Pearson Education.

Polikoff, M. S., Porter, A. C., & Smithson, J. (2011). How well aligned are state assessments of student achievement with state content standards? *American Educational Research Journal*, *48*, 965–995.

Schafer, W. D. (2003). A state perspective on multiple measures in school accountability. *Educational Measurement: Issues and Practice*, *22*(2), 27–31.

Scriffiny, P. L. (2008). Expecting excellence: Seven reasons for standards-based grading. *Educational Leadership*, *66*(2), 70–74.

Waltman, K. K., & Frisbie, D. A. (1994). Parents' understanding of their children's report card grades. *Applied Measurement in Education*, *7*, 223–240.

Willingham, W. W., Pollack, J. M., and Lewis, C. (2002). Grades and test scores: Accounting for observed differences. *Journal of Educational Measurement*, *39*, 1–37.