14

Research on Classroom Summative Assessment

Connie M. Moss

Assessment is unquestionably one of the teacher's most complex and important tasks. What teachers assess and how and why they assess it sends a clear message to students about what is worth learning, how it should be learned, and how well they are expected to learn it. As a result of increased influences from external high stakes tests, teachers are increasingly working to align their CAs with a continuum of benchmarks and standards, and students are studying for and taking more CAs. Clearly, high-stakes external tests shape much of what is happening in classrooms (Clarke, Madaus, Horn, & Ramos, 2000). Teachers design assessments for a variety of purposes and deliver them with mixed results. Some bring students a sense of success and fairness, while others strengthen student perceptions of failure and injustice. Regardless of their intended purpose, CAs directly or indirectly influence students' future learning, achievement, and motivation to learn.

The primary purpose of this chapter is to review the literature on teachers' summative assessment practices to note their influence on teachers and teaching and on students and learning. It begins with an overview of effective summative assessment practices, paying particular attention to the skills and competencies that teachers need to create their own assessments, interpret the results of outside assessments, and accurately judge student achievement. Then, two recent reviews of summative assessment practices are overviewed. Next, the chapter reviews current studies of summative CAs illustrating common research themes and synthesizing prevailing recommendations. The chapter concludes by drawing conclusions about what we currently know regarding effective CA practices and highlighting areas in need of further research.

Setting the Context: The Research on Summative Classroom Assessments

Assessment is a process of collecting and interpreting evidence of student progress to inform reasoned judgments about what a student or group of students knows relative to the identified learning goals (National Research Council [NRC], 2001). How teachers carry out this process depends on the purpose of the assessment rather than on any particular method of gathering information about student progress. Unlike assessments that are formative or diagnostic, the purpose of summative assessment is to determine the student's overall achievement in a specific area of learning at a particular time—a purpose that distinguishes it from all other forms of assessment (Harlen, 2004).

The accuracy of summative judgments depends on the quality of the assessments and

the competence of the assessors. When teachers choose formats (i.e., selected-response [SR], observation, essay, or oral questioning) that more strongly match important achievement targets, their assessments yield stronger information about student progress. Test items that closely align with course objectives and actual classroom instruction increase both content validity and increase reliability so assessors can make good decisions about the kind of consistency that is critical for the specific assessment purpose (Parkes & Giron, 2006). In assessments that deal with performance, reliability and validity are enhanced when teachers specifically define the performance (Baron, 1991); develop detailed scoring schemes, rubrics and procedures that clarify the standards of achievement; and record scoring during the performance being assessed (Stiggins & Bridgeford, 1985).

Teachers' Classroom Assessment Practices, Skills, and Perceptions of Competence

Teacher judgments can directly influence student achievement, study patterns, self-perceptions, attitudes, effort, and motivation to learn (Black & Wiliam, 1998; Brookhart, 1997; Rodriguez, 2004). No serious discussion of effective summative CA practices can occur, therefore, without clarifying the tensions between those practices and the assessment competencies of classroom teachers. Teachers have primary responsibility for designing and using summative assessments to evaluate the impact of their own instruction and gauge the learning progress of their students. Teacher judgments of student achievement are central to classroom and school decisions including but not limited to instructional planning, screening, placement, referrals, and communication with parents (Gittman & Koster, 1999; Hoge, 1984; Sharpley & Edgar, 1986).

Teachers can spend a third or more of their time on assessment-related activities (Plake, 1993; Stiggins, 1991, 1999). In fact, some estimates place the number of teacher-made tests in a typical classroom at 54 per year (Marso & Pigge, 1988), an incidence rate that can yield billions of unique testing activities yearly worldwide (Worthen, Borg, & White, 1993). These activities include everything from designing paper–pencil tests and performance assessments to interpreting and grading test results, communicating assessment information to various stakeholders, and using assessment information for educational decision making. Throughout these assessment activities, teachers tend to have more confidence in their own assessments rather than in those designed by others. And they tend to trust in their own judgments rather than information about student learning that comes from other sources (Boothroyd, McMorris, & Pruzek, 1992; Stiggins & Bridgeford, 1985). But is this confidence warranted?

The CA literature is split on teachers' ability to accurately summarize student achievement. Some claim that teachers can be the best source of student achievement information. Effective teachers can possess overarching and comprehensive experiences with students that can result in rich, multidimensional understandings (Baker, Mednick, & Hocevar, 1991; Hopkins, George, & Williams, 1985; Kenny & Chekaluk, 1993; Meisels, Bickel, Nicholson, Xue, & Atkins-Burnett, 2001). Counterclaims present a more skeptical view of teachers as accurate judges of student achievement. Teacher judgments can be clouded by an inability to distinguish between student achievement and student traits like perceived ability, motivation, and engagement that relate to achievement (Gittman & Koster, 1999; Sharpley & Edgar, 1986). These poor judgments can be further exacerbated when teachers assess students with diverse backgrounds and characteristics (Darling-Hammond, 1995; Martínez & Mastergeorge, 2002; Tiedemann, 2002).

A Gap Between Perception and Competence

For over 50 years, the CA literature has documented the gap between teachers' perceived and actual assessment competence. Teachers regularly use a variety of assessment techniques despite inadequate preservice preparation or in-service professional development about how to effectively design, interpret, and use them (Goslin, 1967; O'Sullivan & Chalnick, 1991; Roeder, 1972). Many teachers habitually include nonachievement factors like behavior and attitude, degree of effort, or perceived motivation for the topic or assignment in their summative assessments. And they calculate grades without weighing the various assessments by importance (Griswold, 1993; Hills, 1991; Stiggins, Frisbie, & Griswold, 1989). When they create and use

performance assessments, teachers commonly fail to define success criteria for the various levels of the performance or plan appropriate scoring schemes and procedures prior to instruction. Moreover, their tendency to record their judgments after a student's performance rather than assessing each performance as it takes place consistently weakens accurate conclusions about how each student performed (Goldberg & Roswell, 2000).

In addition to discrepancies in designing and using their own assessments, teachers' actions during standardized testing routinely compromise the effectiveness of test results for accurately gauging student achievement and informing steps to improve it. Teachers often teach test items, provide clues and hints, extend time frames, and even change students' answers (Hall & Kleine, 1992; Nolen, Haladyna, & Haas, 1992). Even when standardized tests are not compromised, many teachers are unable to accurately interpret the test results (Hills, 1991; Impara, Divine, Bruce, Liverman, & Gay, 1991) and lack the skills and knowledge to effectively communicate the meaning behind the scores (Plake, 1993).

Incongruities in teachers' assessment practices have long been attributed to a consistent source of variance: A majority of teachers mistakenly assume that they possess sound knowledge of CA based on their own experiences and university coursework (Gullikson, 1984; Wise, Lukin, & Roos, 1991). Researchers consistently suggest collaborative experiences with assessments as a way to narrow the gap between teacher perceptions of their assessment knowledge and skill and their actual assessment competence. These knowledge-building experiences develop and strengthen common assessment understandings, quality indicators, and skills. What's more, collaboration increases professional assessment language and dispositions toward reflecting during and after assessment practices events to help teachers recognize how assessments can promote or derail student learning and achievement (Aschbacher, 1999; Atkin & Coffey, 2001; Black & Wiliam, 1998; Borko, Mayfield, Marion, Flexer, & Cumbo, 1997; Falk & Ort, 1998; Gearhart & Saxe, 2004; Goldberg & Roswell, 2000; Laguarda & Anderson, 1998; Sato, 2003; Sheingold, Heller, & Paulukonis, 1995; Wilson, 2004; Wilson & Sloane, 2000).

Two Reviews of Summative Assessment by the Evidence for Policy and Practice Information and Co-Ordinating Centre

Impact of Summative Assessments and Tests on Students' Motivation for Learning

The Evidence for Policy and Practice Information and Co-Ordinating Centre (EPPI-Centre), part of the Social Science Research Unit at the Institute of Education, University of London, offers support and expertise to those undertaking systematic reviews. With its support, Harlen and Crick (2002) synthesized 19 studies (13 outcome evaluations, 3 descriptive studies, and 3 process evaluations). The review was prompted by the global standardized testing movement in the 1990s and sought to identify the impact of summative assessment and testing on student motivation to learn. While a more extensive discussion of CA in the context of motivational theory and research is presented in this volume (see Brookhart, Chapter 3 of this volume), several conclusions from this review are worth mentioning here.

The researchers noticed that following the introduction of the national curriculum tests in England, low achieving students tended to have lower self-esteem than higher achieving students. Prior to the tests, there had been no correlation between self-esteem and achievement. These negative perceptions of self-esteem often decrease students' future effort and academic success. What's more, the high-stakes tests impacted teachers, making them more likely to choose teaching practices that transmit information during activities that are highly structured and teacher controlled. These teaching practices and activities favor students who prefer to learn this way and disadvantage and lower the self-esteem of students who prefer more active and learner-centered experiences. Likewise, standardized tests create a performance ethos in the classroom and can become the rationale for all classroom decisions and produce students who have strong extrinsic orientations toward performance rather than learning goals. Not only do students share their dislike for high-stakes tests but they also exhibit high levels of test anxiety and are keenly aware that the narrow test results do not accurately represent what they understand or can do.

Not surprisingly, student engagement, self-efficacy, and effort increase in classrooms where teachers encourage self-regulated learning (SRL) and empower students with challenging choices and opportunities to collaborate with each other. In these classrooms, effective assessment feedback helps increase student motivation to learn. This feedback tends to be task involved rather than ego involved to increase students' orientation toward learning rather than performance goals.

Impact of Summative Assessments on Students, Teachers, and the Curriculum

The second review (Harlen, 2004), which synthesized 23 studies, conducted mostly in England and the United States, involved students between the ages of 4 and 18. Twenty studies involved embedding summative assessment in regular classroom activities (i.e., portfolios and projects), and eight were either set externally or set by the teacher to external criteria. The review was focused on examining research evidence to learn more about a range of benefits often attributed to teachers' CA practices including rich understandings of student achievement spanning various contexts and outcomes, the capacity to prevent the negative impacts of standardized tests on student motivation to learn, and teacher autonomy in pursuit of learning goals via methods tailored to their particular students. The review also focused on the influence of teachers' summative assessments practices on their relationships with students, their workload, and difficulties with reliability and quality. The main findings considered two outcomes for the use of assessment for summative purposes by teachers: (1) impact on students and (2) impact on teachers and the curriculum.

Impact on Students

When teachers use summative assessments for external purposes like certification for vocational qualifications, selection for employment or further education, and monitoring the school's accountability or gauging the school's performance, students benefit from receiving better descriptions and examples that help them understand the assessment criteria and what is expected of them. Older students respond positively to teachers' summative assessment of their coursework, find the work motivating, and are able to learn during the assessment process. The impact of external uses of summative assessment on students depends on the high-stakes use of the results and whether teachers orient toward improving the quality of students' learning or maximizing students' scores.

When teachers use summative assessments for internal purposes like regular grading for record keeping, informing decisions about choices within the school, and reporting to parents and students, nonjudgmental feedback motivates students for further effort. In the same vein, using grades as rewards and punishments both decreases student motivation to learn and harms the learning itself. And the way teachers present their CA activities may affect their students' orientation to learning goals or performance goals.

Impact on Teachers and the Curriculum

Teachers differ in their response to their role as assessors and the approach they take to interpreting external assessment criteria. Teachers who favor firm adherence to external criteria tend to be less concerned with students as individuals. When teacher assessment is subjected to close external control, teachers can be hindered from gaining detailed knowledge of their students.

When teachers create assessments for internal purposes, they need opportunities to share and develop their understanding of assessment procedures within their buildings and across schools. Teachers benefit from being exposed to assessment strategies that require students to think more deeply. Employing these strategies promotes changes in teaching that extend the range of students' learning experiences. These new assessment practices are more likely to have a positive impact on teaching when teachers recognize ways that the strategies help them learn more about their students and develop more sophisticated understandings of curricular goals. Of particular importance is the role that shared assessment criteria play in the classroom. When present, these criteria exert a positive influence on students and teaching. Without shared criteria, however, there is little positive impact on teaching and a potential negative impact on students. Finally, high stakes use of tests can influence teachers' internal uses of CA

by reducing those assessments to routine tasks and restricting students' opportunities for learning from the assessments.

Review of Recent Research on Classroom Summative Assessment Practices

What follows is a review of the research on summative assessments practices in classrooms published from 1999 to 2011 and gathered from an Education Resources Information Center (ERIC) search on summative assessments. Studies that were featured in the Harlen and Crick (2002) or the Harlen (2004) reviews were removed. The resulting group of 16 studies investigated summative assessment practices in relation to teachers and teaching and/or students, student learning, and achievement. A comparison of the research aims across the studies resulted in three broad themes: (1) the classroom assessment (CA) environment and student motivation, (2) teachers' assessment practices and skills, and (3) teachers' judgments of student achievement. Table 14.1, organized by theme, presents an overview of the studies.

Theme One: Students' Perceptions of the Classroom Assessment Environment Impact Student Motivation to Learn

Understanding student perceptions of the CA environment and their relationship to student motivational factors was the common aim of four studies (Alkharusi, 2008; Brookhart & Bronowicz, 2003; Brookhart, & Durkin, 2003; Brookhart, Walsh, & Zientarski, 2006). Studies in this group examined teacher assessment practices from the students' point of view using student interviews, questionnaires, and observations. Findings noted both assessment environments and student perceptions of CAs purposes influence students' goals, effort, and feelings of self-efficacy.

As Brookhart and Durkin (2003) noted, even though high profile, large-scale assessments tend to be more carefully studied and better funded, the bulk of what students experience in regard to assessment happens during regular and frequent CAs. Investigations in this theme build on Brookhart's (1997) theoretical model that synthesized CA literature, social cognitive theories of learning, and motivational constructs. The model describes the CA environment as a dynamic context, continuously experienced by students, as their teachers communicate assessment purposes, assign assessment tasks, create success criteria, provide feedback, and monitor student outcomes. These interwoven assessment events communicate what is valued, establish the culture of the classroom, and have a significant influence on students' motivation and achievement goals (Ames, 1992; Brookhart, 1997; Harlen & Crick, 2003).

Teachers' Teaching Experience and Assessment Practices Interact With Students' Characteristics to Influence Students' Achievement Goals

Alkharusi (2008) investigated the influence of CA practices on student motivation. Focusing on a common argument that alternative assessments are more intrinsically motivating than traditional assessments (e.g., Shepard, 2000), the study explored the CA culture of science classes in Muscat public schools in Oman. Participants included 1,636 ninth-grade students (735 male, 901 females) and their 83 science teachers (37 males, 46 females). The teachers averaged 5.2 years of teaching ranging from 1 to 13.5 years of experience. Data came from teacher and student questionnaires. Students indicated their perceptions of the CA environment, their achievement goals, and self-efficacy on a 4-point Likert scale. Teachers rated their frequency of use of various assessment practices on a 5-point Likert scale. Using hierarchical linear models to examine variations present in achievement goals, the study suggests that general principles of CA and achievement goal theory can apply to both U.S. and Oman cultures. Teachers became more aware of the "detrimental effects of classroom assessments that emphasize the importance of grades rather than learning and [focused] on public rather than private evaluation and recognition practices in student achievement motivation" (Alkharusi, 2008, p. 262). Furthermore, the aggregate data suggest that the people and actions around them influence students. Specifically, students are more likely to adopt performance goals such as doing better than others rather than mastery goals of learning more, when assessment environments place value on grades. Students' collective experiences regarding the assessment climate influenced patterns of individual student achievement motivation.

Study	*Research Aim*	*Participants*	*Method*	*Summary of Findings*
Theme One: Classroom Practices and Student Motivation				
Alkharusi (2008)	• Examine the effects of CA practices on students' achievement goals.	• 1,636 ninth-grade students (735 males, 901 females) • 83 science teachers from Muscat public schools in Oman (37 males, 46 females)	Survey	• Both individual student characteristics and perceptions and group characteristics and perceptions influence and explain student mastery goals.
Brookhart & Bronowicz (2003)	• Examine students' perceptions of CAs in relation to assignment interest and importance, student self-efficacy for the task and goal orientation behind their effort.	• Seven teachers (five female, two male) from four schools in Western, Pennsylvania (two elementary, two middle, and two high schools) • 161 students from seven different classrooms in four different schools (63 elementary/middle, 98 high school)	Multiple case analysis	• What matters most to a student affects the student's approach to assessment. • There is a developmental progression in student ability to articulate what it means to succeed in school.
Brookhart & Durkin (2003)	• Describe a variety of CA events in high school social studies classes.	• 1 teacher researcher • 96 students from a large urban high school in the United States	Case study	• The design of the CA, process for completing it, and how much time it takes may affect student motivation and effort. • Performance assessments tap both internal and external sources of motivation.
Brookhart, Walsh, & Zientarski (2006)	• Examine motivation and effort patterns associated with achievement in middle school science and social studies.	• Four teachers (two science, two social studies) from a suburban middle school • 223 eighth-grade students from a suburban Pennsylvania middle school	Field study	• CAs differ on how they are handled and how they engage student motivation and effort, and they have a profound effect on student achievement.
Theme Two: Teacher Assessment Practices and Skills				
Black, Harrison, Hodgen, Marshall, & Serret (2010)	• Explore teachers' understanding and practices in their summative assessments.	• 18 teachers (10 mathematics, 8 English) from three schools in Oxfordshire, England	Partially grounded theory	• Teachers' summative practices were not consistent with their beliefs about validity. • Teacher critiques of their own understandings of validity fostered a critical view of their existing practice.

Study	*Research Aim*	*Participants*	*Method*	*Summary of Findings*
McKinney, Chappell, Berry, & Hickman (2009)	• Investigate pedagogical and instructional mathematical skills of teachers in high-poverty elementary schools.	• 99 teachers from high-poverty schools	Survey	• Teachers rely heavily on teacher-made tests to assess mathematics. • Only a small percentage use alternative assessment strategies.
McMillan (2001)	• Describe assessment and grading practices of secondary teachers.	• 1,483 teachers from 53 middle and high schools in urban/metropolitan Virginia	Survey	• Teachers differentiate cognitive levels of assessments as either higher-order thinking or recall. • Higher ability students receive more assessments that are motivating and engaging while lower ability students receive assessments emphasizing rote learning, extra credit, and less emphasis on academic achievement. • English teachers place more emphasis on constructed-response (CR) items and higher order thinking.
McMillan (2003)	• Determine relationships between teacher self-reported instructional and CA practices and scores on a state high-stakes test.	• 79 fifth-grade teachers from 29 K–5 suburban elementary schools	Survey	• English/language arts teachers used objective tests much more frequently than essay, informal, performance, authentic, or portfolio. • Higher usage of essays in math and English was related to higher objective test scores.
McMillan (2005)	• Investigate relationships between teachers' receipt of high-stakes test results and subsequent changes in instructional and CA practices in the following year.	• 722 teachers from seven Richmond, Virginia, school districts	Survey	• Teachers reported making significant changes to their assessment practices as a result of high-stakes test scores. • Teachers reported placing more emphasis on formative assessments.
McMillan & Lawson (2001)	• Investigate secondary science teachers' grading and assessment practices.	• 213 high school science teachers from urban, suburban and rural schools	Survey	• Secondary science teachers used four assessment types: (1) CR, (2) tests created by or (3) supplied to the teacher, and (4) major examinations. • Teachers tended to use self-made tests, assess as much recall as understanding, use more performance assessments with higher ability students, and assess more recall of knowledge with low ability students.

(Continued)

Table 14.1 (Continued)

Study	*Research Aim*	*Participants*	*Method*	*Summary of Findings*
McMillan & Nash (2000)	• Examine the reasons teachers give for their assessment and grading practices and the factors that influence their reasoning.	• 24 elementary and secondary teachers	Interview	• Tension exists between internal beliefs and values teachers hold regarding effective assessment and realities of their classroom environments and external factors imposed upon them.
Rieg (2007)	• Investigate perceptions of junior high teachers and students at risk of school failure on the effectiveness and use of various CAs.	• 32 teachers from three junior high schools in Pennsylvania • 119 students identified by teachers as being at risk (72 at risk of failing two or more subjects; 20 who also had 10% or greater absenteeism; 27 at risk of dropping out) • 329 students not considered at risk	Survey	• Teachers do not use assessment strategies in their practice that they believe to be effective. • There is a significant difference between what the students at risk felt were effective assessment strategies and the strategies they perceived their teachers actually use. • Students at risk rated 82% of the assessment strategies as more effective than teachers' ratings. • Teachers perceived using certain assessment strategies much more frequently than students perceived that their teachers used them.
Zhang, & Burry-Stock (2003)	• Investigate teachers' assessment practices and perceived skills.	• 297 teachers in two school districts in southeastern United States	Self-reports survey	• As grade level increases, teachers rely more on objective techniques over performance assessments and show an increased concern for assessment quality. • Knowledge in measurement and testing has a significant impact on teachers' self-perceived assessment skills regardless of teaching experience.
Theme Three: Teacher Judgments of Student Achievement				
Kilday, Kinzie, Mashburn, & Whittaker (2011)	• Examine concurrent validity of teachers' judgments of students' math abilities in preschool.	• 33 pre–K teachers in Virginia public school classrooms • 318 students identified as being in at-risk conditions	Hierarchical linear modeling	• Teachers misestimate preschool students' abilities in math both in number sense and in geometry and measurement.

Study	*Research Aim*	*Participants*	*Method*	*Summary of Findings*
Martínez, Stecher, & Borko (2009)	• Investigate teacher judgments of student achievement compared to student standardized test scores to learn if CA practices moderate the relationship between the two.	• 10,700 third-grade students • 8,600 fifth-grade students • Teacher reports of use of standardized test scores and their use of standards for evaluating different students	Unconditional hierarchical linear model	• Teachers judged student achievement in relation to the population in their schools thereby circumventing criterion referencing. • Teachers based evaluations on student needs or abilities. • Gaps in performance of students with disabilities were more pronounced on teacher ratings than standardized test scores. • Teacher judgments incorporate a broader set of dimensions of performance than standardized tests and give more comprehensive picture of student achievement but are susceptible to various sources of measurement error and bias.
Wyatt-Smith, Klenowski, & Gunn (2010)	• Investigate teacher judgment, the utility of stated standards to inform judgment, and the social practice moderation.	• 15 teachers (10 primary and 5 secondary) involved in an assessment communities in Queensland, Australia	Analysis of recordings of talk and conversations	• Common assessment materials do not necessarily lead to common practice or shared understandings. • Teachers tended to view criteria as a guide and perceived it as self-limiting to adhere rigidly to the criteria. • Unstated considerations including perceived value of the benefit of the doubt are included in the judgment making process. • Teachers indicated applying unstated standards they carry around in their head and perceived to have in common to reach an agreement on evaluating ability. • Teachers were challenged practically and conceptually when moving between explicit and tacit knowledge regarding their judgments.

Table 14.1 Overview of 1999 to 2011 Studies on Classroom Summative Assessment

NOTE: CA = classroom assessment.

Student Perceptions of Self-Efficacy May Encourage Students to Consider Classroom Assessment as an Important Part of Learning

Brookhart and colleagues (Brookhart & Bronowicz, 2003; Brookhart & Durkin, 2003; Brookhart et al., 2006) authored the three remaining studies in this theme. The studies reported evidence of CAs and related student perceptions "in their habitats" (Brookhart et al., 2006, p. 163) using classroom observations, artifacts from actual assessment events, and interviews with students and teachers. The three studies yielded the following findings:

- What matters most to a student affects how that student approaches an academic assessment (Brookhart & Bronowicz, 2003).
- There may be a developmental progression in students' ability to articulate what it means to succeed in school (Brookhart & Bronowicz, 2003).
- The CA design, the process for completing it, and the amount of time the assessment takes may influence student motivation and perceptions of effort (Brookhart & Durkin, 2003).
- Teachers can stimulate both mastery and performance goals by designing and using interesting and relevant performance assessments in their classrooms (Brookhart & Durkin, 2003).
- CA environments tend to be more clearly defined by perceptions of the importance and value of assessments coupled with mastery goal orientations (Brookhart et al., 2006).

Summary of Theme One

Taken together, the four studies in this theme present evidence of the profound effects that the CA environment has on student motivation to learn. That motivation is influenced by factors that lie outside the teacher's control—an individual student's interests and needs and students' abilities across grades and developmental levels. What teachers test and how they test over time, however, creates a unique classroom climate that either fuels motivation to learn or derails it. These CA practices are more often than not directly under the teacher's control. Further explorations of student perceptions of self-efficacy in relation to the CA environment may help educators understand the factors that encourage students to study more, try harder, or consider CA as an important part of learning.

Theme Two: Teachers' Summative Assessment Practices and Skills Impact Teacher Effectiveness and Student Achievement

Nine studies investigated summative assessment practices of classroom teachers in relation to seven factors: (1) validity in teachers' summative assessments (Black, Harrison, Hodgen, Marshall, & Serret, 2010), (2) summative assessments in mathematics in urban schools (McKinney, Chappell, Berry, & Hickman, 2009), (3) assessment and grading in secondary classrooms (McMillan, 2001; McMillan, & Lawson, 2001), (4) how teachers' assessment practices relate to and are influenced by scores on high-stakes tests (McMillan, 2003, 2005), (5) the reasons teachers give for their assessment practices (McMillan & Nash, 2000), (6) how teachers' perceptions of assessment practices relate to the perceptions of students at risk of school failure, and (7) relationships between actual assessment practices and teachers' perceived assessment skills (Zhang & Burry-Stock, 2003).

Research Through Professional Development Intervention

Black et al. (2010) implemented the King's-Oxfordshire Summative Assessment Project (KOSAP) to examine and then improve the quality of teachers' summative assessments. Their study examined teachers' understandings of validity and the ways teachers explain and develop that understanding as they learn to audit and improve their existing practices (p. 216). The 35-month project (March 2005 through November 2007) involved 18 teachers from three schools (10 mathematics teachers and 8 English teachers) who taught Grade 8 students (ages 12 to 13). In the first year, teachers were asked to analyze the validity of their assessment practices and create student portfolios that included basic assessment evidence. Working together first in their schools and then across schools, teachers negotiated the portfolio's content, designed common

assessment tasks, determined the need for unique assessments for specific purposes, and established procedures for intra- and inter-school moderation. The moderation process occurred as teachers agreed to communal summative assessment standards and grappled with the disparities of their own judgments and those of their colleagues. Data sources included classroom observations of summative assessment events, records of in-school and inter-school moderation meetings, evidence of summative assessments submitted for moderation, and teachers' reflective diaries.

The study revealed the inconsistency between teachers' beliefs about validity and their summative practices; assessment purposes rarely matched assessment practices. Teachers debated assessment validity and their understanding of validity by investigating three issues: (1) the role assessment plays in their judgments of student achievement, (2) the influence these judgments have on learning experiences in their classrooms, and (3) how they deal with the pressure of sharing assessment information with various stakeholders.

While the project impacted teachers' assessment beliefs and practices, the researchers caution that improved assessment competence and skills require sustained commitment over several years. They suggested that interventions should begin with teachers auditing their existing practices, move to engaging communities of teachers in reflection on their individual and shared assessment literacy, and proceed to teachers working together to improve their underlying beliefs and assumptions regarding summative assessment (Black et al., 2010).

Summative Assessments in Mathematics Can Contribute to a "Pedagogy of Poverty"

Historically, traditional and routine instruction and assessment practices dominate mathematics education in urban schools (Hiebert, 2003; Van De Walle, 2006) to produce what Haberman (1991, 2005) framed as the "pedagogy of poverty." McKinney et al. (2009) situated their study in high-poverty schools to investigate current instructional practices in mathematics and compare them to recommendations made by the National Council of Teachers of Mathematics (NCTM) (2000).

They examined practices of 99 elementary teachers from high-poverty schools who attended an NCTM conference and volunteered to complete the *Mathematics Instructional Practices and Assessment Instrument* during the conference. Using a 43-item survey that described effective mathematics instruction (33 indicators) and effective assessment practices (10 indicators), respondents indicated which practices they used and how frequently they used them. Participants were also asked to write in any practices not included in the survey.

The majority of respondents indicated a heavy reliance on traditional teacher-made tests. This finding is in direct opposition to NCTM (2000) principles that encourage its members to match their assessment practices to their CA purpose; be mindful of the ways CA can be used to enhance student learning; and employ alternative strategies like student self-assessments, portfolios, interviews and conferences, analysis of error patterns, and authentic assessments.

As a result of their investigation, McKinney et al. (2009) reported that little had changed in high-poverty mathematics classrooms. Although NCTM encourages its members to employ alternative approaches that allow student inquiry and a concentration on problem solving and reasoning skills, members failed to use them to improve the mathematics success of urban high-poverty students. Only a small number of respondents reported using alternative approaches to mathematics assessment, and even those teachers admitted to using the practices infrequently.

The Influence of High-Stakes Tests on Summative Assessment Practices

Two studies by McMillan (2003, 2005) examined the relationships between high-stakes tests and CA practices. McMillan (2003) warranted the purpose of his first study by citing the lack of empirical evidence about high-stakes testing that relates instructional and CA practices to actual test scores (p. 5). He investigated 70 fifth-grade English and language arts teachers from 29 K–5 suburban elementary schools. The study employed a survey to collect teachers' self-reports of instructional and CA practices. He used average mathematics and reading test scale scores of students in each class as dependent variables and a measure of aptitude as a covariate.

Despite the limitation inherent in self-report data that are not substantiated by classroom observations or artifacts, the findings reveal a positive correlation between the use of essay tests in mathematics and English and higher objective test scores (McMillan, 2003, p. 9). Even given the correlational nature of the findings, the results suggested that essay tests might be a promising CA approach for raising high-stakes test results. This is especially true since the English/language arts teachers in the study reported using objective tests more frequently than essay, performance, authentic, or portfolio assessments.

McMillan's second study (2005), based on previous research (Shepard, 2000) suggesting that tests emphasizing low-level learning influenced more low-level learning practices in classrooms, investigated relationships between teachers' receipt of their students' high-stakes test score results and their revised instructional and CA practices in the following year. McMillan analyzed written survey data from 722 elementary, middle school, and high school teachers from seven Richmond, Virginia, school districts.

Findings showed that teachers believed they had made significant changes to their assessment practices as a direct result of receiving high-stakes test scores (McMillan, 2005, p. 11). Additionally, the teachers reported an increased use of formative assessments, indicating they were more inclined to use assessment data to inform their teaching. Even though changes occurred more often at the elementary level, secondary English teachers were slightly more likely to change their practices than teachers of other subjects. And more secondary social studies teachers seemed to be influenced in their content area practices by the nature of the high-stakes tests since these tests focused on simple knowledge and understanding.

Assessment and Grading Practices in Secondary Classrooms

Most studies examining assessment and grading practices in secondary classrooms use limited sample sizes (ranging from 24 to 150 participants), making it difficult to isolate grade level and subject matter differences and trends (McMillan, 2001, p. 21). In response to this condition, McMillan (2001) and McMillan and Lawson (2001) intentionally used larger participant samples to examine the relationship between assessment and grading in secondary education.

McMillan (2001) examined the practices of 1,438 classroom teachers (Grades 6 through 12) in 53 schools from seven urban/metropolitan school districts in Virginia across a range of content (science, social studies, mathematics, and English). Teachers responded to a questionnaire of closed-form items to indicate the extent to which they emphasized different grading and assessment practices. The questionnaire contained 34 items in three categories (19 items assessed factors teachers used to determine grades, 11 items assessed different types of assessments, and 4 items assessed the cognitive level of the assessments). Three factor analyses reduced the items to fewer components to analyze the relationship among assessment and grading practices and grade level, subject matter, and ability level of the class.

Results indicated an overall tendency for most secondary teachers to differentiate the cognitive level of their assessments into two categories, higher-order thinking, and recall knowledge, with higher-order thinking emphasized more than recall. Analyses of student ability levels and subject matter revealed that class ability level to be a significant variable related to assessment and grading. McMillan (2001) concluded that higher ability students may "experience an assessment environment that is motivating and engaging, because of the types of assessments and cognitive levels of assessments . . . [while] low-ability students [experience] . . . assessment and grading practices that appear to emphasize rote learning" (p. 31).

English teachers differed most from other subject areas when considering types of assessments. These teachers emphasized higher-order thinking more than science and social studies teachers and placed more emphasis on constructed-response (CR) assessments, teacher-developed assessments, and major exams and less reliance on recall items, objective assessments, and quizzes (McMillan, 2001, p. 31). Since teacher reports of their practices were associated to their actions within a specific class and content, McMillan (2001) suggested that future research take subject matter into consideration when examining CA practices since they are "inexorably integrated with instruction and goals for student learning" (p. 32).

McMillan and Lawson (2001) used the survey instrument and data analyses from McMillan's

2001 study to investigate grading and assessment practices of 213 secondary science teachers from urban, suburban, and rural schools. Their findings indicate that though secondary science teachers tended to use teacher-designed, CR assessments, they relied most heavily on objective assessments and emphasized the recall of information nearly as much as they assessed students' understanding. Similar to McMillan's 2001 findings, patterns of differences related to the ability level of the class. Higher-ability students were advantaged by CA environments where teachers used more performance assessments and emphasized higher cognitive levels.

Reasons Teachers Give for Their Assessment and Grading Practices

To better understand the factors that influence teachers' CA and grading, McMillan and Nash (2000) examined those factors in relation to the reasons teachers give for their decisions. They investigated assessment reasoning and decision making of 24 elementary and secondary teachers selected from a pool of 200 volunteers. Teachers were interviewed in their schools during individual sessions that lasted between 45 to 60 minutes. The four-member research team tape-recorded 20 of the interviews and took notes during and after all interviews. Data were coded according to both emerging and preestablished topics identified in the interview guide. The research team organized the coding into five pervasive themes that explained the data and conducted individual case studies for 20 of the 24 teachers adding 10 new categories and one more pervasive theme. The final six themes formed an explanatory model for how and why teachers decided to use specific assessment and grading practices that included the following: (1) teacher beliefs and values, (2) classroom realities, (3) external factors, (4) teacher decision-making rationale, (5) assessment practices, and (6) grading practices. The model illustrated the tension between teachers' internal beliefs and values and the realities of their classrooms along with other mitigating external factors (McMillan & Nash, 2000, p. 9).

The analysis of the reasoning behind teachers' idiosyncratic assessment practices prompted McMillan and Nash (2000) to conclude that the constant tension teachers experience between what they believe about effective CA and the realities of their classrooms, along with pressures from external factors, cause teachers to view assessment as a fluid set of principles that changes each year. Teachers saw assessment and grading as a largely private matter rarely discussed with other teachers, felt most comfortable constructing their own CAs, and often used preassessments to guide their instruction. They reported that learning was best assessed through multiple assessments and that their thinking about how assessments enhance student learning heavily influenced their classroom decisions. Teachers readily admitted that they *pulled for* their students and often used practices that helped them succeed. In fact, their desire to see students succeed was so strong that it prompted the researchers to question whether that desire "promoted assessment practices where students could obtain good grades without really knowing the content or being able to demonstrate the skill" (McMillan & Nash, 2000, p. 36).

Teachers' Perceptions of Their Classroom Assessment Practices and Skills

Teachers routinely use a variety of assessment practices despite being inadequately trained in how to design and use them effectively (Hills, 1991). Two studies in this review investigated this argument by examining teachers' self-perceived assessment skills. In the first study (Rieg, 2007), assessment strategies that teachers perceived to be effective and useful for students who were at risk were compared to the students' view of those same strategies. The second study (Zhang & Barry-Stock, 2003) compared teachers' self-perceived skills with their actual CA practices. A description of each study follows.

Rieg (2007) surveyed 32 teachers from three junior high schools in Pennsylvania. The teachers taught various subjects including language arts, mathematics, social studies, and science. Rieg designed and used two survey instruments (one for teachers and one for students) containing 28 items informed by the literature on students at risk, assessment, grades and motivation, and middle grade students (p. 216). Teachers were asked to rate the effectiveness of the strategies included on the survey and then indicate the frequency with which they used each strategy in their classrooms. She also surveyed 119 students classified as at risk: 72 were at risk of failing two or more subjects, 20 also had 10% or greater

absenteeism, and 27 were at risk of dropping out of school. In addition, surveys were given to 329 students who were not considered to be at risk. Surveys were read aloud to all students to eliminate limitations of individual student reading difficulties that might have interfered with the results.

There were significant differences between teacher and student perceptions of the assessment strategies that were effective and in frequent use. Teachers reported not using many of the assessments and assessment-related strategies that they perceived as effective. Students reported that their teachers rarely used the strategies they felt to be helpful. These strategies included providing in-class time to prepare for assessments, giving a detailed review of what would be on a test, supplying rubrics or checklists before a performance assessment, and furnishing a study guide to help prepare for tests (p. 220). There was a positive mean difference on 23 (82%) of the strategies that the students perceived to be more effective than their teacher, and there was a significant difference on seven (25%) items with teachers' perception of use being greater than the students' perception of the teacher's use of those strategies. Overall, Rieg reported statistically significant differences between the perceptions of students at risk on the helpfulness and use of 26 (93%) of the 28 survey items.

Zhang and Burry-Stock (2003) also examined teachers' perceptions of CA practices to learn more about teachers' assessment skills. Their investigation was framed by the *Standards for Teacher Competence in Educational Assessment of Students* (American Federation of Teachers [AFT], National Council on Measurement in Education [NCME], & National Education Association [NEA], 1990). They administered the *Assessment Practices Inventory* (API) (Zhang & Barry-Stock, 1994) to 297 teachers in two southeastern U.S. school districts. Factor analytical technique was applied to study the relationship between the constructs of assessment practices and self-perceived assessment skills on the self-report survey.

Teachers' assessment practices differed by teaching levels with a general difference between elementary and secondary teachers in terms of assessment methods used and teachers' concerns for assessment quality. Secondary teachers relied more heavily on paper–pencil tests and had greater concern for assessment quality. Elementary teachers reported greater reliance on performance assessments. In addition to variance by grade levels, teachers' assessment practices differed across content areas. This finding prompted a call for increased assessment training at the preservice and in-service levels that is specifically linked to effective instructional strategies for particular areas of content and grade levels. Knowledge in measurement and testing had a significant impact on teachers' perceptions of their CA skills regardless of teaching experience. This impact strongly influenced teachers' ability to interpret standardized test scores, revise teacher-made tests, modify instruction based on assessment feedback, use performance assessments, and communicate assessment results (p. 335). In light of this, the researchers called for increased university coursework in tests and measurement as a way to increase teachers' CA expertise.

Summary of Theme Two

The nine studies in this theme reveal tensions and challenges faced by classroom teachers as they compare their summative assessment practices with their own beliefs about effective summative assessments. There were significant discrepancies between teacher perceptions of effective summative assessment practices and their self-reports of their actual classroom practices (Black et al., 2010; McKinney et al., 2009; McMillan & Nash, 2000; Rieg, 2007). Secondary teachers reported a general trend toward objective tests over alternative assessments (McMillan, 2001; McMillan & Lawson, 2001) even though higher usage of essays in mathematics and English was related to higher objective test scores (McMillan & Lawson, 2001). These discrepancies might be explained in part by the influence of high-stakes testing on the choices teachers make based on their changing views of the essential purposes for summarizing student achievement (McMillan, 2003, 2005). Another influence may lie in the level of assessment knowledge that teachers possess and the grade levels that they teach. This tendency may be partially attributed to the teachers' perceived assessment knowledge—a factor found to exert more influence on a teacher's assessment practices than the teacher's actual teaching experience (Zhang & Burry-Stock, 2003).

Theme Three: Many Factors Impact the Accuracy of Teachers' Judgments of Student Achievement.

The final theme includes four studies (Kilday, Kinzie, Mashburn, & Whittaker, 2011; Martínez, Stecher, & Borko, 2009; McMillan, 2001; Wyatt-Smith, Klenowski, & Gunn, 2010) that examine the validity of teachers' judgments of student achievement and the dimensions they consider when making those judgments. Two of the four studies (Kilday et al., 2011; Wyatt-Smith et al., 2010) compared teacher judgments of student achievement to results from standardized test scores to investigate how teachers understand and use assessment criteria. Each study is discussed in turn.

Misestimates of Student Achievement Stem From Characteristics Inherent to the Teacher

Kilday et al. (2011) used hierarchical linear modeling to examine the concurrent validity of teachers' judgments of students' mathematics abilities in preschool. Data from an indirect rating scale assessment and the children's performance on two direct assessments of their number sense, geometry, and measurement skills were used to gauge teachers' judgments of preschool children's mathematics skills. Thirty-three teachers enrolled in a field study of a curriculum designed to enhance students' knowledge of mathematics and science participated in the study. Approximately 10 students in each teacher's class were assessed resulting in a sample of 313 students who exhibited one or more established risk factors. Each teacher rated the mathematics skills of his or her 10 students using a modified version of the Academic *Rating Scale* (ARS) for mathematics, which was developed by the Early Childhood Longitudinal Study—Kindergarten Class of 1998–99 (ECLS-K).

The teachers tended to misestimate preschool students' abilities in number sense, as well as in geometry and measurement. "Approximately 40% of the variation in teachers' ratings of students' mathematics skills stem[med] from characteristics inherent to the teacher and not the skills of the child" (Kilday et al., 2011, p. 7). The researchers attributed these findings to the inherently subjective nature of the rating scales and the amount of domain variance at the preschool level. Both factors can systematically influence teachers to misestimate the mathematics skills of young children. Based on this explanation, the researchers suggest that early childhood teachers must become more familiar with student learning trajectories in subjects like mathematics.

Teachers Base Their Judgments of Student Performance on a Broader Set of Performance Dimensions Than Standardized Test Scores

Martínez et al. (2009) used data from third- and fifth-grade samples of the *Early Childhood Longitudinal Survey* (ECLS) to investigate teacher judgments of student achievement in mathematics. The data came from the follow-up studies (ECLS-K) involving children (15,305 third graders and 11,820 fifth graders) who entered kindergarten in 1998. Data included two independent measures of student achievement in reading, mathematics, and science—one based on a standardized test and the other based entirely on the judgments of the student's teacher. Also included were data on the characteristics and practices of the children's teachers and descriptions of the children's families, classrooms, and school environments. Teacher judgments were compared to students' standardized test scores to see if the measures produced a similar picture of student mathematics achievement and if CA practices moderated the relationship between the two measures.

Teachers who participated in the ECLS-K study reported the various types of assessments they frequently used during the year and which factors they deemed important for assessing student performance. They also described the availability and usefulness of individual standardized test scores for guiding instructional decisions and the time they spent preparing for standardized tests. In addition, teachers described whether they held the same standards for evaluating and grading all students in their classroom of if they applied different standards to different students depending on perceived student need or ability (Martínez et al., 2009, p. 85).

In spite of limitations inherent in a data set that may not contain important features of what teachers do in classrooms to assess their students, Martínez et al. (2009) were able to draw conclusions and report significant findings. First, teachers' achievement ratings of

students differed from standardized test scores in important ways.

> [Teachers may] explicitly or instinctively use a school-specific *normative scale* in judging the level of achievement of students in their classrooms . . . [and] may rate students high or low in relation to the achievement levels of other students in the same grade at the school and not necessarily in relation to the descriptors of performance [outlined on test scales in relation to national or state standards]. (Martínez et al., 2009, p. 90)

Second, there were discrepancies between teacher appraisals and standardized test scores in relation to student background characteristics. Teachers' achievement ratings showed a larger disadvantage than standardized test scores for students with disabilities, highlighting the complexity of evaluating students with various challenges. And while standardized tests often disadvantage females, students of minority and low socioeconomic status, and those with low English proficiency, the teachers' judgments appeared less susceptible to bias against traditionally disadvantaged student populations in measuring achievement. The researchers suggested that an alternative explanation might also be the case. Teachers might have deliberately adjusted their ratings upward or their criteria and expectations downward to compensate for disadvantage.

Overall, the findings indicated that teacher judgments incorporated a broader set of performance dimensions than standardized test scores, theoretically providing a more comprehensive picture of student achievement. Some teacher appraisals, however, might be more susceptible to error and bias. Certain teachers may be influenced to appraise student achievement more closely to standardized test scores depending on the specific teacher's background and classroom context. Rating accuracy variance might also be related to the teachers' assessment practices in the classroom that influence their ratings of student achievement. In particular, teachers might not judge student achievement in an absolute manner. They tended to judge achievement in relation to the population of third- and fifth-grade students in their schools thereby "circumventing the criterion referenced . . . scale and adopting a school-specific, norm-referenced scale" (Martínez et al., p. 97).

Teacher Judgment of Student Achievement as a Cognitive and Social Practice

Wyatt-Smith et al. (2010) investigated how stated standards frame teacher judgments and how group moderation (face-to-face and through technology) influences a dynamic process of negotiated meaning. Teacher-based assessment is often characterized as having high validity but questionable reliability (Maxwell, 2001). The study was designed to learn if a strong focus on helping teachers develop a common understanding of standards and recognition of the kinds of performances that demonstrate mastery of those standards might be central to improving reliability.

The study took place in Queensland, Australia, where there is a history of moderated standards-based assessment. "*Moderation* as judgment practices is central . . . [and] involves opportunities for teachers to . . . integrate [their own judgments] with those of other teachers and in so doing share interpretations of criteria and standards" (Wyatt-Smith et al., 2010, p. 61). Both qualitative and quantitative analyses were used to interpret survey data from pre- and post-moderation interviews and recorded conversations from moderation meetings. Fifteen primary and secondary teachers were studied as an assessment community. The teachers first met as a group to raise their awareness of the processes and procedures for moderation. They then met in smaller moderation groups involving three to four teachers.

The teachers received three resources: (1) five marked student work samples representing grades A to F; (2) the *Guide to Making Judgments* that included a matrix of task-specific descriptors and assessable elements that they should consider in their assessments; and (3) annotated student work samples for each question or element of the task and an information sheet of the "reviewing process" (Wyatt-Smith et al., 2010, p. 64). Teachers compared their judgments of each student work sample with each other's ratings to achieve consensus about which grade the work should receive. They cited evidence of the quality of the student work and the application of the assessable elements to justify their individual recommendations. The research team shadowed the teams and recorded their comments and conversations.

Simply providing teachers with assessment materials did not necessarily lead to common practices or shared understandings. Quality

standards, no matter how explicitly described, were seen by teachers as inevitably vague or fuzzy. In fact, the teachers' "unstated considerations including the perceived value of 'the benefit of the doubt' were drawn into the judgment-making process" (Wyatt-Smith et al., 2010, p. 69). Teachers needed standards that worked in concert with exemplars to understand how the features of work they were judging satisfied the requirements of a specific level of performance. This might lessen the tendency for teachers to use what Harlen (2005) called "extra-textual considerations" including nonrelevant aspects of student behaviors, work, or performance in their summative assessments (p. 213).

What's more, teachers seemed to have personal standards and criteria that they carry around in their heads. These personal standards come from experience and allow teachers to reach agreement on student ability and what is "average." These *in the head* criteria and standards were not explicitly stated nor elaborated upon. The teachers simply assumed they all held them in common and regarded them as "characteristic of the experienced teacher" (Wyatt-Smith et al., 2010, p. 70). In the head criteria were also assumed to be shared by teachers for summatively judging the characteristics of an average performance.

A tension point emerged as teachers discussed the *fit* of the assessment tasks that yielded the student work samples and the ways the teachers organized their own curriculum and assessed student achievement in their classrooms. Teachers viewed the assessment tasks as distorting and felt that judgments based on them prevented students from getting what they really deserved. This frustration might be attributed to fact that the criteria sheet forced teachers to leave their comfort zone and removed factors they normally employed when judging achievement. Observational data uncovered the ease with which teachers dismissed the assessment criteria preferring to consider student attributes and allowing those attributes to influence their summative judgments. Teachers routinely discussed the merits of linking their assessments to observed student behaviors such as doing a good job, having ability, being deserving, or making an effort.

Although the teachers struggled with biases and flawed judgments, the study ultimately provides insights into the practical and conceptual challenges teachers face. These trials occur daily as teachers try to reconcile their CA practices and beliefs with standardized or common assessments and expectations. These struggles seem to influence teachers to consider both explicit and tacit knowledge about student achievement.

Summary of Theme Three

An accurate and valid description of student achievement is essential to quality teaching and meaningful learning. This knowledge enables teachers to design effective instruction, provide useful feedback, and design effective assessments to collect evidence of student learning. Teachers appear to benefit from talking with each other and experiencing disequilibrium in regard to the validity of their beliefs and practices (Wyatt-Smith et al., 2010). Understanding how teachers view and use assessment criteria provides insights into how their biases and misunderstandings can cause them to misestimate student achievement (Kilday et al., 2011) and prefer their own in the head criteria when it comes to summarizing student achievement (Wyatt-Smith et al., 2010). Teachers may adopt a school-referenced rather than criterion-referenced orientation to summative assessment thereby muddying their decisions and decreasing the reliability and validity of their judgments (Martínez et al., 2009).

Discussion and Recommended Research

The studies reviewed in this chapter reveal areas of need and areas of promise regarding teachers' summative assessment practices. Although teachers are interpreting more test results and testing more frequently, many teachers are underprepared and insufficiently skilled. This leads to summative judgments that are often inaccurate and unreliable. Yet teachers commonly report positive beliefs about and high levels of confidence in their assessments skills and competence despite evidence to the contrary gathered through observations and teacher self-reports (Black et al., 2010; Rieg, 2007). Many teachers misinterpret student achievement or misestimate students' abilities (Kilday et al., 2011). Frequently teachers arrive at their judgments of student achievement

through idiosyncratic methods and interpret assessment results using flexible criteria. These tendencies allow teachers to pull for students who *deserve* better grades or adjust scores down for students with poor attitudes or behavior (Wyatt-Smith et al., 2010). Traditional and routine practices are common across the board with low-level recall and objective tests figuring prominently in the assessment arsenals of teachers regardless of grade level or subject area. Low-level testing can be found in many classrooms where it impacts both the quality of the learning that happens there and the motivation of the students who must engage in those assessments (McKinney et al., 2009). Sadly, the impact of this practice cuts even deeper in classrooms with poorer or less able students. Yet even when teachers recognize effective assessment practices, they often see the realities of their classroom environments and other external factors imposed on them as prohibitive (McMillan & Nash, 2000).

Still, teachers' summative assessment practices have the potential to positively influence students and teachers (McMillan, 2003), do so without the negative effects associated with external tests and examinations, and produce more comprehensive pictures of student achievement (Martínez et al., 2009). The influence of high-stakes test scores may even prompt some teachers to make significant changes to their CA practices (McMillan, 2005). The assessment environment that teachers create in their classrooms influences student motivational factors like self-efficacy and self-regulation (Alkharusi, 2008; Brookhart & Durkin, 2003). When teachers collaborate with each other and are coached by those with expertise in summative assessment practices, they are more likely to recognize the realities of their assessment competencies and begin to address their assessment needs. They can mediate for each other a more systematic and intentional inquiry process into the quality of their assessments and become mindful how the quality of those assessments influence student learning and achievement (Black et al., 2010). Moreover, knowledge in summative assessment has a significant impact on teachers' self-perceived assessment skills regardless of their teaching experience (Zhang & Burry-Stock, 2003).

Given the nature of the studies reviewed and those mentioned for historical context, several suggestions appear warranted. First, there is a need for research designs that go beyond teachers' self-reports, surveys, and inventories. Evidence from classroom interactions with students, criteria-based examinations of actual teacher-made summative assessments, observations of professional discussions about what comprises achievement, and other strong evidence from teachers' decisions would provide a richer and more comprehensive picture of how teachers summarize student achievement. Only seven studies reviewed (Black et al., 2010; Brookhart & Bronowicz, 2003; Brookhart & Durkin, 2003; Brookhart et al., 2006; McMillan & Nash, 2000; Wyatt-Smith et al., 2010) took this approach.

Second, there is a critical need for research into the impact that principals and central office administrators have on the summative assessment practices of teachers in their buildings and districts. Investigations of the roles administrators play in perpetuating mediocre assessments of achievement or spearheading quality CA practices would add to our understanding. Teachers do not assess in a vacuum, yet a review of the CA literature might lead us to conclude otherwise. We know little about how building- and district-level administrators might lead a culture of high quality summative assessment to promote accurate decisions about what students know and can do. And studies of college and university certification programs for educational leadership are sorely needed to identify programmatic factors and approaches that produce administrators who understand quality summative assessment, can recognize it when they see it, and are able to effectively intervene when they don't.

Finally, university programs continue to graduate teachers who are overconfident and under competent when it comes to summarizing achievement and using assessment information to promote improved student learning. These studies could inform the design of teacher preparation programs that make quality assessment a focal point of effective pedagogy. This would be especially true if researchers go beyond counting the number of assessment courses in particular curriculum to examining what actually happens in those courses to develop assessment literacy and follow the graduates into the field to see if those courses impact actual assessment practices.

References

Alkharusi, H. (2008). Effects of classroom assessment practices on students' achievement goals. *Educational Assessment, 13*(4), 243–266.

American Federation of Teachers, National Council on Measurement in Education, & National Education Association. (1990). *Standards for teacher competence in educational assessment of students.* Washington, DC: National Council on Measurement in Education. (ERIC Document Reproduction Service No. ED 323 186)

Ames, C. (1992). Classrooms: Goals, structures, and student motivation. *Journal of Educational Psychology, 84,* 261–271

Aschbacher, P. (1999). Helping educators to develop and use alternative assessments: Barriers and facilitators. *Educational Policy, 8,* 202–223.

Atkin, J. M., & Coffey, J. (Eds.) (2001). *Everyday assessment in the science classroom.* Arlington, VA: National Science Teachers Association Press.

Baker, R. L., Mednick, B. R., & Hocevar, D. (1991). Utility of scales derived from teacher judgments of adolescent academic performance and psychosocial behavior. *Educational and Psychological Measurement, 51*(2), 271–286.

Baron, J. B. (1991). Strategies for the development of effective performance exercises. *Applied Measurement in Education, 4*(4), 305–318.

Black, P., Harrison, C., Hodgen, J., Marshall, B., & Serret, N. (2010). Validity in teachers' summative assessments. *Assessment in Education Principles, Policy & Practice, 17*(2), 215–232.

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

Boothroyd, R. A., McMorris, R. F., & Pruzek, R. M. (1992, April). *What do teachers know about measurement and how did they find out?* Paper presented at the annual meeting of the Council on Measurement in Education, San Francisco. (ERIC Document Reproduction Service No. ED351309)

Borko, H., Mayfield, V., Marion, S., Flexer, R., & Cumbo, K. (1997). Teachers' developing ideas and practices about mathematics performance assessment: Successes, stumbling blocks, and implications for professional development. *Teaching and Teacher Education, 13,* 259–278.

Brookhart, M. S. (1997). A theoretical framework for the role of classroom assessment in motivating student effort and achievement. *Applied Measurement in Education, 10*(2), 161–180.

Brookhart, S. M., & Bronowicz, D. L. (2003). "I don't like writing. It makes my fingers hurt": Students talk about their classroom assessments. *Assessment in Education, 10*(2), 221–241.

Brookhart, S. M., & Durkin, D. T. (2003). Classroom assessment, student motivation and achievement in high school social studies classes. *Applied Measurement in Education 16*(1), 27–54.

Brookhart, S. M., Walsh, J. M., & Zientarski, W. A. (2006). The dynamics of motivation and effort for classroom assessment in middle school science and social studies. *Applied Measurement in Education, 19*(2), 151–184.

Clarke, M. Madaus, G. F., Horn, C. J., & Ramos, M. A. (2000). Retrospective on educational testing and assessment in the 20th century. *Journal of Curriculum Studies, 32*(2), 159–181.

Darling-Hammond, L. (1995). Equity issues in performance-based assessment. In M. T. Nettles & A. L. Nettles (Eds.), *Equity and excellence in educational testing and assessment* (pp. 89–114). Boston: Kluwer

Falk, B., & Ort, S. (1998). Sitting down to score: Teacher learning through assessment. *Phi Delta Kappan, 80,* 59–64.

Gearhart, M., & Saxe, G. B. (2004). When teachers know what students know: Integrating assessment in elementary mathematics. *Theory Into Practice, 43,* 304–313.

Gittman, E., & Koster, E. (1999, October). *Analysis of ability and achievement scores for students recommended by classroom teachers to a gifted and talented program.* Paper presented at the annual meeting of the Northeastern Educational Research Association, Ellenville, NY.

Goldberg, G. L., & Roswell, B. S. (2000). From perception to practice: The impact of teachers' scoring experience on performance-based instruction and classroom assessment. *Educational Assessment, 6,* 257–290.

Goslin, D. A, (1967). *Teachers and testing.* New York: Russell Sage.

Griswold, P. A. (1993). Beliefs and inferences about grading elicited from student performance sketches. *Educational Assessment, 1*(4), 311–328.

Gullikson, A. R. (1984). Teacher perspectives of their instructional use of tests. *Journal of Educational Research, 77*(4), 244–248.

Haberman, M. (1991). The pedagogy of poverty versus good teaching. *Phi Delta Kappan, 73,* 209–294.

Haberman, M. (2005). *Star teachers: The ideology and best practice of effective teachers of diverse children and youth in poverty.* Houston, TX: Haberman Educational Foundation.

Hall, J. L., & Kleine, P. F. (1992). Educators' perceptions of NRT misuse. *Educational Measurement: Issues and Practice, 11*(2), 18–22.

Harlen, W. (2004). A systematic review of the evidence of the impact on students, teachers and the curriculum of the process of using

assessment by teachers for summative purposes. In *Research Evidence in Education Library.* London: Evidence for Policy and Practice Information and Co-Ordinating Centre, Social Science Research Unit, Institute of Education.

Harlen, W. (2005). Teachers' summative practices and assessment for learning—tensions and synergies. *The Curriculum Journal, 16*(2), 207–223.

Harlen, W., & Crick, R. D. (2002). A systematic review of the impact of summative assessment and tests on students' motivation for learning (EPPI-Centre Review, version 1.1*). In *Research Evidence in Education Library, Issue 1.* London: Evidence for Policy and Practice Information and Co-Ordinating Centre, Social Science Research Unit, Institute of Education.

Harlen, W., & Crick, R. D. (2003). Testing and motivation for learning. *Assessment in Education: Principles, Policy & Practice, 10*, 169–207.

Hiebert, J. (2003). What research says about the NCTM standards. In J. Kilpatrick, W. G. Martin, & D. Schifter (Eds.), *A research companion to principles and standards for school mathematics* (pp. 5–23). Reston, VA: National Council of Teachers of Mathematics.

Hills, J. R. (1991). Apathy concerning grading and testing. *Phi Delta Kappa, 72*(7), 540–545.

Hoge, R. D. (1984). Psychometric properties of teacher-judgment measures of pupil attitudes, classroom behaviors, and achievement levels. *Journal of Special Education, 17,* 401–429.

Hopkins, K. D., George, C. A., & Williams, D. D. (1985). The concurrent validity of standardized achievement tests by content area using teachers' ratings as criteria. *Journal of Educational Measurement, 22,* 177–182.

Impara, J. C., Divine, K. P., Bruce, F. A., Liverman, M. R., & Gay, A. (1991). Does interpretive test score information help teachers? *Educational Measurement: Issues and Practice, 10*(4), 319–320.

Kenny, D. T., & Chekaluk, E. (1993). Early reading performance: A comparison of teacher-based and test-based assessments. *Journal of Learning Disabilities, 26,* 227–236.

Kilday, C. R., Kinzie, M. B., Mashburn, A. J., & Whittaker, J. V. (2011). Accuracy of teacher judgments of preschoolers' math skills. *Journal of Psychoeducational Assessment*, *29*(4) 1–12.

Laguarda, K. G., & Anderson, L. M. (1998). *Partnerships for standards-based professional development: Final report of the evaluation.* Washington, DC: Policy Studies Associates, Inc.

Marso, R. N., & Pigge, F. L. (1988, April). *An analysis of teacher-made tests: Testing practices, cognitive demands, and item construction errors.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA. (ERIC Document Reproduction Service No. ED298174)

Martínez, J. F., & Mastergeorge, A. (2002, April). *Rating performance assessments of students with disabilities: A generalizability study of teacher bias.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

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

Maxwell, G. (2001). *Moderation of assessments in vocational education and training.* Brisbane, Queensland: Department of Employment and Training.

McKinney, S. E., Chappell, S., Berry, R. Q., & Hickman, B. T. (2009). An examination of the instructional practices of mathematics teachers in urban schools. *Preventing School Failure: Alternative Education for Children and Youth, 53*(4), 278–284.

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

McMillan, J. H. (2003). *The relationship between instructional and classroom assessment practices of elementary teachers and students scores on high-stakes tests* (Report). (ERIC Document Reproduction Service No. ED472164)

McMillan, J. H. (2005). *The impact of high-stakes test results on teachers' instructional and classroom practices* (Report). (ERIC Document Reproduction Service No. ED490648)

McMillan, J. H., & Lawson, S. (2001). *Secondary science teachers' classroom assessment and grading practices* (Report). (ERIC Document Reproduction Service No. ED450158)

McMillan, J. H. & Nash, S. (2000). *Teacher classroom assessment and grading practices decision making* (Report). (ERIC Document Reproduction Service No. ED447195)

Meisels, S. J., Bickel, D. D., Nicholson, J., Xue, Y., & Atkins-Burnett, S. (2001). Trusting teachers' judgments: A validity study of a curriculum-embedded performance assessment in kindergarten–Grade 3. *American Educational Research Journal, 38*(1), 73–95.

National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics.* Reston, VA: Author.

National Research Council. (2001). *Inquiry and the National Science Education Standards.* Washington, DC: National Academy Press.

Nolen, S. B., Haladyna, T. M., & Haas, N. S. (1992). Uses and abuses of achievement test scores. *Educational Measurement: Issues and Practice, 11*(2), 9–15.

O'Sullivan, R. G., & Chalnick, M. K. (1991). Measurement-related course work requirements for teacher certification and recertification. *Educational Measurement: Issues and Practice, 10*(1), 17–19.

Parkes, J., & Giron, T. (2006). *Making reliability arguments in classrooms.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Plake, B. S. (1993). Teacher assessment literacy: Teachers' competencies in the educational assessment of students. *Mid-Western Educational Researcher, 6*(1), 21–27.

Rieg, S. A. (2007). Classroom assessment strategies: What do students at-risk and teachers perceive as effective and useful? *Journal of Instructional Psychology, 34*(4), 214–225.

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

Roeder, H. H. (1972). Are today's teachers prepared to use tests? *Peabody Journal of Education, 59,* 239–240.

Sato, M. (2003). Working with teachers in assessment-related professional development. In J. M. Atkin & J. E. Coffey (Eds.), *Everyday assessment in the science classroom* (pp. 109–120). Arlington, VA: National Science Teachers Association Press.

Sharpley, C. F., & Edgar, E. (1986). Teachers' ratings vs. standardized tests: an empirical investigation of agreement between two indices of achievement. *Psychology in the Schools, 23,* 106–111.

Sheingold, K., Heller, J. I., & Paulukonis, S. T. (1995). *Actively seeking evidence: Teacher change through assessment development* (Rep. No. MS-94-04). Princeton, NJ: Educational Testing Service.

Shepard, L.A. (2000). The role of assessment in a learning culture. *Educational Researcher, 29,* 4–14.

Stiggins, R. (1991). Relevant classroom assessment training for teachers. *Educational Measurement: Issues and Practice, 10,* 7–12

Stiggins, R. J. (1999). Are you assessment literate? *The High School Journal, 6*(5), 20–23.

Stiggins, R. J., & Bridgeford, N. J. (1985). The ecology of classroom assessment. *Journal of Educational Measurement, 22,* 271–286.

Stiggins, R. J., Frisbie, R. J., & Griswold, P. A. (1989). Inside high school grading practices: Building a research agenda. *Educational Measurement: Issues and Practice, 8*(2), 5–14.

Tiedemann, J. (2002). Teachers' gender stereotypes as determinants of teacher perceptions in elementary school mathematics. *Educational Studies in Mathematics, 50*(1), 49–62.

Van De Walle, J. (2006). *Raising achievement in secondary mathematics.* Buckingham, UK: Open University Press.

Wilson, S. (2004). Student assessment as an opportunity to learn in and from one's teaching practice. In M. Wilson (Ed.), *Towards coherence between classroom assessment and accountability* (National Society for the Study of Education Yearbook, Vol. 103, Part 2, pp. 264–271). Chicago: University of Chicago Press.

Wilson, M., & Sloane, K. (2000). From principles to practice: An embedded assessment system. *Applied Measurement in Education, 13,* 181–208.

Wise, S. L., Lukin, L. E., & Roos, L. L. (1991). Teacher beliefs about training in testing and measurement. *Journal of Teacher Education, 42*(1), 37–42.

Worthen, B. R., Borg, W. R., & White, K. R. (1993). *Measurement and evaluation in schools.* White Plains, NY: Longman.

Wyatt-Smith, C., Klenowski, V., & Gunn, S. (2010). The centrality of teachers' judgment practice in assessment: A study of standards in moderation. *Assessment in Education: Principle, Policy & Practice, 17*(1), 59–75.

Zhang, Z., & Barry-Stock, J. A. (1994). *Assessment Practices Inventory.* Tuscaloosa: The University of Alabama.

Zhang, Z., & Burry-Stock, J. A. (2003). Classroom practices and teachers' self-perceived assessment skills. *Applied Measurement in Education, 16*(4), 323–342.