Journal of Advanced Academics
jaa
Volume 18 ✣ Number 3 ✣ Spring 2007 ✣ pp. 402–423

Coming to Terms With Classroom Assessment

Bruce B. Frey
University of Kansas

Vicki L. Schmitt
Missouri State University

Well-trained and informed classroom teachers have been introducing modern classroom assessment approaches into their classrooms. Performance-based assessment is now commonplace in many classrooms as more teachers become concerned about the authenticity of their assessments and how assessment information can be used as formative feedback to improve teaching and learning. Although these newer approaches are driven by a motivation to make student assessment data more useful and meaningful than some traditional approaches, and their use, therefore, is a positive development, it is hard to judge the theoretical benefit of these changes or even to begin to systematically explore the nature of teachers' modern classroom assessment practices. This difficulty arises because researchers, advocates, and practitioners have not arrived at a consistent definition of what these terms mean or what these practices look like.

At recent meetings of the American Educational Research Association, classroom assessment researchers bemoaned a perceived lack of consensus as to the correct basic definitions for key classroom assessment terms. Marion (2005), Noonan and Duncan (2005), and Shepard (2005) argued, in different contexts, that one limitation in conducting research about classroom assessment is that certain key terms are used in different ways

Summary

As the field of education moves forward in the area of assessment, researchers have yet to come to a conclusion about definitions of commonly used terms. Without a consensus on the use of fundamental terms, it is difficult to engage in meaningful discourse within the field of assessment, as well as to conduct research on and communicate about best assessment practices. For this article, we reviewed journal articles, position papers, thought pieces, and classroom assessment textbooks, focusing on the definitions of the terms *performance assessment, authentic assessment,* and *formative assessment*. We provide a summary of the literal definitions provided and the components, format, and intentions of each type of assessment. In addition, we underscore the important distinctions made by researchers in the field between performance assessment and authentic assessment. Some researchers suggest performance assessment and authentic assessment are synonymous, and others view performance assessment as a component of authentic assessment. Understanding authentic and performance assessments is important to have a sound theoretical basis for decisions made within the classroom. The purpose and benefits of formative assessment represent another area of discussion within the field of assessment. Formative assessment may be used solely to inform the teacher, or it may be used as a powerful means of providing feedback to students, allowing students to alter their strategies to improve learning. We emphasize important distinctions between the formation of learning and the formation of behaviors or strategies that promote learning. Finally, to avoid confusion, classroom assessment should be classified based on the assessment's intended purpose.

Frey, B. B., & Schmitt, V. L. (2007). Coming to terms with classroom assessment. *Journal of Advanced Academics, 18,* 402–423.

to mean different things. Beyond confused researchers, a greater concern might be for teachers, practicing and preservice, who are advised by the field to consider certain assessment approaches.

Most classroom assessment involves tests that teachers have constructed themselves. It is estimated that 54 teacher-made tests are used in a typical classroom per year (Marso & Pigge, 1988), which results in perhaps billions of unique assessments, yearly, worldwide (Worthen, Borg, & White, 1993). Further, teachers place more weight on their own tests in determining grades and student progress than they do on assessments designed by others, or on other data sources (Boothroyd, McMorris, & Pruzek,1992; Fennessey, 1982; Stiggins & Bridgeford, 1985; Williams, 1991). Teachers have reported that they are confident in their ability to produce good student tests (Oescher & Kirby, 1990; Wise, Lukin & Roos, 1991). However, teachers are not particularly good judges of their own abilities or knowledge in test construction (Boothroyd et al., 1992; Oescher & Kirby, 1990) and teachers' own estimates of ability and actual performance (in test construction) have been found to be *negatively* correlated (Marso & Pigge, 1988).

Many teachers believe that they need strong measurement skills (Boothroyd et al., 1992) and believe that their training was inadequate (Wise et al., 1991). They also report a level of discomfort with the quality of their own tests (Stiggins & Bridgeford, 1985). Most state certification systems and half of all teacher education programs have no assessment course requirement nor do they have an explicit requirement that teachers have received training in assessment (Boothroyd et al., 1992; Trice, 2000; Wise et al., 1991). In addition, teachers have historically received little or no training or support after certification (Herman & Dorr-Bremme, 1984). The formal assessment training teachers do receive often focuses on large-scale test administration and standardized test score interpretation, rather than on the test construction strategies or item-writing rules that teachers need to create their own tests (Stiggins & Bridgeford, 1985).

Although there is limited systematic training in assessment strategies for teachers, the classroom assessment field (i.e.,

researchers and teacher educators) routinely advocates new and improved approaches to assessment that are advertised as more valid than past methods. This was the case for performance-based assessment 2 decades ago, and it is also true for authentic assessment and formative assessment more recently. Teachers and others interested in the application of evidence-based practice who wish to adopt these new approaches, however, receive inconsistent messages as to exactly what these approaches are and how they differ from each other.

Specific instances of where differing use of these terms may prevent the generalization of research to the classroom include questions about:

- whether the category of performance-based assessments should include traditional essay exams. For example, Frey and Schmitt (2005) reported that when assessing students, classroom teachers in Kansas use performance-based assessments about 28% of the time. Their study, however, categorized essay questions as traditional, not performance-based. For those who would define that item format differently, the real frequency of performance-based assessments in Kansas is still a question, and
- whether formative assessment and assessment for learning (a term implying that assessment improves learning) are synonymous concepts, different ways of advocating for the same practices for the same reasons.

For this study, we reviewed journal articles, position papers, thought pieces, and textbooks focusing on the definitions of the terms *performance assessment*, *authentic assessment*, and *formative assessment*. We were interested in the literal definitions, components, format, and intentions of each type of assessment as presented by different authors. The different ways that the terms are understood have implications for the field of classroom assessment research and teacher preparation and training. We present the important distinctions in the field between performance assessment and authentic assessment and the purpose and ben-

efits of assessment that drive the two primary uses of the term formative assessment.

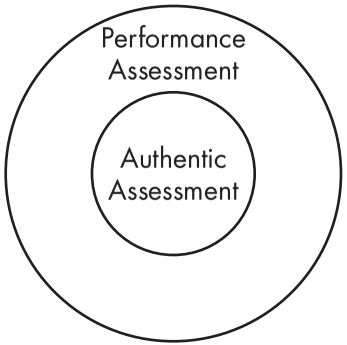
Terms of Classroom Assessment

There are important differences in how performance-based assessment is conceptualized in the literature and in textbooks. A key distinction also exists in what the perceived purpose and benefit of formative assessment is meant to be. We look first at performance assessment and its perceived relationship to authentic assessment. Later we discuss formative assessment.

Performance and Authentic Assessment

A key clarification of the term performance assessment, as it relates to authentic assessment, is needed. In our review, it is clear that two categorical schemas exist as to the relationship between authentic assessment and performance assessment. At the risk of adding to the circular nature of any definitional debate, we offer this fairly well-accepted brief description of authentic assessment: assessments that specifically address real-world applications (e.g., Mertler, 2003). One camp views performance assessment as a category that includes authentic assessment, but does not categorize all performance assessment as authentic (e.g., Mertler, 2003; Meyer, 1992; Oosterhof, 2003; Popham, 2002; Stiggins, 1991). The other view is that, by its nature, performance assessment is authentic (Airasian, 2001; Kubiszyn & Borich, 2003; Mueller, 2005; Taylor & Bobbit-Nolen, 2005). These perspectives are summarized in Figure 1.

In their "short history" of performance assessment published in 1999, Madaus and Dwyer treat authentic assessment as a true synonym for performance assessment, at least as it is used by the field, citing another equivalent phrase as "the 3 P's: performance, portfolios and products" (p. 688). Their treatment of the two terms as equivalent is particularly interesting because their definition of performance assessment is the broadest we have found:



- Authentic assessments are performance assessments, but the inverse is not true (Oosterhof, 2003).
- Many performance-based assessments are also classified as authentic assessments (Mertler, 2003).
- In some instances . . . school tasks rather than real-world tasks may be suitable for performance assessment (Popham, 2002).
- Performance assessment is the use of a performance criteria to measure achievement (Stiggins, 1992).

- Performance assessment measures outcomes in more authentic contexts (Kubiszyn & Borich, 2003).
- As authentic performances, performance assessment seems more relevant to the real world (Taylor & Bobbit-Nolen, 2005).
- Performance assessments may be called authentic assessments. They permit students to show what they can do in real situations (Airasian, 2001).
- Virtually all performance tasks should be authentic in nature or they lose some relevance to the students. This distinction between performance and authentic assessments becomes insignificant and unnecessary (Mueller, 2003).

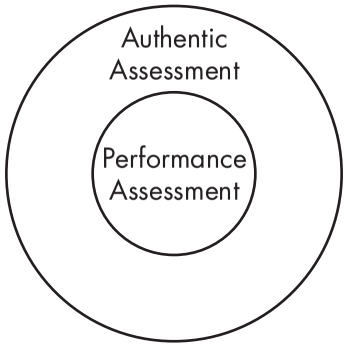


Figure 1. Perspectives on performance assessment.

"Performance assessment requires examinees to construct/supply answers, perform or produce something for evaluation" (p. 690). Conversely, Wiggins, a leader in the field of authentic assessment, stated flatly that "'authentic' (assessment) is not synonymous with 'performance assessment'" (Newman, Brandt & Wiggins, 1998, p. 21), whereas Brandt (Newman, et al., 1998) believed performance assessment is a useful synonym to describe what would otherwise be referred to as authentic assessment.

The practice of performance assessment is much older than the use of a specific term to describe it. The approach is at least as old as the beginnings of the oft-cited (on the first page of the first chapter of many textbooks) Chinese civil examinations more than 2,000 years ago. Madaus and Dwyer (1999) provided a handy year, 210 BCE, as the nominal birth of performance assessment in China. As a type of assessment, the use of written essays and other performance-based demonstrations in the U.S. throughout the 1800s demonstrates that performance assessment was the original approach and was surpassed in popularity with what is now somewhat incorrectly termed traditional assessment starting in about 1914 with the use of multiple-choice testing.

Performance assessment as a philosophical approach regained momentum in the 1980s (Madaus & Dwyer, 1999), but the earliest references to "performance" tasks we could find occurred in the context of discussions about the nature of criterion-referenced measurement (Harris, Alkin & Popham, 1974). Criterion-referenced assessments evaluate students against a defined set of objectives or standards, and unlike norm-referenced assessments, do not compare students to each other. Alkin (1974) described these tests as "usually referenced to a performance objective or a behavioral objective" (p. 3) and typically developed from "well-defined performance domains or objectives" (p. 4). The term *performance* can, of course, be used to simply mean how one scored on a test, as in the sense of "I performed well on the GRE," but the term seems to be used somewhat consistently at the time to mean observable behaviors. For example, in the Alkin article, Glaser and Nitko (1971) were referenced as describing criterion-referenced tests as providing directly interpretable information

related to a domain of tasks to be performed. Also, Harris and Stewart (1971) described criterion-based tests as consisting of a sample of production tasks from a population of performances. These early discussions of criterion-referenced testing as performance-based assessment also suggest the connections between performance assessment and authentic assessment. Nitko (1984), for example, wrote that criterion-referenced tests hope to reveal what kind of behaviors students can demonstrate and one can almost hear the "... in the real world" extension that would come a decade and a half later.

Criterion-based measures are not always, nor even substantially, performance-based, either then nor now. Most of the test theorists at the time emphasized content or knowledge domains as the source of items for these tests (e.g., Baker, 1974, Popham, 1974), but it is worth noting that several tended to use performance tasks and the assessment of skills as illustrative examples when addressing the measurement issues involved in criterion-referenced testing (e.g., Nitko, 1984; Skager, 1974).

The earliest reference to authentic tests was likely made by Archbald and Newman in 1988, in a book critical of standardized testing that sought to promote assessment centered on meaningful real-world problems or tasks. Later, Newman et al. (1998) suggested that assessment is authentic when it measures products or performances that "have meaning or value beyond success in school" (p. 19). According to Newman, assessments that ask questions and pose problems that have "real-world" meaning to students meet one criteria for being authentic intellectual work, but there are two others related to disciplined inquiry that are unrelated to the realism of the assessment tasks. Wiggins (1989) was also an early proponent for the use of the term *authentic* to describe assessment with real-world application. "'Authentic' refers to the situational or contextual realism of the proposed tasks" he emphasized (Newman et al., 1998, p. 20).

Terwilliger (1998) expressed concerns with Wiggins' and others' use of the term, viewing the label of authentic as a thinly veiled criticism of traditional assessment approaches as somehow less authentic or inauthentic. Wiggins' position was essen-

tially that traditional assessment is not inauthentic, it is simply less direct and, probably, less meaningful to students. Wiggins (1993) argued that traditional assessment is not faithful to the domains of performances and contexts that are most important for higher order thinking and learning. As he used the term, authenticity is akin to fidelity.

Bergen (1993) identified three qualities of "good" authentic assessment. Referring to assessment that is both performance and authentic, one criterion is that it is often group-based with each individual contribution required for success. The other two qualities refer to the complexity of the task—it measures many facets simultaneously and it is applied in a way that reflects the complex roles of the real world.

The attribute of directness, the closeness of the connection between the assessment task and the actual real-world task, is a useful one because it allows us to theoretically place an item on a continuum with two anchors—the task is the actual real-world task (at one end) or the task does not represent a real-world at all (at the other end). By the common definition of authentic, performance tasks near the real-world task end of the continuum are authentic and those that are nearer the middle or the other end (only representative of real-world tasks or not "realistic" at all) are not authentic. Gronlund (2003) suggested a similar continuum for determining the "appropriate degree of realism" (p. 124) when designing performance assessment tasks in the classroom.

Meyer (1992) argued that it is assessors who should determine whether a given assessment is authentic, using the criteria that seem most crucial to them. Criteria of authenticity could include, among other aspects, the nature of the stimuli, the complexity of the task, conditions, resources, consequences, and whether the specific tasks or activities are determined by the student or the assessor.

Formative Assessment

Formative assessment's varied descriptions begin with differences in such basic concepts as to why it is even called *forma-*

tive. Some authors suggest that because the feedback from these assessments can help form teacher or student behavior, the formative descriptor is used (e.g., Airasian, 2001; Black & Wiliam, 1998). Although such data certainly can help form teacher and student behavior and that is its strength and, perhaps, its purpose, the term is older than its current popularity as a modern classroom assessment approach and philosophy (Scriven, 1967). Historically, formative assessment was so named to distinguish it from summative assessment. One occurred while learning was still occurring or *form*ing; the other occurred at the end of learning. To be accurate, the term was initially used by Scriven to apply to a program evaluation approach, and was contrasted with summative evaluation. The concept was attached to assessment, apparently, first by Bloom (1968), who saw a relationship between formative assessment and mastery learning.

The important distinction we highlight here is between the formation of learning and the formation of behaviors or strategies that promote learning. The most common use of the term formative assessment that we find today assumes the latter goal. Typically, "assessments become formative when the information is used to adapt teaching and learning to meet students' needs" (Boston, 2002, ¶ 2). This key disagreement in the purpose of formative assessment does exist, however, in the textbooks and scholarly writings used to prepare classroom teachers and centers around whether the feedback produced is for the use of teachers or students or both. Most current textbooks describe the purpose of formative assessment as informing the teacher, and seldom mention providing feedback to students. Many researchers and advocates for formative assessment, on the other hand, argue that its primary benefit is in allowing students to control and improve their own learning (e.g., Stiggins, 2002). Which of these perspectives one adopts has important implications on the format of formative assessment (its frequency, whether it is formal or informal, and whether observation of students counts). Figure 2 summarizes these contrasting perspectives.

The idea that classroom assessment can provide feedback for students so they can affect their own learning or even any refer-

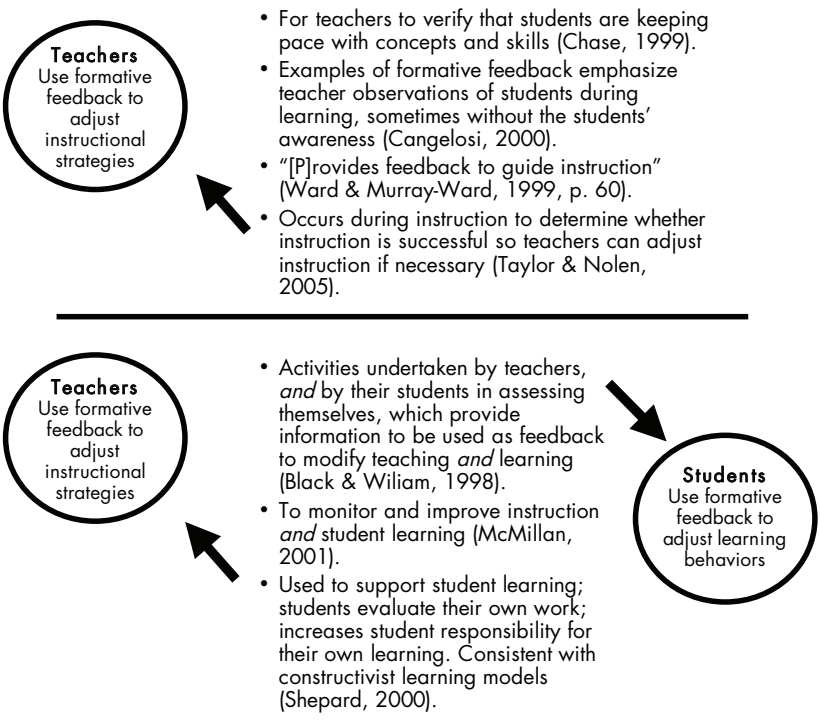


Figure 2. Perspectives on the purpose of formative assessment.

ence to the possibility that quality assessment can directly affect learning is relatively new. A 1974 listing by Skager of the ways that tests provide information and meet information needs did not include any reason among the six given that students might benefit from the information. Although labeled as a classroom management purpose, one benefit of assessment identified is that "the present learning status of the pupils in terms of the objectives" (Skager, 1974, p. 48) can be determined. This is contrasted with terminal learning status, so a distinction between summative and formative assessment might be inferred, and this writing may have been influenced by the formative assessment concept as it took place after Scriven (1967) and Bloom (1968). After Scriven and Bloom's introduction of the term, formative assessment became a category of assessment that occurred during instruction, but the purpose was not to provide feedback to students for the purpose of directly affecting learning. The original

purpose was to evaluate instruction, and then, improve or alter it. Fifteen years later, little had changed as Millman and Greene's (1989) comprehensive survey of test development methods did not include student self-evaluation or self-monitoring as a reason for assessing during instruction.

A more important disagreement exists in the consideration of the purpose of formative assessment. Is it meant to inform the teacher, who can improve instruction, the student, who can alter learning strategies, or both?

Discussion

Bligh (2001) has editorialized about gaps between theory and practice in classroom assessment. He referred to findings that instructors in medical schools held beliefs about assessment that were consistent with evidence-based best practices, but the instructors actual practices were not consistent with their beliefs. He wrote:

> It appears that assessment is an example of a subject where there are two camps: one full of well meaning, earnest teachers and researchers immersed in the language and culture of assessment practice (validity, generalizability, psychometrics are examples of the words they commonly use); the other full of well meaning, earnest teachers facing the day to day practical problems of running assessments. . . . (p. 312)

The disparate use of terms described here may reflect a similar dilemma, a wish by researchers and teacher education faculty to encourage what they view as evidence-based best practice while providing purposefully broadly drawn definitions that include a variety of classroom assessment activities in which teachers can and will engage.

Of course, the use of a term to describe an assessment strategy does not somehow infuse the operationalization of that

strategy with validity or effectiveness. Referring to another common, at least in the context of British education, inexact confusion of terminology, Black and Wiliam (2004) wrote, "The terms classroom assessment and formative assessment are often used synonymously, but . . . the fact that an assessment happens in the classroom . . . says very little about either the nature of the assessment or the functions that it can serve" (p. 183).

Authentic Performance Assessment

It is clear that performance-based assessment is perceived by different researchers and theorists as suggesting slightly, but importantly, different formats, created to solve slightly different problems and useful for slightly different purposes. Distinctions in conceptualizations focus on the measurement objectives, the nature of the student tasks required, and the scoring mechanisms.

Knowing what words mean is critical for researchers, practitioners, and trainers to understand each other. To engage in best practices requires a shared understanding of what different assessment practices are, what they look like, and what the critical components are in order to expect outcomes suggested by theory or empirical research. Take, for example, advocates for performance-based assessment. If they view all performance assessment as authentic assessment, there are certain logical consequences that follow. For instance, any test that assesses skill or ability or, perhaps, any test that uses a subjective scoring rubric (common defining components of performance assessments; e.g., Frey & Schmitt, 2005) may be treated as a celebratory example of real-world assessment. From a validity perspective, if the inferences made from authentic assessments are appropriate because the assessment tasks fairly represent some domain of real-world tasks that students will have to perform, then performance assessments need not provide much validity evidence; they are theoretically valid because they are all authentic.

Assessment for Learning

We would guess that most educational researchers have concluded that formative assessment positively and almost directly affects learning, and this is likely due to the oft-cited Black and Wiliam study (1998). Not all advocates for formative assessment are using Black and Wiliam's (1998) classification of formative as being those assessments that provide information to be used as feedback to modify teaching and learning. As Stiggins (2002) emphasized, it is assessment *for* learning which most directly affects learning.

Consider teachers or textbook authors who value "formative assessment" because they believe it is designed to provide the teacher with feedback during instruction. They might suggest to teachers that they observe or listen to students during cooperative learning activities to see if students are "getting it." This activity is formative assessment under their paradigm. It might well provide helpful feedback to teachers and might guide them to improve instruction and, therefore, improve student learning. It is not the same "formative assessment" that Black and Wiliam (1998) found so convincingly effective in increasing student learning, however. Their formative assessment is a collaborative activity involving student self-assessment that alters both student and teacher behavior. Providing feedback only to alter instruction, while useful, is not the same formative assessment that helps students improve and control their own learning through reflection on assessment criteria (Fontana & Fernandes, 1994; Frederiksen & White, 1997) and the development of self-monitoring skills (McCurdy & Shapiro, 1992; Sawyer, Graham & Harris, 1992).

The purpose of formative assessment should be to increase student learning. We are not alone in believing that this is most effectively done when students use the data to adjust their own learning behaviors and teachers use the information to adjust their own teaching behaviors. Formative assessment works best when it is treated as assessment for learning, not as assessment of learning. Stiggins (2002) stressed that if formative assessment

is designed to provide feedback to teachers only, then formative assessment is not the same as assessment for learning. He agreed with Black and Wiliam (1998) that assessment for learning must heavily involve students in the process or learning is merely assessed, it is not produced.

Definitions Based on Assessment Purpose

When choosing labels for performance-based assessment, authentic assessment, and formative assessment, potential confusion might be best avoided by focusing on the intent of the assessment. Those who develop assessments, or choose them, do so for a certain purpose. This purpose is the core of all validity concerns that center on the interpretations of performance as it relates to an intended use (American Educational Research Association, American Psychology Association, & National Council on Measurement and Evaluation, 1999). For classroom assessment, the purpose may be to measure knowledge, to measure skill or ability, to provide feedback to the instructor, to provide feedback for students, or some combination of all these purposes and more. By categorizing assessments as performance-based assessment, authentic assessment, formative assessment, or assessment for learning, based on what the intended purpose of the assessment is, the common definitional criteria remain but the distinctions are clearer. The categorization scheme shown in Table 1 works well.

Under this classification scheme, some performance assessment is authentic, but not all. Any assessment that asks the student to demonstrate a skill or produce a product is a performance assessment. Consider the assignment of giving a speech. Its purpose as an assignment is likely to assess public speaking skills, so it is a performance-based assessment. The assessment is authentic, though, only if the conditions under which the speech is given match real-world contexts. It might be authentic, for example, if the student chose the topic of the speech, had time to prepare and revise, and had a purpose for the speech that is reasonably similar to the reasons people give speeches in the real world.

Table 1

Defining Assessment Types Based on Purpose

Purpose	Assessment Type
To measure a skill or ability	Performance Assessment
To measure ability on tasks which represent real-world problems or tasks	Authentic Assessment
To provide feedback to the teacher to assess the quality of instruction or to improve teaching behaviors, or to provide feedback to the student to assess the quality of learning and to improve learning behaviors	Formative Assessment
To provide feedback to students to assess the quality of learning and to improve learning behaviors	Assessment for Learning

Likewise, some formative assessment is assessment for learning, but not all. Black and Wiliam's research supports assessment for learning, but not all formative assessment approaches. For example, a teacher might take class time in the middle of a weeklong unit for a "practice" quiz that does not affect a student's grade and covers the content that is currently being taught. Because this assessment occurs while learning is occurring and provides feedback to teachers or students, it is formative assessment. If the feedback is provided to students in a way that allows them to evaluate and alter learning strategies, then the assessment is an example of assessment for learning. Notice also that the quiz format is not an example of authentic assessment, as it likely does not mirror activities typically engaged outside the classroom.

It is not our intent to criticize any current definition or categorization scheme for the variety of assessments. We are simply suggesting that it might be useful and more straightforward to classify each of these types or approaches to assessment based on what the purpose of the assessment is. Classroom assessments, of course, can have a complex set of purposes and be administered for multiple purposes. Indeed, that is probably more often the case than not. We hope only to clarify the distinctions between

three assessment terms—performance, authentic, and formative assessment, and we have focused only on the conceptual overlap between authentic and performance and the use of the term formative assessment. There are more distinctions among these three terms that could be explored. Even possible and reasonable definitions of authentic assessment and formative assessment are not distinct. Wiggins (1989), for example, described quality authentic assessment as including particular attention to the scoring criteria used, the role of the student as self-evaluator, and an observable display of mastery that might preclude cheating. If one accepts Wiggins conception of authentic assessment with its emphasis on the student as a partner in the assessment process, learning the criteria for quality performance, and taking part in self-assessment, then one could conclude that authentic assessment is formative assessment. But, that is a problem for another time.

References

American Educational Research Association (AERA), American Psychology Association (APA), & National Council on Measurement and Evaluation (NCME). (1999). *Standards for educational and psychological testing.* Washington, DC: Author.

Airasian, P. W. (2001). *Classroom assessment: Concepts and applications.* New York: McGraw-Hill.

Alkin, M. C. (1974). "Criterion-referenced measurement" and other such terms. In C. W. Harris, M. C. Alkin, & W. J. Popham (Eds.), *Problems in criterion-referenced measurement* (Center for the Study of Evaluation Monograph Series in Evaluation, No. 3; pp. 3–12). Los Angeles: Center for the Study of Evaluation, University of California.

Archbald, D. A., & Newmann, F. M. (1988). *Beyond standardized testing: Assessing authentic academic achievement in the secondary school.* Reston, VA: National Association of Secondary School Principals.

Baker, R. L. (1974). Measurement considerations in instructional product development. In C. W. Harris, M. C. Alkin, & W. J. Popham (Eds.), *Problems in criterion-referenced measurement* (Center for

the Study of Evaluation Monograph Series in Evaluation, No. 3; pp. 37–46). Los Angeles: Center for the Study of Evaluation, University of California.

Bergen, D. (1993). Authentic performance assessments. *Childhood Education, 70*, 99–102.

Black, P. J., & Wiliam, D. (1998). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan, 80*, 139–148.

Black, P. J., & Wiliam, D. (2004). Classroom assessment is not (necessarily) formative assessment (and vice-versa). *Yearbook of the National Society for the Study of Education, 103*, 183–188.

Bligh, J. (2001). Assessment: The gap between practice and theory. *Medical Education, 35*, 312.

Bloom, B. S. (1968). Learning for mastery. *Evaluation Comment, 1*(2), 1–5.

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 National Council on Measurement in Education, San Francisco, CA. (ERIC Document Reproduction Service No. ED351309)

Boston, C. (2002). The concept of formative assessment. *Practical Assessment, Research and Evaluation, 8,* 9. Retrieved March 7, 2006, from http://PAREonline.net/getvn.asp?v=8&n=9

Cangelosi, J. S. (2000). *Assessment strategies for monitoring student learning.* New York: Addison Wesley Longman.

Chase, C. I. (1999). *Classroom assessment for educators.* New York: Addison-Wesley.

Fennessey, D. (1982, July). *Primary teachers' assessment practices: Some implications for teacher training.* Paper presented at the annual meeting of the South Pacific Association for Teacher Education, Frankston, Victoria, Australia. (ERIC Document Reproduction Service No. ED229346)

Fontana, D., & Fernandes, M. (1994). Improvements in mathematics performance as a consequence of self-assessment in Portuguese primary school pupils. *British Journal of Educational Psychology, 64*, 407–417.

Frederiksen, J. R., & White, B. J. (1997, March). *Reflective assessment of students' research within an inquiry-based middle school science curriculum*. Presented at the annual meeting of the American Educational Research Association, Chicago.

Frey, B. B., & Schmitt, V. (2005, April). *Teachers' classroom assessment practices.* Presented at the Annual Meeting of the American Educational Research Association, Montreal, Quebec, Canada.

Glaser, R., & Nitko, A. J. (1971). Measurement in learning and instruction. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 625–670). Washington, DC: American Council on Education.

Gronlund, N. E. (2003). *Assessment of student achievement* (7th ed.). Needham Heights, MA: Allyn & Bacon.

Harris, C. W., Alkin, M. C., & Popham, W. J. (Eds.). (1974). *Problems in criterion referenced measurement* (Center for the Study of Evaluation Monograph Series in Evaluation, No. 3). Los Angeles: Center for the Study of Evaluation, University of California.

Harris, M. L., & Stewart, D. M. (1971, February). *Application of practical strategies to criterion-referenced tests*. Paper presented at the Annual Meeting of the American Educational Research Association, New York.

Herman, J. L., & Dorr-Bremme, D. W. (1984, February). *Teachers and testing: Implications from a national study. Draft.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA. (ERIC Document Reproduction Service No. ED244987)

Kubiszyn, T., & Borich, G. (2003). *Educational testing and measurement: Classroom application and practice* (7th ed.). Hoboken, NJ: Wiley.

Madaus, G. F., & Dwyer, L. M. (1999). Short history of performance assessment: Lessons learned. *Phi Delta Kappan, 80*, 688–695.

Marion, S. F. (Discussant). (2005, April). *New directions in classroom assessment practices.* Session at the Annual Meeting of the American Educational Research Association, Montreal, Quebec, Canada.

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)

McCurdy, B. L., & Shapiro, E. S. (1992). A comparison of teacher-, peer-, and self-monitoring with curriculum-based measurement in reading among students with learning disabilities. *Journal of Special Education, 26*, 162–180.

McMillan, J. H. (2001). *Classroom assessment: Principles and practice for effective instruction* (2nd ed.). Needham Heights, MA: Allyn & Bacon.

Mertler, C. A. (2003). *Classroom assessment: A practical guide for educators.* Los Angeles: Pyrczak.

Meyer, C. A. (1992). What's the difference between authentic and performance assessment? *Educational Leadership, 49*(8), 39–40.

Millman, J., & Greene, J. (1989). The specification and development of tests of achievement and ability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 335–366). Phoenix, AZ: American Council on Education, Oryx Press.

Mueller, J. (2005). *Authentic assessment toolbox.* Retrieved August 1, 2005, from http://jonathan.mueller.faculty.noctrl.edu/toolbox

Newman, F., Brandt, R., & Wiggins, G. (1998). An exchange of views on "Semantics, psychometrics, and assessment reform: A close look at 'authentic' assessments." *Educational Researcher, 27*(6), 19–22.

Nitko, A. J. (1984). Problems in the development of criterion-referenced tests: The IPI Pittsburgh experience. In C. W. Harris, M. C. Alkin, & W. J. Popham (Eds.), *Problems in criterion-referenced measurement* (Center for the Study of Evaluation Monograph Series in Evaluation, No. 3; pp. 59–82). Los Angeles: Center for the Study of Evaluation, University of California.

Noonan, B. W., & Duncan, C. R. (2005, April). *Peer and self-assessment in high schools.* Paper presented at the Annual Meeting of the American Educational Research Association, Montreal, Quebec, Canada.

Oescher, J., & Kirby, P. C. (1990, April). *Assessing teacher-made tests in secondary math and science classrooms.* Paper presented at the annual meeting of the National Council on Measurement in Education, Boston. (ERIC Document Reproduction Service No. ED322169)

Oosterhof, A. (2003). *Developing and using classroom assessments.* Upper Saddle, NJ: Merrill Prentice Hall.

Popham, W. J. (1974). Thus spake Psychometrika. . . . In C. W. Harris, M. C. Alkin, & W. J. Popham (Eds.), *Problems in criterion-referenced measurement* (Center for the Study of Evaluation Monograph Series in Evaluation, No. 3; pp. 95–97). Los Angeles: Center for the Study of Evaluation, University of California.

Popham, W. J. (2002). *Classroom assessment: What teachers need to know* (3rd ed.). Needham Heights, MA: Allyn & Bacon.

Sawyer, R. J., Graham, S., & Harris, K. R. (1992). Direct teaching, strategy instruction, and strategy instruction with explicit self-regulation: Effects on the composition skills and self-efficacy of students with learning disabilities. *Journal of Educational Psychology, 84,* 340–352.

Scriven, M. (1967). The methodology of evaluation. In R. Tyler, R. Gagne, & M. Scriven (Eds.), *Perspectives of curriculum evaluation* (pp. 39–83). Chicago: Rand McNally.

Shepard, L. (2000). The role of assessment in a learning culture. *Educational Researcher 29*, 7, 4–14.

Shepard, L. A. (2005, April). *Competing paradigms for classroom assessment: Echoes of the tests-and-measurement model*. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Quebec, Canada.

Skager, R. W. (1974). Generating criterion-referenced tests from objectives-based assessment systems: Unsolved problems in test development, assembly, and interpretation. In C. W. Harris, M. C. Alkin, & W. J. Popham (Eds.), *Problems in criterion-referenced measurement* (Center for the Study of Evaluation Monograph Series in Evaluation, No. 3; pp. 47–58). Los Angeles: Center for the Study of Evaluation, University of California.

Stiggins, R. J. (1987, Fall). Design and development of performance assessments. *Instructional Topics in Educational Measurement (ITEMS),* 1-9.

Stiggins, R. J. (1991). Assessment literacy. *Phi Delta Kappan*, *72*, 534–539.

Stiggins, R. J. (2002). Assessment crisis: The absence of assessment for learning. *Phi Delta Kappan, 83*, 758–765.

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

Taylor, C. S., & Bobbit-Nolen, S. (2005). *Classroom assessment: Supporting teaching and learning in real classrooms.* Upper Saddle River, NJ: Prentice Hall.

Terwilliger, J. (1997). Semantics, psychometrics, and assessment reform: A close look at "authentic" assessments. *Educational Researcher, 26*(8), 24–27.

Trice, A. D. (2000). *A handbook of classroom assessment*. New York: Addison Wesley Longman.

Ward, A. W., & Murray-Ward, M. (1999). *Assessment in the classroom.* London: Wadsworth.

Wiggins, G. (1989). A true test: Toward more authentic and equitable assessment. *Phi Delta Kappan, 70*, 703–713.

Wiggins, G. (1993). *Assessing student performance: Exploring the purpose and limits of testing.* San Francisco: Jossey-Bass.

Williams, J. M. (1991). *Writing quality teacher-made tests: A handbook for teachers.* (ERIC Document Reproduction Service No. ED349726)

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

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