Educational Measurement: Issues and Practice
Fall 2009, Vol. 28, No. 3, pp. 32–37

Commentary: Evaluating the Validity of Formative and Interim Assessment

Lorrie A. Shepard, *University of Colorado at Boulder*

In many school districts, the pressure to raise test scores has created overnight celebrity status for formative assessment. Its powers to raise student achievement have been touted, however, without attending to the research on which these claims were based. Sociocultural learning theory provides theoretical grounding for understanding how formative assessment works to increase student learning. The articles in this special issue bring us back to underlying first principles by offering separate validity frameworks for evaluating formative assessment (Nichols, Meyers, & Burling) and newly-invented interim assessments (Perie, Marion, & Gong). The article by Heritage, Kim, Vendlinski, and Herman then offers the most important insight of all; that is, formative assessment is of little use if teachers don't know what to do when students are unable to grasp an important concept. While it is true that validity investigations are needed, I argue that the validity research that will tell us the most—about how formative assessment can be used to improve student learning—must be embedded in rich curriculum and must at the same time attempt to foster instructional practices consistent with learning research.

Keywords: formative assessment, interim assessment, validity

Formative assessment is recognized to be a powerful tool for improving student learning. A group of assessment researchers in the British Educational Research Association (Assessment Reform Group, 1999)—keenly aware of the important differences between accountability assessment and assessment used in classrooms to inform instruction—commissioned the now famous review by Black and Wiliam (1998). Since that time, virtually no scholarly or popular treatment of formative assessment can begin without acknowledging Black and Wiliam's compelling research synthesis demonstrating the power of formative assessment to raise student achievement. Unfortunately, this obeisance has become so ritualized that a more detailed recollection of the bodies of research literature they examined has been lost. Attention has been focused on the positive effect sizes they reported, rather than the underlying theories that explain how formative assessment works.

The articles in this issue of *Educational Measurement: Issues and Practice* are loosely connected to one another, but all are directly relevant to the larger research literature on formative assessment. Nichols, Meyers, and Burling (this issue) offer a validity framework for evaluating formative assessment applications. Heritage, Kim, Vendlinski, and Herman (this issue) address very specific aspects of teacher knowledge needed to enact formative assessment effectively. Perie, Marion, and Gong (this issue) draw important distinctions between classroom formative assessment and more recently developed district-level interim assessments and offer a framework for complementary uses of both. Although authors of two of the papers invoke No Child Left Behind (NCLB) as context for their work, accountability testing is not the focus of their contributions. Rather, it is the research evidence and theoretical understandings underlying formative assessment that most need to be considered in responding to their arguments. Therefore, it is helpful by way of background to offer a brief reprise of the major findings from the Black and Wiliam formative assessment review.

Black and Wiliam drew together diverse bodies of research including studies addressing: teachers' assessment practices, students' self-perception and achievement motivation, classroom discourse practices, quality of assessment tasks and teacher questioning, and the quality of feedback. Because of the close overlap between formative assessment and feedback, they provided considerable detail regarding the features of effective feedback, drawing from both the cognitive and motivational literatures. For example, feedback is more likely to lead to improved student learning if it is directed toward successful completion of the learning task, with clear guidance about how to improve, but also requires thinking and "mindfulness" on the part of the student (Bangert-Drowns, Kulik, Kulik,

Lorrie A. Shepard is Dean and Professor at the School of Education, University of Colorado at Boulder, Campus Box 249, Boulder, CO 80309; Lorrie.Shepard@Colorado.edu.

& Morgan, 1991). Merely correcting errors is less effective. Focusing feedback on elements of the task is also consistent with findings from research on intrinsic and extrinsic motivation, showing the generally negative effects of praise and other forms of evaluation that cue students to focus more on themselves and their own abilities rather than mastery of the task.

Black and Wiliam were well aware that culture makes a difference in how students interpret and respond to various assessment practices. This is true in international comparisons, but is also true in the classroom expectations set up between any given teacher and her students. In this vein, they quote Gipps (1994), who called for a paradigm shift from a testing culture to an assessment culture, and Perrenoud (1991) who most famously argued that "Every teacher who wants to practice formative assessment *must reconstruct the teaching contracts so as to counteract the habits acquired by his pupils*" (p. 92, italics in original).

Black and Wiliam did not begin with a learning theory framework for analyzing these diverse bodies of research, but they drew inferences throughout highlighting the relationships between specific assessment strategies and assumptions about learning. For example, behaviorist learning theory underlying mastery learning approaches treats learning as isomorphic with test performance. As a result the studies in this literature often look very much like a teaching-the-test regime, where the pretest, instructional worksheets, and posttest look highly similar. Not surprisingly, the effect sizes for mastery learning are 1.17 for formative tests and .6 on teacher-made summative tests (Kulik, Kulik, & Bangert-Drowns, 1990), but essentially zero on standardized tests (Slavin, 1987). Behaviorist theory supports the use of external rewards and less challenging tasks to ensure the opportunity to reinforce successful performance (Skinner, 1954) and would not suggest the need for novel tasks to extend student learning or to ensure transfer.

Elsewhere I have argued for social-constructivist or sociocultural theories of learning (Shepard, 2000) as a framework for explaining and making coherent findings and insights from the many bodies of work reviewed by Black and Wiliam (1998). These theories are addressed most directly in Black and Wiliam's summary of research on self and peer assessment. Sociocultural theories can incorporate cognitive theories that focus on mental representations, schema-theory, and the like, but based on Vygotsky's (1978) cultural theory of development, they also account for the social and interactive ways that language constructions and ways of thinking are practiced, developed, and taken into the mind. Vygotsky's zone of proximal development also explains how with social support or scaffolding students can try out and then master competencies that are initially beyond their reach.

Reform efforts in mathematics and science education intended to change patterns of classroom discourse, to "make thinking visible" and help students develop strategies for reasoning from evidence, are explicitly derived from these theories. Whereas psychologists at one time treated cognitive development and motivation as wholly separate, activity theory and communities of practice (Lave & Wenger, 1991) help us understand how entwined are the development of increasing competence and an identity of meaningful participation. With an understanding of these theories, it is not just "cute" to have a second grader sit in "author's chair" and hear from classmates what they thought of his story, but critical to the development of his identity as a writer. These theories provide straightforward explanations as to why feedback improves learning in some studies and actually harms it in others, and they are critical to an understanding of how formative assessment works, when it does work.

Formative Assessment Validity Framework: Nichols, Meyers, and Burling

Nichols et al. begin their validity framing with important reminders. Just because it's labeled formative assessment doesn't make it so. At a time when the label formative assessment has been misappropriated with abandon (Shepard, 2008), this admonition seems especially critical. Moreover, it is the use of an instrument rather than the instrument itself that must be shown, with evidence, to warrant the claim of formative assessment. Nichols et al. also set down, from Wiliam and Black (1996), the more demanding of definitions for formative assessment, namely that it not only produce information amenable to instructional intervention, but that such intervention be effective in improving student learning.

> To sum up, in order to serve a formative function, an assessment must yield evidence that, with appropriate construct-referenced interpretations, indicates the existence of a gap between actual and desired levels of performance, *and suggests action that are in fact successful in closing the gap.* (p. 543, emphasis added)

Interestingly, proof of success was a hotly contested element in the drafting of a definition of formative assessment by the Formative Assessment for Teachers and Students (FAST) State Collaborative on Assessment and Student Standards (SCASS) sponsored by the Council of Chief State School Officers (CCSSO). The FAST SCASS definition focuses on the adjustment of ongoing instruction, and finesses the issue as to whether responses to formative assessment must be successful each and every time.

> Formative assessment is a process used by teachers and students during instruction that provides feedback to adjust ongoing teaching and learning to improve students' achievement of intended instructional outcomes. (McManus, 2008, p. 3)

Clearly, however, and consistent with Nichols et al.'s validity framework, formative assessment interventions must have, on average, a positive effect on learning to substantiate the validity claim.

Nichols et al. use the validity framework from Messick (1989) and Kane (2006) to lay out a validity argument detailing how formative assessment is expected to work. The point of a validity investigation, then, is to test whether the argument as a whole, as well as specific elements, works as intended. Their framework begins in the *assessment phase* with a domain model, which represents the knowledge and skills intended to be taught, to which the student model is compared, which represents skills actually mastered by the student as inferred from observations and student data. In the *instructional phase*, the teaching model is used to come up with an instructional prescription and instructional plan to close the gap between a student's current understanding and the targeted understanding. This follows closely Sadler's (1989) gap-closing model of formative assessment and Black and Wiliam's (1998) elements required for

effective feedback. Finally, a *summative phase* repeats data collection from the formative phase to determine whether the process has been effective.

This is all quite sensible, but disappointing. The presentation is at such a high level of generality that it is hard to imagine that a reader, who does not already know how to go about evaluating validity claims, would be able to specify the necessary model in a particular case. The authors warn us that "values play a role in interpreting student responses using statistical algorithms" and that "assumptions about the psychology of learning, including assumptions about the motivations and self-perceptions of students, may be implicit or explicit in the teaching model." But they do not explain how an acknowledgement of these assumptions would be made concrete in the collection of validity evidence. For example, item response theory is mentioned as a statistical method for aggregating student data to estimate the student model. No further guidance is offered. Yet, we know that the relevance to a teaching model is much more straightforward when this is done with reasonably unidimensional constructs and anchored both logically and empirically with key instructional tasks as in the development of learning progressions (Masters & Forster, 1996). In contrast, when heterogeneous pools of items are scaled and a student is said to have a score of 190 on "data and probability," there is not an obvious match to a teaching model that can be defended psychometrically (even if Rasch assumptions are satisfied). Lack of concrete guidance was especially disappointing given important earlier work by Nichols (1994) in which he distinguished between traditional measurement designs, which average over a test taker's responses, and cognitively diagnostic assessments, which are based on a carefully developed substantive model of the domain.

Nichols et al. intend to answer my questions about specific instantiation of their model by talking through an algebra tutoring example, but this too was disappointing. The authors do not explain why they chose to analyze the Sleeman, Kelly, Martinak, Ward, and Moore (1989) study. To be fair, however, it appears that they were most interested in the degree of experimental control and ruling out any other explanations for study results, while I was hoping for a good example of formative assessment. As described by Sleeman et al., the students in their series of three studies "received algebra instruction that was largely procedural; that is, algebra was treated as a series of transformations without extensive reference to possible meaning" (p. 554). Each of the two treatments being compared in study one lasted 35 minutes. While there were significant gains by the time of posttest 1 (approximately two items), the scores reverted to baseline by the time of posttest 2, which means there was no learning. This was such an impoverished example of teaching and learning, that it can hardly be used to form conclusions about whether attention to specific errors could help to improve understanding.

Sleeman et al. went on to conduct a second study hypothesizing that students in the first study might not have been "sufficiently involved with their learning (i.e., they were passive listeners to the tutor's instructions)." So, in the second study, "by having students verbally repeat the correct procedure back to the tutor, it was hoped to engage them more in their learning" (p. 556). The treatments in these studies do not reflect an adequate understanding of either cognitive theory or motivational theory even in 1989. This raises the pressing question of how underlying psychological theories, referred to by Nichols et al. as well as by Black and Wiliam (1998), are to be brought to bear in a validity evaluation. Should the framework only address the steps in feedback and gap closing? Or should implementation of feedback and ensuing interventions be tested for their fidelity to underlying theory?

While I find the substantive example troubling, I agree with the type of interpretive reasoning offered by Nichols et al. to evaluate each step in the validity framework. And there are many other studies, as cited in Black and Wiliam (1998), to support their inference that "to serve a formative function, the information must fit as a component of a system of coordinated assessment and instruction." They lose me, however, at the end of the article when they speak to "assessment developers." Surely they don't mean for assessment developers to build a remedial instructional model independent of regular instruction? Why not tap the enormous resources at Pearson and build a "coordinated assessment and instruction system" in the context of *Connected Mathematics*, for example, which is possibly the most widely implemented National Science Foundation sponsored, mathematics reform curriculum and now published by Pearson? *Connected Mathematics* has a well-developed teaching model and it has well-developed instructional tasks that also serve as assessment tasks. Studies have been done to date that address the overall effects of the curriculum on student achievement, but finer grained studies have not been done to examine whether insights from student work are used formatively to adjust instruction or whether particular ways of using formative information are more or less effective in accord with learning theory.

Teacher Knowledge for Formative Assessment: Heritage, Kim, Vendlinski, and Herman

The generalizability study by Heritage et al. (this issue) can be thought of as an investigation of a specific element in the Nichols et al. framework, that is, the instructional prescription step. "Given the student model and the teaching model, what instructional method should be implemented?" (Nichols et al., Figure 1). Heritage et al. developed a measure of teacher knowledge for formative assessment in which teachers responded to student answers on assessment items. Each teacher received three scores indicating how well she or he could: (a) identify the key mathematical principles being assessed, (b) characterize the student's level of understanding, and (c) determine appropriate next instructional steps. The results of the study showed that teachers were much more adept at the first two tasks but generally had difficulty saying what instructional interventions should be used given evidence of what a student did or did not understand. For some teachers, this difficulty was apparent for all of the mathematical topics (the distributive property, solving equations, and rational number equivalence), whereas other teachers had difficulty specifying instructional next steps for some but not all of the topics.

Obviously, as noted by Black and Wiliam (1998), "For assessment to be formative the feedback information *has to be used*—which means that a significant aspect of any approach will be the *differential treatments* which are incorporated in response to the feedback" (p. 16, italics added). Later, citing the curriculum-based assessment efforts by

Fuchs, Fuchs, Hamlett, and Stecker (1991), Black and Wiliam (1998) concurred with Heritage et al. regarding the importance of learning progressions for addressing the problem of what to do next.

> . . .teachers need more than good assessment instruments—they also need help to develop methods to interpret and respond to the results in a formative way. One requirement for such an approach is a sound model of students' progression in the learning of the subject matter, so that the criteria that guide the formative strategy can be matched to students' trajectories of learning. (p. 37)

One word of caution should be offered regarding the development of learning progressions to support student assessment and instruction. Sorting observed teacher responses to develop a scoring rubric makes sense when developing a measure of *teacher* knowledge based on these particular assessment items. However, to develop *student* learning progressions, even in these same domains, would require both expert knowledge and empirical evidence of student progress under conditions of effective instruction. In particular, it would be important to bring to bear evidence of conceptual understanding and the ability to apply and generalize knowledge along with more typical measures of procedural skill development, to identify particular misconceptions or obstacles, and to begin to link these with strategies or challenges shown to be effective in addressing these specific misconceptions.

Interim Assessment Validity Framework: Perie, Marion, and Gong

Perie et al. (this issue) consider appropriate uses of interim or benchmark assessments, which fill a niche in between state-level, once-per-year, summative tests, and day-to-day formative assessments used as part of classroom instruction. Like Nichols et al., they emphasize the conditions that instructionally grounded assessments would have to meet to satisfy the research-based definition of formative assessment, and they explain why most commercial products and district-developed periodic assessments fit, instead, their definition of interim assessment. The evaluative framework provided by Perie et al. is remarkably comprehensive. It asks the superintendent or school board members contemplating investment in an interim assessment system to identify their purposes and then to select carefully an assessment product that will serve these purposes well. In general, I like the *Consumer Reports* flavor of their analysis, which helps the reader by identifying criteria consistent with each particular desired outcome, such as minimizing dishwasher noise versus energy efficiency or choosing the right car for mostly in-town or highway driving.

One concern with the Perie et al. framework, however, is that it takes for granted the need for interim assessments of some kind. In their introduction they say that because of NCLB and because state tests cannot provide diagnostic information for individual students, "educators and policymakers have realized that other forms of assessments are necessary to inform instruction during the school year." Later in the article, the authors provide a stringent set of criteria that interim assessments intended for instructional purposes must meet, but the reader who does not follow their reasoning carefully is allowed to assume that this wish by policymakers is generally satisfied by typically available interim assessments. Not until the end of the article do we get some hint of skepticism on the part of Perie et al. indicating that, "it is not worth spending scarce resources on interim assessments that simply administer a series of mini-summative assessments," (p. 12), and even for instructional purposes, they argue that "resources would be better spent helping teachers learn formative assessment techniques, including using the information to intervene with students who do not yet understand key concepts" (p. 13). Interim assessments could sometimes be a good thing, but they are brand new and wholly unexamined. As recently as 2001, a significant National Research Council report, *Knowing What Students Know* (Pellegrino, Chudowsky, & Glaser, 2001), addressed coherence between large-scale and classroom assessments but did not even recognize intermediate, interim assessments. Therefore, some amount of skepticism and search for evidence is warranted.

The authors have distinguished three different purposes for interim assessments—instructional, evaluative, and predictive—and have identified the criteria to be met in each instance if interim assessments are to work as intended. Note that Perie et al. use the term "theory of action" (Argyris & Schon, 1978), which is the more familiar term in policy and evaluation circles, but this idea is directly parallel to the notion of a validity argument in measurement theory (Kane, 2006; Messick, 1989). How is the information from the test expected to be used, what inferences will be drawn with what evidentiary warrant, and is there evidence that consequences are as intended? Again, the authors' treatment of the issues is quite comprehensive, so I offer only additional points of clarification.

The evaluative use of interim assessments is broadest because it provides information about programs rather than about individual students. For example, when standard-based mathematics reforms were first introduced, a district might have wanted to evaluate how well schools were doing in teaching measurement and geometry as well as traditional number sense topics. District writing assessments are also a classic example of an assessment program that can be closely tied to curriculum reform and teacher professional development efforts. In my experience, programmatic uses of focused assessments can be the most helpful when they are initiated by subject-matter experts in a district, because the science coordinator or the literacy coaches are more likely to attend to the substantive quality of the assessment tasks and their fidelity in capturing curricular goals. In some cases, commonly administered science inquiry tasks or common writing prompts can be central to teacher professional development workshops focused on examining student work, even if the assessments are never formally scored and reported. But then, we would classify this as an instructional use of assessment, as discussed a bit later. If the decision is made to adopt formal interim assessments for program evaluation purposes, then good questions to ask would be, "what can this assessment tell us about the quality of our curriculum and instructional programs beyond what we have already learned from the state test, and are the subscales of the proposed assessment reliable enough to support such inferences?"

As Perie et al. suggest, predictive purposes for interim assessments seem straightforward but their usefulness tends to erode in practice. A school board member says we need an early warning system to know which students

won't meet the proficiency standard at the end of the year so they can receive extra help. (This assumes, first and foremost, that a significant majority of teachers does not already know which students are at risk. This is an assumption that should be directly tested as part of any validity investigation.) Technical experts, then, build mini versions of the end-of-year tests, complete with equated or estimated cut scores for proficiency. The only problem is that when the interim test is given in October, two-thirds of the content has not yet been covered. While it is more difficult to come up with an equivalent cut score on a test that covers only September and October content, administering a test relevant to specific curriculum units is much more likely to provide useful information; and unit-specific interim tests can correlate as highly with the end-of-year test as do carbon-copy tests.

Unfortunately, the criteria identified for predictive assessments may lead in contradictory directions. Perie et al. say that predictive assessments should have a similar mix of item types as the criterion measure and should be designed from the same or similar blueprint. These requirements make the most sense if the district is going to do formal progress reporting and growth modeling for schools mid year and therefore need the same kind of rigorous technical accuracy as for accountability tests. Perie et al. then go on to say that predictive assessments should contain "enough diagnostic information so that remediation can be targeted for students predicted to score below the cut on the criterion measure," but this leads to the problem of diagnosing as weaknesses topics that have not yet been taught. Moreover, accountability tests have already been said to cover too broad a range of topics to provide meaningful diagnostic information, which must then also be true of parallel versions. Such tests can't pinpoint what particular skills a student is lacking. They tell us only which students are the most at risk. Once examined, districts may find that their real purposes for interim assessments are instructional rather than predictive.

Perie et al. identify seven general criteria and eight specific instructional criteria that can be used to evaluate the adequacy of interim assessments. If all of these criteria were met—especially regarding the quality of items in capturing learning goals and the close link between assessment results and instructional decisions—then such uses would be formative at least for the week or two following administration of the test. These criteria are formidable, however, and as yet are seldom realized. As Perie et al. point out, essentially all of today's commercially available systems are of a very different character from the formative assessments evaluated by Black and Wiliam (1998). So what do policymakers get if they invest in less-than-ideal interim assessments with the intention of informing instruction? Some of the available item banks, it can be argued, are as good as end-of-chapter tests in many current textbooks, so why not provide teachers with the flexibility of electronically available materials? There can't be much harm in using these predominantly multiple-choice item banks; at least we should admit that they are not that much different from the quizzes and worksheets used routinely in many traditional classrooms. Policymakers should be aware, however, that such products do not capture conceptual understanding and problem-solving goals of more reform-oriented curricula, and they do not provide any diagnostic insight about what in particular a student is not understanding.

In contrast to the many multiple-choice interim assessment products available, focused primarily on scoring and ranking of students, there are a few collections of instructionally focused assessment tasks that districts may also want to consider. These products will not necessarily always provide scores to compare schools or predict end-of-year state test results, but they will enable greater teacher learning about student learning. The Silicon Valley Mathematics Assessment Collaborative, for example, uses performance assessments developed by the New Standards Project and the Mathematics Assessment Resource Service (MARS) to test student reasoning, problem solving, and communication skills in the context of five core mathematics ideas identified for each grade level. These assessments are administered formally once per year in the participating districts and are used explicitly to address the what-to-do-next problem identified by Heritage et al. Teachers participate in scoring workshops where student errors and misconceptions are surfaced and discussed. Then a "Tools for Teachers" document is developed that specifically links common misconceptions with classroom practices that may be "contributing to errors, poor communication, or genuine lack of understanding" (Foster & Noyce, 2004). More importantly, teachers use released assessment tasks throughout the school year, at the beginning and end of instructional units aimed at each core idea. Math coaches lead discussions focused on the mathematics that students need to know to complete the tasks, and in many cases address weaknesses in teachers' mathematical understanding as well.

Full option science system (FOSS) kits are another example of curriculum-embedded assessment resources closely tied to instructional activities. FOSS kits provide assessment tasks intended for both mid-unit formative purposes and end-of-unit summative assessments. These materials strengthen teacher knowledge by highlighting key understandings that need to be demonstrated when observing mid-unit checking-for-understanding tasks, and they offer advice about what prior instructional strategies to return to when understanding breaks down. Perie et al. argue that if policymakers want to adopt an interim assessment primarily for the purposes of helping teachers improve instruction, "the resources would be better spent helping teachers learn formative assessment techniques." I would agree, except that the ideal would be to invest in materials like the MARS or FOSS assessments so that teachers would have the support of well-developed assessment materials along with professional development to learn more productive ways of using feedback and self-assessment as part of instructional practice.

Conclusion

NCLB does indeed create a context for examining formative assessment because in many settings the pressure to raise test scores has created overnight celebrity status for formative assessment. Its powers to raise student achievement have been touted, however, without attention to the research on which these claims were based. The articles in this issue bring us back to first principles by offering separate validity frameworks for evaluating formative assessment (Nichols et al.) and newly invented interim assessments (Perie et al.). The Heritage et al. article then offers the most important

insight of all, that is, formative information is of little use if teachers don't know what to do when students are unable to grasp an important concept.

While it is true that validity investigations are needed, I argue that this work will be of limited value if it is conducted by measurement specialists separate from curriculum and subject matter expertise. This is one reason that research on learning progressions is so important currently, not just because it considers longitudinal development of proficiency, rather than cross-sectional status measures, but because in the development of learning progressions it becomes impossible to address *measurement* questions without also considering corresponding *content and learning* questions. What are the core understandings and skills to be mastered, what activities support student engagement with these ideas, how is mastery demonstrated, what are typical errors and misconceptions, how can teachers back up or try another strategy when learning falters, what social and emotional supports are needed along with cognitive help? I believe that the validity research that will tell us the most about how formative assessment can be used to improve student learning must be embedded in rich curriculum and must at the same time attempt to foster instructional practices consistent with research on learning.

References

Argyris, C., & Schon, D. A. (1978). *Organizational learning: A theory of action perspective*. Reading, MA: Addison-Wesley.

Assessment Reform Group (1999). *Assessment for learning: Beyond the black box*. Cambridge: University of Cambridge School of Education.

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

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy, and Practice*, *5*(1), 7–74.

Foster, D., & Noyce, P. (2004). The mathematics assessment collaborative: Performance testing to improve instruction. *The Silicon-Valley Mathematics Initiative*. Available at http://www.noycefdn.org/publications.html Retrieved April 2, 2009.

Fuchs, L. S., Fuchs, D., Hamlett, C. L., & Stecker, P. M. (1991). Effects of curriculum-based measurement and consultation on teacher planning and student achievement in mathematics operations. *American Educational Research Journal*, *28*, 617–641.

Gipps, C. (1994). *Beyond testing: Towards a theory of educational assessment*. London: Falmer Press.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Washington, DC: The National Council on Measurement in Education and the American Council on Education.

Kulik, C. L. C., Kulik, J. A., & Bangert-Drowns, R. L. (1990). Effectiveness of mastery-learning programs: A meta-analysis. *Review of Educational Research*, *60*, 265–299.

Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation*. Cambridge, UK: Cambridge University Press.

Masters, G., & Forster, M. (1996). *Progress maps: Assessment resource kit*. Melbourne, Victoria, Australia: The Australian Council for Educational Research.

McManus, S. (Ed.) (2008). *Attributes of effective formative assessment*. Washington, DC: Council of Chief State School Officers.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (pp. 13–103). Washington, DC: American Council on Education and National Council on Measurement in Education.

Nichols, P. D. (1994). A framework for developing cognitively diagnostic assessments. *Review of Educational Research*, *64*(4), 575–603.

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (Eds.) (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.

Perrenoud, P. (1991). Towards a pragmatic approach to formative evaluation. In P. Weston (Ed.), *Assessment of pupil achievement: Motivation and school success* (pp. 79–101). Amsterdam: Swets & Zeitlinger.

Sadler, R. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, *18*, 119–144.

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

Shepard, L. A. (2008). Formative assessment: Caveat emptor. In C. A. Dwyer (Ed.), *The future of assessment: Shaping teaching and learning*. New York: Lawrence Erlbaum Associates.

Skinner, B. F. (1954). The science of learning and the art of teaching. *Harvard Educational Review*, *24*, 86–97.

Slavin, R. E. (1987). Mastery learning reconsidered. *Review of Educational Research*, *57*, 175–214.

Sleeman, D., Kelly, A. E., Martinak, R., Ward, R. D., & Moore, J. L. (1989). Studies of diagnosis and remediation with high school algebra students. *Cognitive science*, *13*, 551–568.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.

Wiliam, D., & Black, P. (1996). Meaning and consequences: A basis for distinguishing formative and summative functions of assessment? *British Educational Research Journal*, *22*(5), 537–548.