Formative assessment: Caveat emptor

Lorrie A. Shepard

University of Colorado at Boulder

ETS Invitational Conference 2005

The Future of Assessment: Shaping Teaching and Learning

New York

October 10-11, 2005

Formative assessment as a part of good teaching has been around for a long time. In the past two decades, however, formal theory about this type of assessment – used to further students' developing understandings and to engage students in taking responsibility for their own learning – was developed in other countries (Black & Wiliam, 1998; Cowie & Bell, 1999; Sadler, 1989), in part to counter the negative effects of external accountability tests exported by the U.S. Recently, this robust and well-researched knowledge base has made its way back across the oceans, offering great promise for shifting classroom practices toward a culture of learning (Shepard, 2000; Stiggins, 2002).

Unfortunately, the arrival of formative assessment in America was ill timed. This potentially powerful classroom-based learning and teaching innovation was overshadowed almost immediately by the No Child Left Behind Act (January 2002) with its intense pressure to raise scores on external accountability tests. The title of my chapter is prompted by the recent burgeoning of so-called "formative assessments" offered by commercial test publishers to help raise test scores for NCLB. "Everyone knows that formative assessment improves learning," said one anonymous test maker, hence the rush to provide and advertise "formative assessment" products. But are these claims genuine? Dylan Wiliam (personal communication, 2005) has suggested that prevalent interim and benchmark assessments are better thought of as "early-warning summative" assessments rather than as true formative assessments. Commercial item banks may come closer to meeting the timing requirements for effective formative assessment, but they typically lack sufficient ties to curriculum and instruction to make it

possible to provide feedback that leads to improvement. The misappropriation of the formative assessment label has become so pervasive that one assessment CEO invested in a series of essay-length ads in *Education Week* (Kahl, 2005a, 2005b, 2006a, 2006b, 2006c) to warn educators that what vendors are selling are not truly formative assessments.

Because of the widespread confusion in terminology, I begin this chapter with a definition of formative assessment and contrast it with formative program evaluation and with testing for remedial placement. I argue that benchmark and interim assessments are better suited for making instructional program decisions and gross remedial placement decisions rather than day-to-day, individual student adjustments in instruction. Although I do not have any individual authority to insist that the definition of formative assessment that I propose is the correct one, my argument is that the official definition of formative assessment should be the one that best fits the research base from which its claims of effectiveness are derived. One might think of this as a "truth in labeling" definition of test validity (Shepard, 1993). Following the discussion of definitions, I provide a brief overview of the research base for formative assessment focusing on those features that directly link to learning theory and thereby help to explain how formative assessment works to improve learning. Then I turn to a quite different research literature. While the effects of benchmark and interim assessments do not have a foundation in research, it is plausible that findings from the implementation of other external summative tests would generalize to this new application. Therefore, I review the teaching-the-test research as a lens for thinking about the possible effects of administering standardized tests more frequently.

In the concluding sections of the chapter, I offer criteria that help to clarify further the distinction between formative assessment and formative program evaluation. If met, these criteria also ensure that each type of formative inquiry is effective. Finally, I propose solutions for test makers interested in ensuring the integrity and efficacy of their products and for state and district policymakers interested in enhancing teachers' formative assessment skills.

Distinguishing formative assessment from formative program evaluation and remedial placement tests

The terms assessment and evaluation are used interchangeably in many contexts. Here, however, we want to distinguish the type of formative assessment that helps students learn during the course of instruction from other types of testing or data gathering. I find it useful, therefore, to adopt the clear distinction made by the Office of Economic Co-operation and Development (OECD) in a study of formative assessment in eight countries.

> For purposes of this study, assessment refers to judgments of student performance, while evaluation refers to judgements of programme or organizational effectiveness (OECD, 2005, p. 25)

Similarly, in a review of the formative assessment literature from French-speaking countries, Allal and Lopez (2005) traced the history of formative assessment from Scriven's (1967) original definition of "formative evaluation" of educational programs, noting that the term "assessment" had "progressively replaced 'evaluation' when the object is student learning in the classroom" (p. 241).

Formative assessment is defined as assessment carried out during the instructional process for the purpose of improving teaching or learning (Shepard, Hammerness, Darling-Hammond, & Rust, 2005). Similarly, OECD authors (2005) said that "Formative assessment refers to frequent, interactive assessments of student progress and understanding to identify learning needs and adjust teaching appropriately" (p. 21). When Kahl (2005a) argued against misuse of the term by test vendors, he emphasized that formative assessment is a "midstream" tool that teachers use "to measure student grasp of the specific topics and skills they are teaching" (p. 38). What makes formative assessment *formative* is that it is immediately used to make adjustments so as *to form* new learning. As Sadler (1989) explained in his early contribution to the theory of formative assessment, feedback is a critical element, requiring that teachers (and ultimately students) have a clear vision of the skills to be learned, appraise current student progress, and make clear to students how to improve.

Benchmark and interim assessments have been adopted by many school districts to help monitor progress during the school year toward meeting state standards and NCLB performance goals. Typically these assessments are formal, machine-scored instruments administered at the end of every quarter, or sometimes as frequently as once per month. They serve as formative program evaluation tools by providing teachers with information about which content standards have been mastered well and which will require additional instructional attention. In addition, benchmark and interim assessments may report the specific content standards mastered by each student, thereby identifying individual students' strengths and weaknesses. The individual profile data from these assessments are not directly formative however, for two reasons: the data

available are at too gross a level of generality and feedback for improvement is not part of the process.

Benchmark and interim assessments function much more as remedial placement tests rather than as substantive formative assessments. For example, if a fourth-grade student is low on the "number, operation, and quantitative reasoning" standard, a teacher would have to work with that student further and do additional assessment to find out whether this meant that the student was having problems with understanding place value, representing fractions, or understanding multiplication and division. For most teachers, scores on benchmark tests simply signal which students are most at risk and therefore require the most attention rather than indicating the specific learning area that is in need of improvement. Such focusing of effort may indeed be one of the primary purposes for using these assessments, but the scores do not provide substantive insights about how to intervene.

Because of the grossness of the information from reliable subtest scores, interim assessment results can only be used to make relatively gross instructional-program-level decisions. For example, if class results show a relative weakness on the math subtest "Statistics and Probability" and the teacher notices that many of the items on the test involved bar graphs using objects, bars, and tally marks, then the teacher might plan a review lesson on bar graphs. One might think that responses to individual test items would provide more insight for specific students, but the item-level information is unreliable and only loosely coupled with instructional lessons and units of study. This is what I call the "1000 mini-lessons problem." Over the course of a year it would take a thousand mini-lessons to respond item-by-item and student-by-student to all of the

missed items on interim assessments. And resulting lessons would be incoherent and decontextualized for the students as well as impractical for the teacher. The truth is that teachers do not have time to go back and reteach lessons, after the fact, for all of the topics missed on interim assessments, without jeopardizing the curriculum for the next months of the school year. In contrast, true formative assessment, which involves natural questioning and follow-up as teachers interact with students during the course of instruction, is both more targeted to specific student needs in the context of meaningful lessons and more time-efficient because it occurs as a part of normal teaching.

Research base for formative assessment

Formative assessment has an extensive research base that draws on both cognitive and motivational research. An early review provided by Crooks (1988) from the University of Otago in New Zealand, for example, was noteworthy because it brought together findings from the literatures in educational measurement, motivational psychology, learning theory (both behaviorist and cognitive), and research on teaching – literatures that at that time rarely acknowledged one another. The recommendations Crooks offered for educational practice already contained most of the important features of more comprehensive present-day research syntheses. For example, classroom assessments guide student judgments about what is important to learn and affect students' self-perceptions of competence. Greater learning occurs when assessments focus on deep learning rather than surface or memorization approaches to learning. Useful feedback is much more important for learning than is maximizing the reliability of summative evaluations. Cooperative learning contributes to students' active engagement and helps to develop valuable peer and self-assessment skills.

The landmark review by Black and Wiliam (1998) is the most widely cited reference on formative assessment and stands behind the common knowledge that "Everyone knows that formative assessment improves learning." Black and Wiliam examined 250 studies from research literatures addressing current classroom practices; student motivation and student participation in assessment practices; learning theory; specific classroom strategies such as discourse and questioning; and the properties of effective feedback. They concluded that formative assessment has a more profound effect on learning than do other typical educational interventions, producing effect sizes of between .4 and .7. Moreover, formative assessment practices tend to help low-achieving students more than they help high-achieving students. One way to think about this latter finding is that formative assessment helps to develop metacognitive skills and enhance motivation differentially for low-achieving students because high-achieving students already have these resources intuitively or through other supports.

Close examination of the research literature helps us identify the features of formative assessment, or causal mechanisms, that make it work to improve learning. For example, we know from cognitive research that having students become self-aware in monitoring their own learning, also referred to as *meta-cognition*, improves achievement. In the case of Palincsar and Brown's (1984) *reciprocal teaching*, for example, teaching reading comprehension strategies -- like thinking about the story and making predictions about what comes next -- dramatically improved the reading proficiency of low-performing middle school students. Similarly in the formative assessment literature, teaching students to self-assess so they can internalize and use criteria as they carry out their work increases both the quality of student projects and conceptual understanding

(White & Frederickson, 2000). Other bodies of work in the cognitive literature demonstrate the importance of engaging students' *prior knowledge* to support new learning and the effectiveness of focusing on principled understanding to enable *transfer* and knowledge generalization. Formative assessment processes connect directly to these learning strategies when they address Sadler's questions, where are you now, and where do you want to go? In addition, transfer is supported when a rich array of tasks is used both for assessment and for instruction (Shepard, 1997).

Understanding the cognitive and motivational theories underlying formative assessment is essential because these theories explain why formative assessment works when it works. Feedback is the most obvious feature of formative assessment and the one with the strongest research base (i.e., the largest number of studies). Surprisingly, however, feedback is not always or even usually successful. Kluger and DeNisi's (1996) meta-analysis cautions that in one-third of studies feedback worsens performance, when evaluation focuses on the person rather than the task. In one-third of comparisons there is no difference in outcomes with and without feedback. Only in the one-third of studies where the feedback focused on substantive elements of the task, giving specific guidance about how to improve, did feedback consistently improve performance. Thus, merely telling students their score or proficiency category is not the type of feedback endorsed by the formative assessment literature.

Understanding the theoretical basis of formative assessment is also important because it provides coherence, thus helping to ensure that separate elements of effective practice make sense and work together. If we think of teachers as learners, then our goal should be a deeper and more coherent understanding of learning theory as a means to tie

together not only formative assessment strategies but also to aid in seeing how formative assessment relates to discourse reforms in mathematics, comprehension strategies in reading, inquiry methods in science, and so forth. Although teachers and teacher education students often have little patience with theory, big-picture understandings are especially important when we are trying to change our teaching practices. Theory helps us think about what to do when we can't rely on past experience.

Findings from the research on motivation provide additional insights, especially regarding the relationship between classroom summative and formative assessments. Research on motivation might also have significant implications for the increased frequency of external testing. We know that extrinsically motivated students work toward "performance goals," i.e., to get good grades, to please the teacher, and to appear competent to others. In the literature this is termed a "performance orientation." Performance-oriented students pick easy tasks and are less likely to persist once they encounter difficulty. In contrast, intrinsically motivated students, or students with a learning orientation, work toward "learning goals," i.e., to feel an increasing sense of mastery and to become competent (in contrast to appearing competent). Learning-oriented students are more engaged in schoolwork, use more self-regulation, and develop deeper understanding of subject matter. The most alarming finding from this literature is that students can learn to be extrinsically motivated, or to become extrinsically motivated in some contexts and not in others. Normative grading practices and extrinsic rewards produce performance-oriented students (Stipek, 1996). Obviously, not all mastery-oriented students will give up their love of learning because of a teacher's comparative grading practices, but the evidence is substantial that many students learn to focus on

grades because grades have been used so pervasively as rewards to control behavior and direct student effort.

A goal in developing a formative assessment classroom culture is to counteract students' obsession with grades and to redirect interest and effort toward learning. Motivation research on self efficacy and children's beliefs about ability also teaches us valuable lessons about how day-to-day uses of feedback and praise can shape children's confidence about their abilities as learners. Praising children for "being smart" when they perform well on tasks can have negative consequences for learning because such praise fosters students' implicit beliefs that intelligence and ability are fixed. In studies over the course of three decades, Carol Dweck (2002) has found that students who believe that intelligence is an unchangeable characteristic they were born with, what she calls an "entity" theory of self, are flummoxed by difficult problems and tend to avoid academic challenges. In contrast, students who have been taught that ability can be increased by effort, who hold an "incremental" theory of self, are more likely to seek academic challenges and to persist when faced with difficult problems. Feedback that focuses on a student's level of effort, evidence of alternative reasoning strategies used, and the specifics of work products fosters incremental beliefs about ability and results in more constructive behavior in the face of learning obstacles. Similar to Claude Steele's (Steele & Aronson, 1995) research on stereotype threat, Dweck and other attribution researchers find that female and minority students are more likely to hold entity theories of intelligence and to lack confidence in their ability to perform difficult tasks. Importantly, praise focused on effort and strategies can change children's adherence to "entity" beliefs, which in turn increases their resilience and learning.

Insights from the cognitive and motivation literatures can be drawn together in the more encompassing sociocultural theory of learning. According to sociocultural theory, children develop cognitive abilities through social interactions that let them try out language and practice their reasoning. Instead of being born with a fixed level of intelligence, children become "smart" through what Barbara Rogoff (1990) calls an "apprenticeship in thinking." In various learning contexts – talking at the dinner table, helping in the kitchen, doing math in classrooms – learners have both expert models and supports from adults or peers to enable them to participate in that activity. This process of providing support to help the learner attempt and then master increasingly complex tasks on their own is called *scaffolding*. When Ed Gordon talks about the idea of creating an environment or a culture where we support students' learning and their ability to participate in demanding academic contexts, he is talking about this theory. Sociocultural theory folds together an understanding of how children learn and at the same time how they develop identities as capable learners. When implemented by master teachers, formative assessment practices further cognitive goals and at the same time draw students into participation in learning for its own sake.

Research on teaching the test

A well known finding from the cognitive literature is that principled learning and transfer are aided when learning takes place across multiple contexts (Brown, Collins, & Duguid, 1989). In a sense, transfer is made possible when it is built into instructional routines, thereby allowing students to gain experience with tasks that look different (superficially) but that tap the same underlying principles. By contrast, to permit their frequent use at reasonable cost, benchmark and interim assessments are typically

multiple-choice, machine-scoreable instruments and are therefore quite limited in the knowledge representations they offer. There is reason, therefore, to be concerned that the increased frequency of standardized test administrations will narrow conceptions of subject matter and thereby harm student learning. A brief review of the literature on teaching the test helps to document how this narrowing works and what impacts it has.

After the first decade of high-stakes testing in the 1980's, the U. S. Congressional Office of Technology Assessment (1992) produced a report on *Testing in American Schools*, which concluded that test-driven reforms produce "test-score inflation" and "curriculum distortion." Test score inflation is a useful term that reminds us that it is possible for test scores to go up without a commensurate increase in learning. Curriculum distortion occurs when teachers teach what is on the test and ignore other content. Recent declines in science test scores, for example, have been attributed to neglect of science because of increased pressure to raise test scores in reading and mathematics. Another, potentially more serious, meaning of curriculum distortion is to distort even the way that reading and mathematics are taught, conceiving of knowledge in these subject areas only in the ways they are represented on the test. It is this type of fundamental curriculum distortion that explains how test-score inflation happens. Unhappily, another significant finding from the teaching-the-test literature is that these negative impacts are greatest for poor and minority children because the poorer the school, the more time is devoted to instruction that imitates the test (Madaus, West, Harmon, Lomax, & Viator, 1992).

Figure 1 illustrates the phenomenon of test score inflation. These data are from third graders in a very large urban district in a high-stakes testing environment. Prior to

1987, the district had been administering Test C, a well-known standardized achievement measure. In 1987 a new standardized test was adopted and scores dropped dramatically. The two standardized tests looked very much alike. Subtest names were almost identical and items on both tests were all multiple-choice. Almost immediately after the first administration of the new tests, test scores went up and continued to rise until they reached the same high level of the previous test. In 1990, as an additional check, Koretz, Linn, Dunbar, and Shepard (1991) administered the old test to a random subsample of district third graders. Now that the old test had become unfamiliar, to the current third graders, performance fell off dramatically. We believe that these comparisons illustrate the non-generalizability of test-score gains in this high-stakes context Students had become more proficient on the exact test formats, without the conceptual understanding that good test performance should signify.

INSERT FIGURE 1 ABOUT HERE

FIG. 1. Performance on familiar and unfamiliar standardized tests with very similar content and format

This idea of being able to do well on a test without really understanding the concepts is difficult to grasp. Indeed, many educational reformers believe that teaching the test might not be all bad: "At least they'll know what's on the test." The two sets of questions in Figure 2 are examples from a much larger set of items used in a study by Koczor (1984). Koczor's findings illustrate sharply how it is possible to look as if you understand Roman numerals without understanding them at all. In the Koczor study, students were randomly assigned to one of two conditions. One group learned and practiced translating Roman to Arabic numerals. The other group learned and practiced

Arabic to Roman translations. At the end of the study each group was randomly subdivided again (now there were four groups). Half of the subjects in each original group got assessments in the same format as they had practiced. The other half got the reverse. Within each instructional group, the drop off in performance, when participants got the assessment that was not what they had practiced, was dramatic. Moreover, the amount of drop-off depended on whether participants were low, middle, or high achieving. For low-achieving students, the loss was more than a standard deviation. Students who were drilled on one way of translation appeared to know the material, but only so long as they were not asked to translate in the other direction. Koczor's findings show clearly the harm of teaching content using only a narrow range of problem types.

INSERT FIGURE 2 ABOUT HERE

FIG. 2. Examples of items used in both teaching materials and in testing materials in the study by Koczor (1984)

In principle, multiple-choice test questions can be written to elicit higher-order cognitive processes. However, it is more difficult to write such items and even more difficult to have them survive pilot testing because high-inference items are often found to be ambiguous by reviewers and examinees. These difficulties are multiplied a hundred-fold by the sheer quantity of test items being generated for item banks and high-frequency testing. For example, the prompt and questions in Figure 3 were written to imitate as closely as possible the type of items displayed on a current "formative assessment" website. Although these reading passages resemble sophisticated cloze techniques, in fact, they are based on unnatural paragraphs. Instead of emphasizing comprehension of the overall meaning of the passage, repeated use of the same format invites students to

learn the strategy that the answer is nearly always in the sentence before the blank. Given the large volume of items currently being generated by test publishers, it is not surprising that interim tests are disproportionately low-level, fact-type questions. Use of such questions increases the likelihood that students will correspondingly adjust their learning strategies to conform to what the tests tell them are the goals for learning.

INSERT FIGURE 3 ABOUT HERE

FIG. 3. An example constructed to imitate items on a "Formative Assessment" website

The motivation literature cited previously also warns us that teaching to the test is likely to have negative motivational consequences as well as negative cognitive outcomes. The Fall Conference Sheet and Student "Self-Assessment" in Figure 4 illustrates how well-intentioned efforts to raise test scores can, perversely, lead to a performance orientation and to an emphasis on evaluation, which we know reduces students' intrinsic motivation and interest in material for its own sake (Stipek, 1996). The claim of self-assessment in this example is thus a distortion of key principles from the research on self-assessment, because the rating form does not engage students in thinking about the substantive features of their work. Indeed the only inference that could be drawn from the rating task about how to improve would be to try harder (good advice if you were not making much effort before, but frustratingly unhelpful if you were already trying hard). Learning that the purpose of learning is to perform on examinations exacerbates what Lave and Wenger (1991) pointed to as the commoditization of learning, which often occurs in school when knowledge and skills to be learned are entirely removed from any context of use. When this is the case, students have no compelling reason to participate except to produce for the test.

PLACE FIGURE 4 ABOUT HERE

FIG. 4. A recent example of materials used to motivate students to raise their test scores.

There are no definitive studies demonstrating either the harm or the benefit of benchmark or interim assessments. The widespread use of these instruments is too new to have been studied systematically. Indeed, these instruments are for the most part being sold without even the minimum validity evidence required for standardized tests (American Educational Research Association, American Psychological Association & National Council on Measurement in Education, 1999). It is not unreasonable, however, to generalize from the findings from research on high-stakes accountability tests, noting in particular that negative impacts on learning will be greatest when assessments based on limited item formats are administered at frequent intervals.

Criteria for effective interim assessments and formative assessment

Knowing What Students Know (Pellegrino, Chudowsky, & Glaser), the landmark report on assessment issued by a National Research Council committee in 2001, called for "balanced assessment systems" to redress the balance of resources between classroom and external forms of assessment. Key features recommended for a balanced assessment system were comprehensiveness, coherence, and continuity. *Comprehensiveness* refers to the need for multiple sources of evidence to draw inferences about an individual student's proficiency. The property of *coherence* refers to the need for a shared model of learning linking curriculum, instruction, and assessment within the classroom and also linking classroom assessments and external, large-scale assessments. *Continuity* extends the underlying model of learning to allow for a longitudinal assessment of learning

progress over time. As recently as 2001, in *Knowing What Students Know*, there was no mention of interim assessments as necessary components of a balanced assessment system.

Very recently, at the 2006 CCSSO Large-Scale Assessment Conference, for example, a new use of the term "comprehensive assessment system" has been adopted to try to bring coherence to a landscape that now includes three levels of assessments: state accountability tests, district interim tests, and classroom formative and summative assessments. Superintendents, school board members, and other policy leaders at the state and local level should be cautioned that interim assessments are not essential to an effective assessment system and, as stated previously, they lack a research base. Benchmark and interim assessments are an invention of the testing industry that has been welcomed by policymakers as a way to "do something" immediately in response to NCLB. The decision to invest in interim assessments should be weighed against other potentially more effective uses of the same resources.

If the decision is made to purchase benchmark or interim assessments, then meeting the criteria detailed below will help to increase the likelihood that interim tests will provide useful information and avoid negative side effects for students. Criteria are also offered for effective formative assessment in classrooms. Comparing the two sets of criteria, as illustrated in Table 1, shows how the two types of assessment can be coherent while at the same time emphasizing what each should do uniquely.

INSERT TABLE 1 ABOUT HERE

TABLE 1. Criteria for effective interim assessments and formative assessment.

Assessments must embody learning goals. The first criterion, desired of both interim assessments and day-to-day formative assessment, is that they "embody learning goals" and fully represent what it is that we want students to master. The term *authentic assessment* is often used to convey this idea that students be engaged in demonstrating their skills and "know-how" in the context of realistic tasks that reflect the "core challenges of the field of study, not the easily scored" (Wiggins, 1998, p. 23). In classrooms, formative assessment can readily be done in the context of mathematics problems, history papers, and science experiments, focusing on the key concepts and competencies that are the aims of a given instructional unit. Interim tests could similarly present mastery tasks calling on students to apply the knowledge and skills developed during a quarter or semester's time and would therefore be coherent with the preceding instruction and with classroom assessment. For example, the Center for Research on Evaluation, Standards, and Student Testing (CRESST) is developing POWERSOURCE assessments to tap powerful principles, such as representation, equivalence, and transformation that underlie mathematical understanding across problem types (Niemi, Vallone, & Vendlinski, 2006).

By contrast, typical interim tests being sold today do not provide rich conceptual tasks, primarily because of the cost of scoring more open-ended problem types. The cost of developing and field testing more challenging items is also a limiting factor. Choosing an interim test that is "aligned" with a district's curriculum should ensure adequate content coverage. Unfortunately, the meaning of the term *alignment* has been debased so that items on the test can be mapped to the list of content standards, but they do not necessarily reflect the full range of cognitive competencies implied by the standards.

Districts and teachers are advised to conduct their own evaluations of interim tests and item pools to determine whether test content goes beyond rote-level knowledge and formulaic problem types.

Assessments should be timed to be instructionally-linked or instructionally-embedded. As noted earlier, timing is one of the key dimensions on which formative program evaluation instruments and formative assessments differ. To be formative, assessment insights must be used immediately as part of the instructional process; For example, a teacher sees that several students are confused and intervenes immediately, or a student receives feedback in a writing conference and works to rewrite his essay accordingly. Formative assessment is effective, then, when it is timed so that the information can be used. Comments on a term paper, for example, are not formative if students do not have the opportunity to use feedback to improve the particular piece of work or a subsequent assignment.

Interim assessments are not a part of on-going instruction but they can be effective as program evaluation tools if they are instructionally-linked. In other words, the objectives tested should match those taught in the preceding weeks and months. Although this may seem obvious, some interim tests are merely parallel forms of the end-of-year accountability test, and cover the same content whether they are administered in October or January. Repeat administrations of the end-of-year test is the least effective and most incoherent form of interim testing because it means that students are being tested on content that has not yet been taught. Benchmark or interim tests are more effective if they are substantively linked to instructional units and timed to be an external summative check on students' mastery of a particular unit of study. This criterion is a

reasonable “Consumer Reports” type of requirement for this new type of test product, but just like the switch from maximum horsepower to energy efficiency for automobiles, the desire for better instructional links will require some retooling by the industry.

Assessments must satisfy their respective definitions, by providing program insights or supporting learning processes. By definition, program evaluation tools and formative assessments have different purposes and their effectiveness can be judged by how well they accomplish those respective purposes. Given the extensive amount of testing that takes place in schools today, it is reasonable to require that new benchmark and interim assessments meet a cost/benefit test, i.e., the program evaluation insights gained about objectives that need to be retaught, for example, should be greater than the instructional time lost and other potential negative side effects. At a minimum, interim tests must yield new insights beyond what has been learned from the state assessment administered as part of NCLB. It is surprising that in many districts currently adopting interim assessment instruments there has not been a systematic effort to first learn as much as possible from state assessment results at the individual student level or by content strand. Similarly, formative assessments are to be judged by how well they accomplish their intended purpose and work to enhance student learning. Claims from the research literature can be used to evaluate whether formative assessment practices are working as intended. For example, is feedback provided that helps students to see how to improve performance over time? Is self-assessment used as a means to support internalization of criteria and personal ownership of the learning process?

Assessments should produce coherent improvements in teaching and learning. Ultimately the effectiveness of both program evaluation tools and formative assessment

will be determined by how well they guide efforts to improve teaching. Knowing that a student performs poorly on an interim test is hardly a new insight because teachers almost always know who their low performing students are. For an interim test to be effective, it has to provide new information that is coherent and can feasibly be acted upon by the teacher. Most significantly, it must avoid the "1000-mini-lessons" problem. Many publishers produce a class roster for teachers showing objectives mastered; and their advertisements display deceptively simple examples where only one or two students have significant gaps or the class as a whole missed only one or two objectives. For many teachers, however, these grids are actually a checkerboard of checks and zeros, and even veteran teachers may find it difficult to plan engaging lessons that will address multiple objectives. They may be tempted, instead, to gather groups of students for drill on the items missed. As suggested by both the embodiment and timing criteria above, interim assessments are likely to be the most effective as formative evaluation tools, if they are tied in a coherent way (aligned in the original sense of the term) to the district curriculum. Then using the curriculum as a guide, teachers can use interim test results formatively to see which parts of the curriculum are not working or which subgroup of students needs special help to catch up.

Because formative assessments are embedded in instruction, they should more naturally lead to coherent, theoretically sound improvements in teaching. Unlike more formal assessments intended to produce a score, formative assessment, grounded in specific instructional activities, provides much more detail as well as *qualitative* insights about students' understandings and misconceptions. For example, a typical interim

assessment might report that a student had or had not mastered the following algebra objective:

> Develop an understanding of function. Translate among verbal, tabular, graphic, and algebraic representations of functions. Identify relations and functions as linear or nonlinear. Find, identify, and interpret the slope (rate of change) and intercepts of a linear relation. Interpret and compare properties of linear functions from tables, graphs, or equations.

By contrast, formative assessment in an algebra class might occur as students are working in groups to solve problems. In conversation with one student, the teacher notes that the student is thinking about the steepness of a line in terms of its angle above the x axis, but she is not thinking about the change in y related to the change in x. The student can also give a memorized definition of slope, but has not learned what it means until the teacher asks her to show on the graph how change in y and change in x relate to the steepness of the slope. Then to make sure the student is understanding, the teacher asks a follow-up question, "So what would the change in x need to be, in order to make the slope flatter?"

Conclusion: Potential solutions for test publishers and for states and school districts

Although formative assessment and interim assessments could peacefully coexist, with each serving its respective purpose, in the current NCLB context the risk is great that interim assessments will prevent implementation of real formative assessment. Interim assessments are easier to install than classroom-based formative assessment practices. More significantly, when labeled as formative assessment, purchasing interim assessment data systems diverts attention and resources that might otherwise be directed toward professional development needed to implement formative assessment reforms.

Ideally, testing companies would stop using the term "formative assessment" to market interim and benchmark tests. Occasionally in the past, when confronted by

ethical rather than technical challenges, test publishers have taken very public ethical stands. For example, when Mehrens and Kaminski (1989) showed that using test preparation materials such as *Scoring High* was tantamount to practicing on a parallel form of the actual test, the parent company for test-maker CTB McGraw-Hill divested of its ownership of *Scoring High*. Gregory Anrig, president of ETS, refused to sell the National Teacher Examination (NTE) to states or school boards that used the test inappropriately to determine the futures of practicing teachers (Owen, 1984). Benchmark and interim assessments are immensely popular with local school boards because, in theory, they provide an early indication of what test results will be at the end of the year in time for teachers and students to do something about them. It is unlikely that this enthusiasm would abate if test publishers stopped using the term formative assessment, but-truth-in advertising would improve. Publishers can get equal mileage from concepts such as data-driven decision-making or program evaluation, without falsely promising Black and Wiliam (1998) results.

In the midst of the flurry of assessment activity related to NCLB,, states and districts want to know how best to help teachers target and improve instruction. The choice between investing in interim data systems or formative assessment is not a 50-50 proposition – whether to buy product A or buy product B. This asymmetry, in fact, makes it particularly difficult to further the use of real formative assessment. On the one hand, purchasing an interim assessment system is relatively straightforward. A school board agrees to the cost of the product plus the additional costs for technical support and for a limited amount of teacher professional development. In contrast, because real formative assessment is so entwined with instruction and pedagogical processes, much

more sustained professional development and support are needed to help teachers make more fundamental – and more effective – changes in their teaching practices. In more recent work, based on their famous review, Black and Wiliam (2004) and Black, Harrison, Lee, Marshall, and Wiliam (2003) have demonstrated directly the positive impact of using formative assessment as an instructional intervention, with an average gain in achievement across classrooms of .46 standard deviations (equivalent to an extra half grade level of growth). In contrast, as noted earlier, research on the impact of interim assessments on student achievement is not yet available.

Black and Wiliam's (2004) teacher professional development focuses on specific formative assessment strategies: questioning, feedback, sharing criteria, and student self-assessment, all of which lead to significant changes in teaching repertoires and in subsequent student learning. States and districts may not have considered investing in professional development to introduce teachers to formative assessment because they are already heavily vested in teacher professional development that is focused on implementation of new, standards-based literacy, mathematics, or science curricula. Rather than imagining that learning about formative assessment would need to be a new, entirely separate initiative, states and districts might consider building formative assessment ideas and processes into their subject-specific professional development offerings. In this way, both the theory of the reforms and the specific instructional strategies would be more coherently tied together for teachers attempting to try out these reforms for the first time. Literacy, mathematics, and science curriculum experts in each state are often deeply knowledgeable about formative assessment strategies that are uniquely tailored to the pedagogical demands of their respective disciplines. Running

records, author's chair, and conferencing are all examples of formative assessment strategies specific to literacy instruction. In mathematics, showing your solution on a white board or coming to the overhead to explain your reasoning are assessment strategies that also fit with the reform goal of developing students' abilities to communicate mathematically.

To date, districts have had the lion's share of responsibility for purchasing interim assessment data systems, and occasionally for investing in subject-specific curricular reforms with formative assessment components. This makes sense because both interim and formative assessment reforms should be implemented at the organizational level that has curricular authority. Not only can districts choose the most effective interim assessment system and formative assessment reform using the criteria developed in this chapter, they can also engage in the follow-on strategies that ensure maximum effectiveness. For example, consider the effectiveness criteria in Table 1 requiring that interim assessments should be "timed to be instructionally linked" and that formative assessment should be "curriculum-embedded both in timing and substance." In addition to picking an interim assessment product that has the capacity to be tailored to specific instructional units, districts can also foster instructional linkage and effective use of interim assessment results by convening professional development workshops focused on what to do in response to specific patterns of results. Because districts have control over curriculum, they can also support the curriculum-embedded power of formative assessment either by picking rich curricular materials in the first place or by providing rich conceptual tasks to supplement more procedurally oriented traditional textbooks.

States also have a key role to play by providing leadership to help local school boards and educators understand what is at stake in choosing among myriad assessment products all promising to boost student test scores. The argument about whether benchmark and interim assessments can legitimately be called formative assessment is more than a debate among pointy-headed academicians. Understanding the difference is essential for understanding what each type of assessment can do, for investing in either type, and for making effective use of assessment results and practices once the investment has been made.

The research on formative assessment is compelling and shows us explicitly how formative assessment works to improve learning -- by helping students internalize the features of good work, by showing them specifically how to improve, by developing habits of thinking and a sense of competency, and so forth. An understanding of how these formative assessment processes are tied to standards-based reform in each of the disciplines makes it possible to coordinate and integrate reform efforts so that they need not be assembled as a laundry list of new approaches. Benchmark and interim assessments can also be very helpful to teachers as program evaluation tools and as a means to identify students who need special help, but professional development may be needed to avoid interpreting the results to mean reteach everything. States should also be alert to the ways that interim and benchmark systems may exacerbate the problems of teaching the test. The literature on test-score inflation has taught us not to celebrate dramatic test score gains until their credibility has been assured by demonstrations of student competencies beyond overly-practiced, multiple-choice formats.

Note

I wish to thank Sara Y. Bryant for her assistance in collecting examples of products advertised as formative assessments, for constructing the sample item in Figure 3, and for thoughtful comments in response to earlier versions of the paper.

References

Allal, L., & Lopez, L. M. (2005). Formative assessment of learning: A review of publications in French. In Office of Economic Co-operation and Development, *Formative assessment: Improving learning in secondary classrooms*. Paris: OECD Publishing.

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

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

Black, P., & Wiliam, D. (2004). The formative purpose: Assessment must first promote learning. In M. Wilson (Ed.), *Towards coherence between classroom assessment and accountability: 103rd yearbook of the National Society for the Study of Education*, Part II (pp. 20-50).

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2003). *Assessment for learning: Putting it into practice*. Maidenhead, Berkshire, England: Open University Press.

Brown, J. S., Collins, A., & Duguid, P. (1989). Situated cognition and the culture of learning. *Educational Researcher*, *18*(1), 32-42.

Cowie, B. & Bell, B. (1999). A model of formative assessment in science education, *Assessment in Education*, *6*(1), 101-116.

Crooks, T. J. (1988). The impact of classroom evaluation practices on students. *Review of Educational Research*, *58*(4) 438-481.

Dweck, C. S. (2002). Messages that motivate: How praise molds students' beliefs, motivation, and performance (in surprising ways). In J. Aronson (Ed.), *Improving academic achievement: Classic and contemporary lessons from psychology*. New York: Academic Press.

Kahl, S. (2005a, October 26). Where in the world are formative tests? Right under your nose! *Education Week*, *25*(9), p. 38.

Kahl, S. (2005b, November 30). Coming to terms with assessment. *Education Week*, *25*(13), p.26.

Kahl, S. (2006a, January 25). Helping Teachers make the connection between assessment and instruction. *Education Week*, *25*(30), p.27.

Kahl, S. (2006b, February 22). Beware of quick-fix tests and prescriptions. *Education Week*, *25*(24), p. 31.

Kahl, S. (2006c, April 26). Self-directed learning *plus* formative assessment *equals* individualized instruction. *Education Week*, *25*(33), p. 29.

Kluger, A. N., & DeNisi, A. (1996). The effect of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, *119*, 254-284.

Koczor, M.L. (1984). *Effects of varying degrees of instructional alignment in posttreatment tests on mastery learning tasks of fourth grade children*. Unpublished doctoral dissertation, University of San Francisco, CA.

Koretz, D., Linn, R. L., Dunbar, S.B., & Shepard, L. A. (1991, April). *The effects of high-stakes test: Preliminary evidence about generalization across tests*. Presented at the annual meetings of the American Educational Research Association and the National Council on Measurement in Education, Chicago.

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

Madaus, G. F., West, M. M., Harmon, M. C., Lomax, R. G., & Viator, K. A. (1992). *The influence of testing on teaching math and science in Grades 4-12: Executive summary*. Chestnut Hill, MA: Center for the Study of Testing, Evaluation, and Educational Policy, Boston College.

Mehrens, W. A., & Kaminski, J. (1989). Methods for improving standardized test scores: Fruitful, fruitless, or fraudulent?, *Educational Measurement: Issues and Practice, 8*(1), 14-22.

Niemi, D., Vallone, J., & Vendlinski, T. (2006). The power of big ideas in mathematics education: Development and pilot testing of POWERSOURCE Assessments, CSE Report 697. Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Office of Economic Co-operation and Development. (2005). *Formative assessment: Improving learning in secondary classrooms*. Paris: OECD Publishing.

Owen, D. (1984). Testing teachers, *APF Reporter*, *7*(3), 1-5.

Palincsar, A. S., & Brown, A. L. (1984). Reciprocal teaching of comprehension-fostering and comprehension-monitoring activities. *Cognition and Instruction, 1*(2), 117-175.

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.

Rogoff, B. (1990). *Apprenticeship in thinking: Cognitive development in social context*. New York: Oxford University Press.

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

Scriven, M. (1967). The methodology of evaluation, *AERA Monograph Series on Evaluation, 1*, 39-83.

Shepard, L. A. (1993). Evaluating test validity. In L. Darling-Hammond (Ed.), *Review of Research in Education* (Vol. 19, pp. 405-450). Washington, DC: American Educational Research Association.

Shepard, L. A. (1997). *Measuring achievement: What does it mean to test for robust understanding? William H. Angoff Memorial Lecture Series*. Princeton, NJ: Educational Testing Service.

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

Shepard, L., Hammerness, K., Darling-Hammond, L., Rust, F. (2005). Assessment. In L. Darling-Hammond & J. Bransford (Eds.), *Preparing teachers for a changing*

world: What teachers should learn and be able to do (pp. 275-326). San Francisco: Jossey-Bass.

Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans, *Journal of Personality and Social Psychology*, *69*(5), 797-811.

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

Stipek, D. J. (1996). Motivation and instruction. In D. C. Berliner & R. C. Calfee (Eds.), *Handbook of Educational Psychology* (pp. 85-113). New York: Simon & Schuster Macmillan.

U.S. Congress, Office of Technology Assessment. (1992). *Testing in American schools: Asking the right questions* (OTA-SET-519). Washington, DC: U. S. Government Printing Office.

White, B. Y., & Frederiksen, J. R. (2000). Metacognitive facilitation: An approach to making scientific inquiry accessible to all. In J. Minstrell & E. van Zee (Eds.), *Inquiring into inquiry learning and teaching in science* (pp. 33-370). Washington, DC: American Association for the Advancement of Science.

Wiggins, G. (1998). *Educative assessment: Designing assessments to inform and improve student performance*. San Francisco: Jossey-Bass.

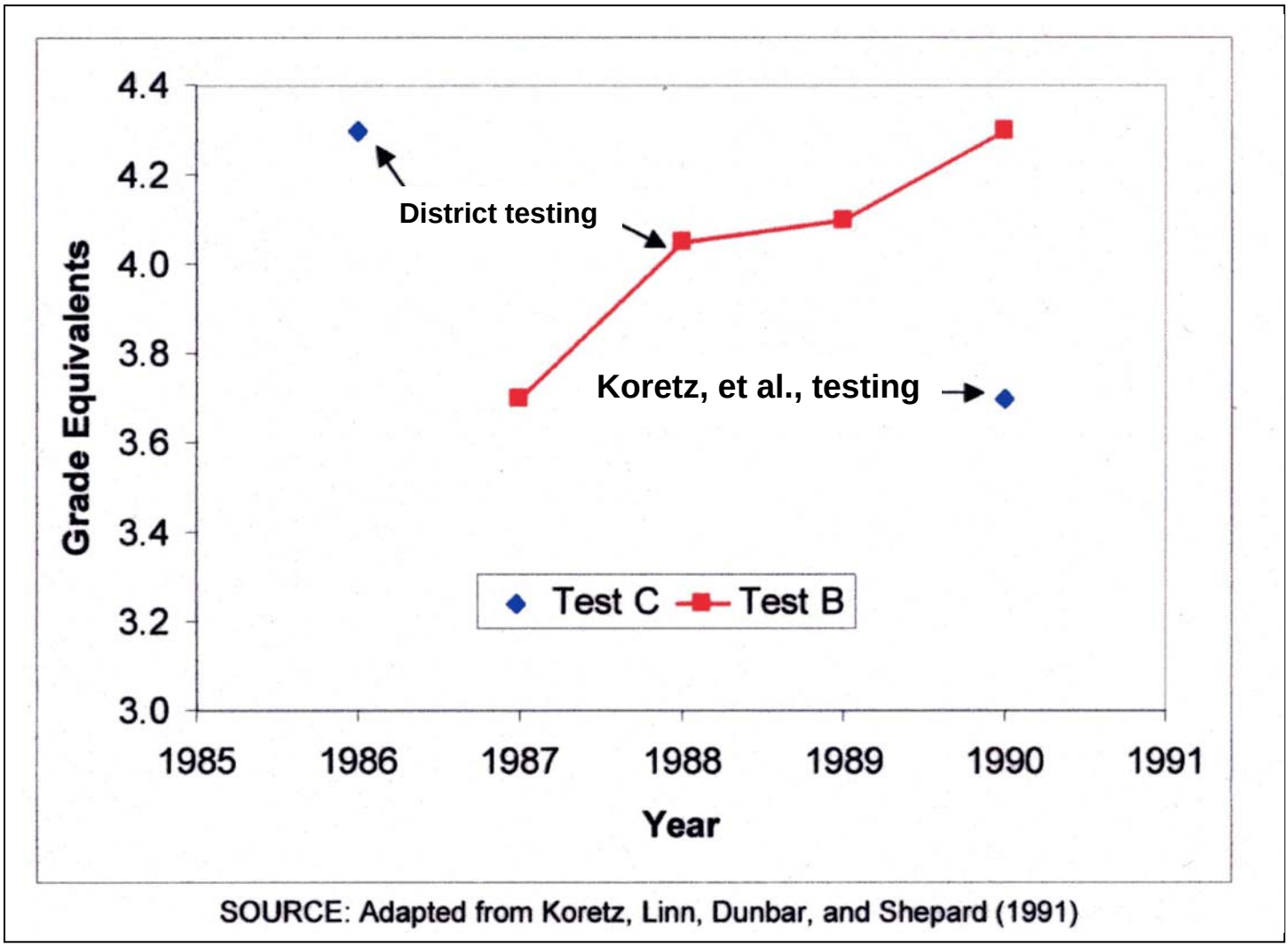
Table 1.

Criteria for effective interim assessments and formative assessment

Criteria for Effective Interim Assessments	Criteria for Effective Formative Assessment
• More than simplistic alignment, they must "embody" learning goals.	• More than simplistic alignment, tasks must "embody" learning goals.
• They should be timed to be instructionally linked.	• It should be curriculum-embedded (both in timing and substance). Tasks should be instructional tasks to provide insights about learning as it is occurring.
• They should meet a cost/benefit test, i.e., instructional insights must be greater than instructional time lost and negative side effects. (At a minimum they must yield new insights beyond NCLB accountability test.)	• By definition, it must enable the supportive learning processes invoked in the formative assessment literature.
• Instructional insights should lead to coherent, theoretically sound improvements in teaching.	• Instructional insights should lead to coherent, theoretically sound improvements in teaching.

FIG. 1

Performance on familiar and unfamiliar standardized tests with very similar content and format



SOURCE: Adapted from Koretz, Linn, Dunbar, and Shepard (1991)

FIG. 2

Examples of items used in both teaching materials and in testing materials in the study by Koczor (1984)

DIRECTIONS: Write the Arabic numerals for the following Roman numerals.

1. XXI _____ **5. DCLXXXIX** _____

2. LXVIII _____ **6. DCCLIX** _____

3. XIV _____ **7. MCMLI** _____

DIRECTIONS: Write the Roman numerals for the following Arabic numerals.

1. 11 _____ **5. 546** _____

2. 20 _____ **6. 417** _____

3. 89 _____ **7. 1608** _____

FIG. 3

An example constructed to imitate items on a "Formative Assessment" website

Cats are fun animals to have as pets. Read the following passage and fill in the missing words.

Cats make great pets. They are soft and cuddly. They play with toys. Sometimes they get tired and need to rest. They ___{1}___. There are many kinds of cats. Cats have four legs.

Cats like to play outside. Dogs chase them. Cats have to climb trees to get away. They get ___{2}___.

{1}	O	bite	{2}	O scared
	O	sleep		O hurry
	O	ski		O cold
	O	close		O hungry

Note: These reading passages resemble a cloze technique but, in fact, are based on unnatural paragraphs. Instead of inference, they invite learning the strategy that the answer is nearly always in the sentence before.

FIG. 4

A recent example of materials used to motivate students to raise their test scores

Fall Conference Sheet and Self-Assessment

Student Name:_____ Date:_____

Math

My fall RIT score is _____.

In the spring, my target goal will be _____.

Here is how I rate myself.

1. Paying attention in class excellent/good/okay/need to improve
2. Effort on homework excellent/good/okay/need to improve
3. Effort on tests excellent/good/okay/need to improve
4. Class participation excellent/good/okay/need to improve
5. Behavior excellent/good/okay/need to improve
6. Attendance excellent/good/okay/need to improve