# The Role of Classroom Assessment in Teaching and Learning

CSE Technical Report 517

Lorrie A. Shepard
CRESST / University of Colorado at Boulder

February 2000

Center for the Study of Evaluation
National Center for Research on Evaluation,
Standards, and Student Testing
Graduate School of Education & Information Studies
University of California, Los Angeles
Los Angeles, CA 90095-1522
(310) 206-1532

and

Center for Research on Education, Diversity and Excellence
University of California, Santa Cruz
1156 High Street
Santa Cruz, CA 95064
(408) 459-3500

Project 2.4 Assessment of Language Minority Students Lorrie Shepard, Project Director
CRESST / University of Colorado at Boulder



The work reported herein was supported in part by grants from the Office of Educational Research and Improvement, U.S. Department of Education to the Center for Research on Evaluation, Standards, and Student Testing (CRESST) (Award No. R305B60002) and to the Center for Research on Education, Diversity and Excellence (CREDE)(Award No. R306A60001).

The findings and opinions expressed in this report do not reflect the positions or policies of the National Institute on Student Achievement, Curriculum, and Assessment, the Office of Educational Research and Improvement, or the U.S. Department of Education.

# THE ROLE OF CLASSROOM ASSESSMENT IN TEACHING AND LEARNING

**Lorrie A. Shepard**[1]
**CRESST/University of Colorado at Boulder**

## Introduction and Overview

Historically, because of their technical requirements, educational tests of any importance were seen as the province of statisticians and not that of teachers or subject matter specialists. Researchers conceptualizing effective teaching did not assign a significant role to assessment as part of the learning process. The past three volumes of the *Handbook of Research on Teaching*, for example, did not include a chapter on classroom assessment nor even its traditional counterpart, tests and measurement. Achievement tests were addressed in previous handbooks but only as outcome measures in studies of teaching behaviors. In traditional educational measurement courses, preservice teachers learned about domain specifications, item formats, and methods for estimating reliability and validity. Few connections were made in subject matter methods courses to suggest ways that testing might be used instructionally. Subsequent surveys of teaching practice showed that teachers had little use for statistical procedures and mostly devised end-of-unit tests aimed at measuring declarative knowledge of terms, facts, rules, and principles (Fleming & Chambers, 1983).

The purpose of this chapter is to develop a framework for understanding a reformed view of assessment, where assessment plays an integral role in teaching and learning. If assessment is to be used in classrooms to help students learn, it must be transformed in two fundamental ways. First, the content and character of assessments must be significantly improved. Second, the gathering and use of assessment information and insights must become a part of the ongoing learning process. The model I propose is consistent with current assessment reforms being advanced across many disciplines (e.g., International Reading Association / National Council of Teachers of English Joint Task Force on Assessment, 1994; National Council for the Social Studies, 1991; National Council of Teachers of Mathematics,

[1] I wish to thank Margaret Eisenhart, Kenneth Howe, Gaea Leinhardt, Richard Shavelson, and Mark Wilson for their thoughtful comments on drafts of this chapter.

1995; National Research Council, 1996). It is also consistent with the general argument that assessment content and formats should more directly embody thinking and reasoning abilities that are the ultimate goals of learning (Frederiksen & Collins, 1989; Resnick & Resnick, 1992). Unlike much of the discussion, however, my emphasis is not on external accountability assessments as indirect mechanisms for reforming instructional practice; instead, I consider directly how classroom assessment practices should be transformed to illuminate and enhance the learning process. I acknowledge, though, that for changes to occur at the classroom level, they must be supported and not impeded by external assessments.

The changes being proposed for assessment are profound. They are part of a larger set of changes in curriculum and theories of teaching and learning, which many have characterized as a paradigm change. Constructivist learning theory, invoked throughout this volume, is at the center of these important changes and has the most direct implications for changes in teaching and assessment. How learning occurs, in the minds and through the social experience of students, however, is not the only change at stake. Equally important are epistemological changes that affect both methods of inquiry and conceptions of what it means to know in each of the disciplines. Finally, there is a fundamental change to be reckoned with regarding the diverse membership of the scholarly community that is developing this emergent paradigm. It includes psychologists, curriculum theorists, philosophers, experts in mathematics, science, social studies, and literacy education, researchers on teaching and learning to teach, anthropologists, and measurement specialists. How these perspectives come together to produce a new view of assessment is a key theme throughout this chapter.

The chapter is organized as follows. Three background sections describe first, underlying curriculum and psychological theories that have shaped methods of instruction, conceptions of subject matter, and methods of testing for most of this century; second, a conceptual framework based on new theories and new relationships among curriculum, learning theory, and assessment; and third, the connections between classroom uses of assessment and external accountability systems. In the fourth and fifth sections, I elaborate a model for classroom assessment based on social-constructivist principles, arguing, respectively, for the substantive reform of assessment and for its use in classrooms to support learning. In the concluding section, I outline the kinds of research studies that will be needed to help realize a reformed vision of classroom assessment.

## Historical Perspectives: Curriculum, Psychology, and Measurement

Assessment reformers today emphasize the need for a closer substantive connection between assessment and meaningful instruction. They are reacting against documented distortions in recent decades where teachers in the contexts of high-stakes accountability testing have reshaped instructional activities to conform to both the content and format of external standardized tests, thereby lowering the complexity and demands of the curriculum and at the same time reducing the credibility of test scores. In describing present-day practice, for example, Graue (1993) suggests that assessment and instruction are "conceived as curiously separate," a separation which Graue attributes to technical measurement concerns. A longer-term span of history, however, helps us to see that those measurement perspectives, now felt to be incompatible with instruction, came from an earlier, highly consistent theoretical framework in which conceptions of "scientific measurement" were closely aligned with curricula underpinned by behaviorist learning theory and directed at social efficiency.

Figure 1 was devised to show in broad brush the shift from the dominant twentieth-century paradigm (on the left) to an emergent, constructivist paradigm (on the right), in which teachers' close assessment of students' understandings, feedback from peers, and student self-assessment are a part of the social processes that mediate the development of intellectual abilities, construction of knowledge, and formation of students' identities. The middle portion of the figure, intended to represent present-day teaching practices, adapts a similar figure from Graue (1993) showing a sphere for instruction entirely separate from the sphere for assessment. According to this model, instruction and assessment are guided by different philosophies and are separated in time and place. Even classroom assessments, nominally under the control of teachers, may be more closely aligned with external tests than with day-to-day instructional activities. Although there is ample evidence that the intermediate model describes current practice, this model has no theoretical adherents. The best way to understand this mismatch is to see that instructional practices (at least in their ideal form) are guided by the new paradigm, while traditional testing practices are held over from the old.

It is important to know where traditional views of testing came from and to appreciate how tightly entwined they are with past models of curriculum and instruction, because new theories are defined and understood in contrast to prior

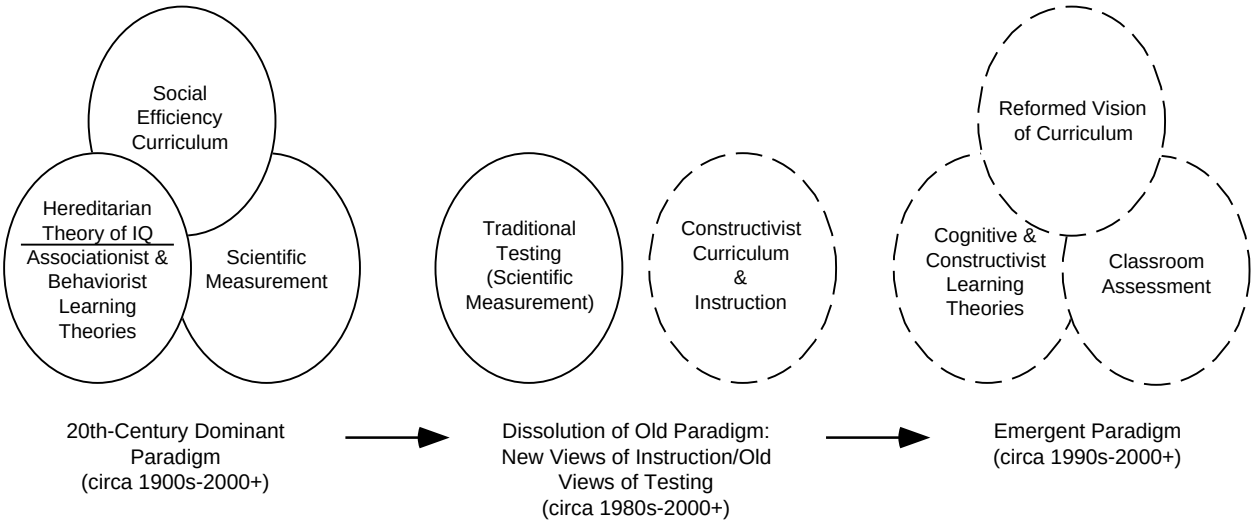



*Figure 1.* An historical overview illustrating how changing conceptions of curriculum, learning theory, and measurement explain the current incompatibility between new views of instruction and traditional views of testing.

theories. More importantly, however, dominant theories of the past continue to operate as the default framework affecting current practices and perspectives. Belief systems of teachers, parents, and policy makers are not exact reproductions of formal theories. They are developed through personal experience and from popular cultural beliefs. Nonetheless, formal theories often influence implicit theories held and acted upon by these various groups; and because it is difficult to articulate or confront formal theories once they have become a part of the popular culture, their influence may be potent but invisible long after they are abandoned by theorists. For example, individuals who have been influenced by behaviorist theories, even if not identified as such, may believe that learning in an academic subject is like building a brick wall, layer by layer. They may resist reforms intended to show connections between multiplication and addition or between patterns and functions because they disrupt the traditional sequencing of topics. Most importantly, adherence to behaviorist assumptions leads to the postponement of instruction aimed at thinking and reasoning until after basic skills have been mastered.

A more elaborated version of the twentieth-century dominant paradigm is presented in Figure 2. The central ideas of social efficiency and scientific management were closely linked, in the first case, to hereditarian theories of individual differences and, in the second case, to associationist and behaviorist

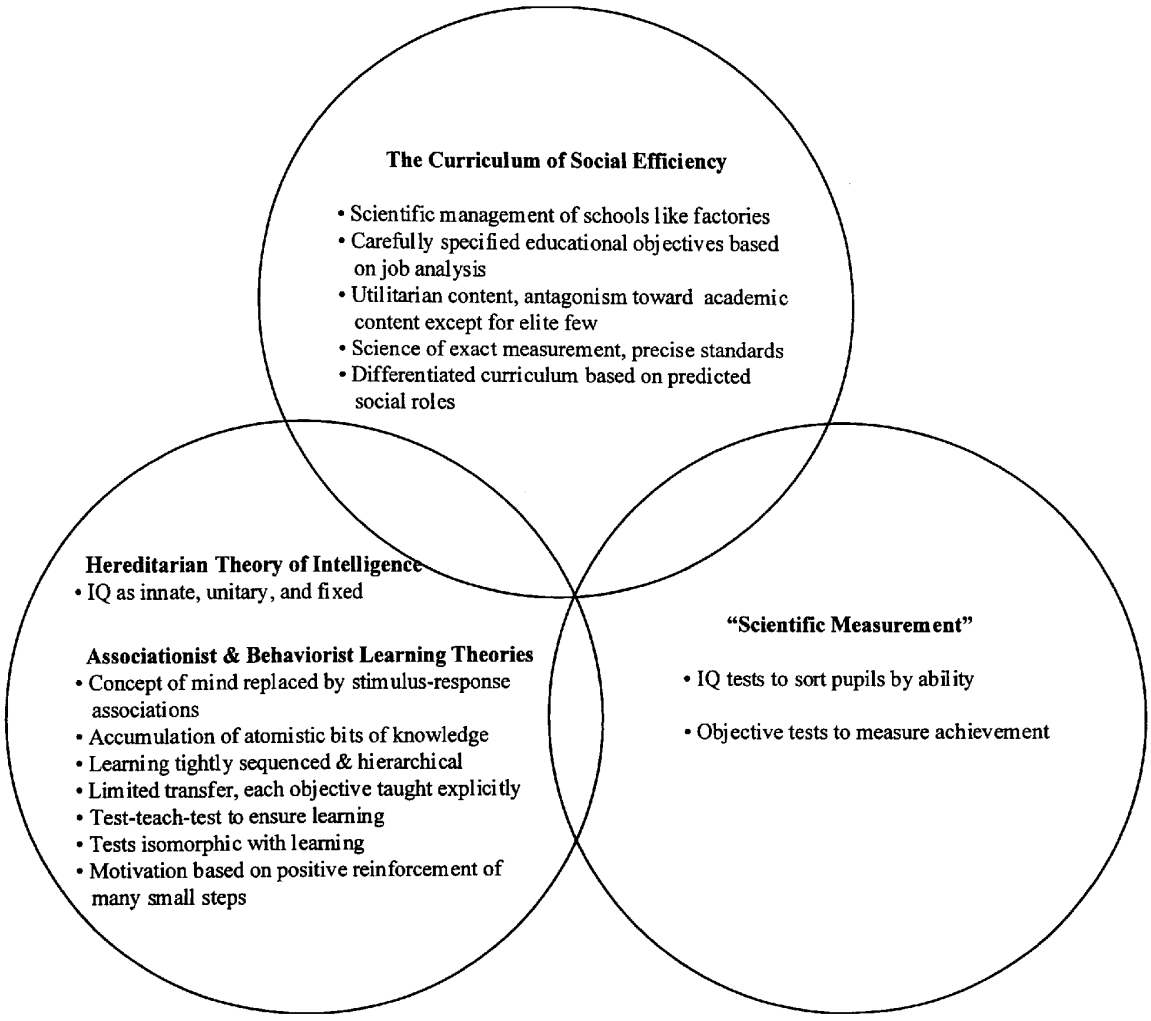



*Figure 2.* Interlocking tenets of curriculum theory, psychological theories, and measurement theory characterizing the dominant twentieth-century paradigm.

learning theories,[2] which saw learning as the accumulation of stimulus-response associations. These respective psychological theories were, in turn, served by scientific measurement of ability and achievement. The interlocking components of this historic and extant paradigm are summarized in the following sections with particular attention to the legacy of these ideas for classroom assessment practices.

[2] This is not to suggest that hereditarian and behaviorist theories were compatible with each other. Behaviorists strongly favored environmental over genetic explanations for human ability. However, these psychological theories co-existed throughout the century, and both exerted great influence over educational and testing practices.

## The Curriculum of Social Efficiency

In the early 1900s, public concerns about education were shaped by industrialization, fears of the loss of community, and the need to absorb and "Americanize" large numbers of immigrants (Callahan, 1962; Kliebard, 1995; Tyack, 1974). The social efficiency movement grew out of the belief that science could be used to solve these problems. It was led by sociologists and psychologists but was equally embraced by business leaders and politicians.

According to this theory, modern principles of scientific management, intended to maximize the efficiency of factories, could be applied with equal success to schools. This meant taking Taylor's example of a detailed analysis of the movements performed by expert bricklayers and applying similar analyses to every vocation for which students were being prepared (Kliebard, 1995). Then, given the new associationist or connectionist psychology with its emphasis on fundamental building blocks, every step would have to be taught specifically. Precise standards of measurement were required to ensure that each skill was mastered at the desired level. And because it was not possible to teach every student the skills of every vocation, scientific measures of ability were also needed to predict one's future role in life and thereby determine who was best suited for each endeavor. For John Bobbitt, a leader in the social efficiency movement, a primary goal of curriculum design was the elimination of waste (1912), and it was wasteful to teach people things they would never use. Bobbitt's most telling principle was that each individual should be educated "according to his capabilities." These views led to a highly differentiated curriculum and a largely utilitarian one that disdained academic subjects for any but college preparatory students.

Thus, scientific management and social efficiency launched two powerful ideas: the need for detailed specifications of objectives and tracking by ability. Although social efficiency began to lose popularity among sociologists and psychologists after the 1930s, these ideas continued to have profound influence on educational practice because they were absorbed in eclectic versions of curricula, such as life adjustment education and work-oriented curriculum, that had strong appeal with school administrators (Kliebard, 1995). These ideas contributed to and were buttressed by concomitant developments in psychology and measurement.

## Hereditarian Theory and IQ Testing

When intelligence tests were brought to the United States in the early 1900s, their interpretation and use were strongly influenced by the eugenics movement and prevalent racist beliefs. Binet (1909, pp. 100-101), who had developed the first IQ tests in France, believed in "the educability of the intelligence" and deplored the "brutal pessimism" of those who thought it to be a fixed quantity. His program of "mental orthopedics" was intended to improve the use of mental resources and thus help the student become more intelligent than before. In contrast, American psychologists such as Terman, Goddard, and Yerkes promoted IQ test results as a scientifically exact measure of a fixed trait that conformed to the laws of Mendelian genetics.

In a climate of fear about degeneration of the race and the threat of immigration from southern and eastern Europe (Cronbach, 1975; Gould, 1981), most American psychologists emphasized the biological nature of IQ. Goddard (1920, p. 1) referred to intelligence as a "unitary mental process . . . which is inborn" and "determined by the kind of chromosomes that come together with the union of the germ cells." Terman (1906, p. 68) asserted without evidence his belief in "the relatively greater importance of endowment over training as a determinant of an individual's intellectual rank among his fellows" (cited in Gould, 1981, p. 175). Both men also pursued the exact ordering of individuals on the scale of IQ, which they believed accounted for moral behavior as well as cognitive performance. Goddard fine-tuned distinctions among the feeble-minded, creating the categories of idiot, imbecile, and moron. Terman (1916) saw a precise deterministic relationship between IQ score and lot in life: "an IQ below 70 rarely permits anything better than unskilled labor," "the range of 70-80 is preeminently that of semiskilled labor; from 80-100 that of ordinary clerical labor" (p. 27), and so forth.

Because measured differences were taken to be innate (and because society would not agree to a program of sterilization), the only way to cope with inexorable differences in capacity was a highly differentiated curriculum. For example, having attributed the higher rate of "border-line deficiency" scores among "Spanish-Indian, Mexicans in the Southwest, and Negroes" to inherited differences that were most likely racial, Terman (1916) urged that "children of this group should be segregated in special classes and be given instruction which is concrete and practical. They cannot master abstractions, but they can often be made efficient workers, able to look out for themselves" (pp. 91-92).

These beliefs and policies were advocated almost 100 years ago, yet they continue to have a profound effect on school practices and public understandings about education. Streaming or tracking by ability began in the 1920s and has continued with only slight diminution in recent decades. As Cronbach (1975) explained, the most extreme nativist claims had received widespread attention in the popular press. In contrast, for half a century, more temperate scholarly debates about the fallibility of measures, contributions of environment, and the self-fulfilling consequences of test-based sorting were conducted out of the public eye—until Jensen's (1969) work rekindled the controversy. It was relatively late in the century before scholars or public officials gave attention to the potential harm of labeling children (Hobbs, 1975), to the inaccuracy of classifications based on single tests (Education for All Handicapped Children Act of 1975, P.L. 94-142), and to the possible ineffectiveness of special placements (Heller, Holtzman, & Messick, 1982).

Now at the end of the century, superficially at least, the tide has changed. Most scientists and educated citizens assign a much more limited role to heredity, recognize the multidimensional nature of ability, and are aware of the large effect of past learning opportunities on both test performance and future learning. Herrnstein and Murray's (1994) argument—that inherited cognitive differences between races account for apparent differences in life chances—is an obvious carrying forward of earlier ideas but no longer has support in the current scientific community. Such a summary, however, ignores the persistence of underlying assumptions in popular opinion and cultural norms.

As Wolf and Reardon (1996) point out, enduring beliefs about the fixed nature of ability and the need to segregate elite students explain why there is such a conflict in American education between excellence and equity. Although group IQ tests are no longer routinely used to determine children's capabilities, many teachers, policymakers, and lay people implicitly use family background and cultural difference as equally fixed characteristics that account for school failure (Valencia, 1997). The use of readiness measures and achievement tests to categorize students' learning capacity still has the same negative effects as tracking based on IQ, because of the assumption that students in the lower strata should receive a simplified curriculum.

More subtly perhaps, the sorting and classification model of ability testing for purposes of curriculum differentiation has left a legacy that also affects the conception of assessment within classrooms. Even when aptitude measures are

replaced by achievement tests, there is still the tendency to use test results to assign students to gross instructional categories rather than having the test tell something particular about what a student knows or how he is thinking about a problem. It is as if achievement is seen as a uni-dimensional continuum and tests are "locator" devices. In this regard, the tradition of ranking by ability became curiously entwined with lock-step assumptions about learning sequences discussed in the next section.

**Associationist and Behaviorist Learning Theories**

Edward Thorndike's (1922) associationism and the behaviorism of Hull (1943), Skinner (1938, 1954) and Gagne (1965) were the dominant learning theories for the greater part of the 20th century. Their views of how learning occurs focused on the most elemental building blocks of knowledge. Thorndike was looking for constituent bonds or connections that would produce desired responses for each situation. Similarly, behaviorists studied the contingencies of reinforcement that would strengthen or weaken stimulus-response associations. The following quotation from Skinner (1954) is illustrative:

> The whole process of becoming competent in any field must be divided into a very large number of very small steps, and reinforcement must be contingent upon the accomplishment of each step. This solution to the problem of creating a complex repertoire of behavior also solves the problem of maintaining the behavior in strength. . . . By making each successive step as small as possible, the frequency of reinforcement can be raised to a maximum, while the possibly aversive consequences of being wrong are reduced to a minimum. (p. 94)

Although it is not possible to give a full account of these theories here, several key assumptions of the behavioristic model had consequences for ensuing conceptualizations of teaching and testing: 1. Learning occurs by accumulating atomized bits of knowledge; 2. Learning is sequential and hierarchical; 3. Transfer is limited to situations with a high degree of similarity; 4. Tests should be used frequently to ensure mastery before proceeding to the next objective; 5. Tests are the direct instantiation of learning goals; and 6. Motivation is externally determined and should be as positive as possible (Greeno, Collins, & Resnick, 1996; Shepard, 1991b; Shulman & Quinlan, 1996).

Behaviorist beliefs fostered a reductionistic view of curriculum. In order to gain control over each learning step, instructional objectives had to be tightly specified just as the efficiency expert tracked each motion of the brick layer. As explained by Gagne (1965), "to 'know,' to 'understand,' to 'appreciate' are perfectly good words,

but they do not yield agreement on the exemplification of tasks. On the other hand, if suitably defined, words such as to 'write,' to 'identify,' to 'list,' do lead to reliable descriptions" (p. 43). Thus, behaviorally-stated objectives became the required elements of both instructional sequences and closely related mastery tests. Although it was the intention of behaviorists that learners would eventually get to more complex levels of thinking, as evidenced by the analysis, synthesis, and evaluation levels of Bloom's (1956) Taxonomy, emphasis on stating objectives in behavioral terms tended to constrain the goals of instruction.

Rigid sequencing of learning elements also tended to focus instruction on low-level skills, especially for low-achieving students and children in the early grades. Complex learnings were seen as the sum of simpler behaviors. It would be useless and inefficient to go on to ABC problems without first having firmly mastered A and AB objectives (Bloom, 1956). For decades, these principles undergirded each educational innovation: programmed instruction, mastery learning, objectives-based curricula, remedial reading programs, criterion-referenced testing, minimum-competency testing, and special education interventions. Only later did researchers begin to document the diminished learning opportunities of children assigned to drill-and-practice curricula in various remedial settings (Allington, 1991; Shepard, 1991a).

For all learning theories, the idea of transfer involves generalization of learning to new situations. Yet because behaviorism was based on the building up of associations in response to a particular stimulus, there was no basis for generalization unless the new situation was very similar to the original one. Therefore, expectations for transfer were limited; if a response were desired in a new situation, it would have to be taught as an additional learning goal. Cohen (1987), for example, praised the effectiveness of closely aligning tests with instruction, citing a study by Koczor (1984) in which students did remarkably better if they were taught to convert from Roman to Arabic numerals and then were tested in that same order. If groups were given "misaligned" tests, however, asking that they translate in reverse order, from Arabic to Roman numerals, the drop-off in performance was startling, from 1.10 to 2.74 standard deviations in different samples. Consistent with the behaviorist perspective, Cohen and Koczor considered Roman-to-Arabic and Arabic-to-Roman conversions to be two separate learning objectives. They were not troubled by lack of transfer from one to the other, nor did they wonder what this implied about students' understanding.

Testing played a central role in behaviorist instructional systems. To avoid learning failures caused by incomplete mastery of prerequisites, testing was needed at the end of each lesson, with reteaching to occur until a high level of proficiency was achieved. In order to serve this diagnostic and prescriptive purpose, test content had to be exactly matched to instructional content by means of the behavioral objective. Because learning components were tightly specified, there was very limited inference or generalization required to make a connection between test items and learning objectives. Behaviorists worked hard to create a low-inference measurement system so that if students could answer the questions asked, it was proof that they had fully mastered the learning objective.

The belief that tests could be made perfectly congruent with the goals of learning had pervasive effects in the measurement community despite resistance from some. For decades, many measurement specialists believed that achievement tests only required content validity evidence and did not see the need for empirical confirmation that a test measured what was intended. Behavioristic assumptions also explain why, in recent years, advocates of measurement-driven instruction were willing to use test scores themselves to prove that teaching to the test improved learning (Popham, Cruse, Rankin, Sandifer, & Williams, 1985), while critics insisted on independent measures to verify whether learning gains were real (Koretz, Linn, Dunbar, & Shepard, 1991).

Behaviorist viewpoints also have implications for assessment in classrooms. For example, when teachers check on learning by using problems and formats identical to those used for initial instruction, they are operating from the low-inference and limited transfer assumptions of behaviorism. For most teachers, however, these beliefs are not explicit, and, unlike Koczor and Cohen in the example above, most teachers have not had the opportunity to consider directly whether a student “really knows it” if he can solve problems only when posed in a familiar format.

Behaviorism also makes important assumptions about motivation to learn. It assumes that individuals are externally motivated by the pursuit of rewards and avoidance of punishments. In particular, Skinner’s (1954) interpretation of how reinforcement should be used to structure learning environments had far-reaching effects on education. As expressed in the earlier quotation, it was Skinner’s idea that to keep the learner motivated, instruction should be staged to ensure as much success as possible with little or no negative feedback. It is this motivational purpose

as much as the componential analysis of tasks that led to the idea of little steps. In Individually Prescribed Instruction (Education U.S.A., 1968), for example, lessons were designed around skills that the average student could master in a single class period.

**"Scientific Measurement" and Objective Examinations**

It is no coincidence that Edward Thorndike was both the originator of associationist learning theory and the "father" of "scientific measurement"[3] in education (Ayers, 1918). Thorndike and his students fostered the development and dominance of the "objective" test, which has been the single most striking feature of achievement testing in the United States from the beginning of the century to the present day. Recognizing the common paternity of the behaviorist learning theory and objective testing helps us to understand the continued intellectual kinship between one-skill-at-a-time test items and instructional practices aimed at mastery of constituent elements.

Borrowing the psychometric technology of IQ tests, objective measures of achievement were pursued with the goal of making the study of education more scientific. According to Ralph Tyler (1938), "The achievement-testing movement provided a new tool by which educational problems could be studied systematically in terms of more objective evidence regarding the effects produced in pupils" (p. 349). Objective tests were also promoted for classroom use as a remedy for embarrassing inconsistencies in teachers' grading practices documented by dozens of research studies. In one classic study, for example, the same geometry paper was distributed to 116 high school mathematics teachers and received percentage grades ranging from 28 to 92 (Starch & Elliott, 1913). Many of the arguments made in favor of teacher-developed objective tests suggest issues that are still relevant today. For example, in addition to solving the problem of grader subjectivity, discrete item types also allowed "extensive sampling" (better content coverage) and "high reliability per unit of working time" (Ruch, 1929, p. 112). The emphasis on reliability, defined as the consistency with which individuals are ranked, followed naturally from the application to achievement tests of reliability and validity coefficients developed in the context of intelligence testing.

[3] Note that "scientific measurement" was the term used historically but is based on the conception of science and scientific inquiry held at the turn of the previous century. The honorific label does not imply that early achievement tests were scientific according to current-day standards for scientific inquiry.

Examples from some of the earliest “standard” tests and objective-type classroom tests are shown in Figure 3. Looking at any collection of tests from early in the century, one is immediately struck by how much the questions emphasize rote recall. To be fair this was not a distortion of subject matter caused by the adoption of objective-item formats. Rather, the various recall, completion, matching, and multiple-choice test types fit closely with what was deemed important to learn in the first part of the century. Nonetheless, once knowledge of curriculum became encapsulated and represented by these types of items, it is reasonable to say that these formats locked-in and perpetuated a particular conception of subject matter. Also shown in Figure 3 is an example of the kind of essay question asked alongside of objective questions in a 1928 American history test. Little data exist to tell us how often the two types of examinations were used or to document their relative quality. For example, Ruch (1929) defended his new-type objective examination against the complaint that it only measured memory, by saying that “teachers and educators pay lip service to the thought question and then proceed merrily to ask pupils to `Name the principal products of New England’ or to `List the main causes of the Revolutionary War’” (p. 121).

Present-day calls for assessment reform are intended to counteract the distorting effects of high-stakes accountability tests. Under pressure to improve scores, teachers have not only abandoned untested content but have reshaped their classroom instruction to imitate the format of standardized tests (Darling-Hammond & Wise, 1985; Madaus, West, Harmon, Lomax, & Viator, 1992; Shepard & Dougherty, 1991; Smith et al., 1990). By hearkening to a day before standardized achievement measures had such serious consequences, reformers seem to imply that there was once a golden era when teachers used more comprehensive and challenging examinations to evaluate student knowledge. A longer-term historical view suggests, however, that the current propensity to focus on low-level skills is merely an exaggeration of practices that have continued without interruption throughout the twentieth century.

There has been a long-term, abiding tendency to think of subject matter in a way that is perfectly compatible with recall-oriented test questions. The 1946 *National Society for the Study of Education Yearbook*, for example, was devoted to “The Measurement of Understanding.” In introducing the volume, William Brownell

**New Stone Reasoning Tests in Arithmetic, 1908**

1. James had 5 cents. He earned 13 cents more and then bought a top for 10 cents. How much money did he have left? *Answer*: _____

2. How many oranges can I buy for 35 cents when oranges cost 7 cents each? *Answer*: _____

**Sones-Harry High School Achievement Test, Part II, 1929**

1. What instrument was designed to draw a circle?...(_____)1
2. Write "25% of" as "a decimal times." ..... ..... . (_____)2
3. Write in figures: one thousand seven and four hundredths..... ..... ..... ..... ..... ...(_____)3

**The Modern School Achievement Tests, Language Usage**

1. I borrowed a pen a. off / b. off of / c. from my brother. _____

2. Every student must do a. your / b. his / c. their best. _____

3. He a. has got / b. has / c. has gotten his violin with him. _____

**The Barrett-Ryan Literature Test: Silas Marner**

1. ( ) An episode that advances the plot is the--a. murdering of a man. b. kidnapping of a child. c. stealing of money. d. fighting of a duel.

2. ( ) Dolly Winthrop is--a. an ambitious society woman. b. a frivolous girl. c. a haughty lady. d. a kind, helpful neighbor.

3. ( ) A chief characteristic of the novel is--a. humorous passages. b. portrayal of character. c. historical facts. d. fairy element.

**Examples of True-False Objective Test (Ruch, 1929)**

1. Tetanus (lockjaw) germs usually enter the body through open wounds. *True* *False*

2. Pneumonia causes more deaths in the United States than tuberculosis. *True* *False*

3. White blood corpuscles are more numerous than are the red ones. *True* *False*

**Examples of Best-Answer Objective Test (Ruch, 1929)**

1. Leguminous plants play an important role in nature because:
   Bacteria associated with their roots return nitrogen to the soil.
   They will grow on soil too poor to support other crops.
   The economic value of the hay crop is very large.

2. The best of these definitions of photosynthesis is:
   The action of sunlight on plants.
   The process of food manufacture in green plants.
   The process by which plants give off oxygen.

**American History Examination, East High School**
**Sam Everett and Effey Riley, 1928**

I. Below is a list of statements. Indicate by a cross (X) after it, each statement that expresses a social heritage of the present-day American nation.
Place a (0) after each statement that is not a present-day social heritage of the American nation.
1. Americans believe in the ideal of religious toleration. __
2. Property in land should be inherited by a man's eldest son. __
3. Citizens should have the right to say what taxes should be put upon them. __

II. To test your ability to see how an intelligent knowledge of past events help us to understand present-day situations, and tendencies.
(Note: Write your answer in essay form on a separate sheet of paper.)
Some one has said that we study the past relationships in American life in order to be able to understand the present in our civilization and that we need to understand the present so as to influence American national development toward finder things.

*State your reasons for every position assumed.*

4. Take some economic fact or group of facts in American History about which we have studied and briefly show what seems to you to be the actual significance of this fact in the past, present and future of America.
5. Show this same three-fold relationship using some political fact or facts.
6. Show this same three-fold relationship using a religious fact or facts.

III. 7. The rise of manufacturing in New England was greatly aided by the fact that their physical environment furnished: (a) cold temperature, (b) all kinds of raw materials, (c) many navigable rivers, (d) easy communication with the West, (e) water power.
8. The wealth of Colonial South Carolina came chiefly from: (a) rice, (b) tobacco, (c) cotton, (d) furs, (e) wheat.

IV. 9. The Constitution represents a series of compromises rather than a document considered perfect by its signers. R ? W
10. Since a great number of the colonists had come to America for political freedom and to found governments on democratic ideals, full manhood suffrage was granted in every colony from the first. R ? W
11. The major reason why slavery did not flourish in the New England colonies was because it was not a good financial proposition. R ? W

V. 12. As part of your education you have been studying in American history about the Constitutional Convention. Has the study of that historical event meant to you simply memorizing a list of the facts or events, --or has it given you (a) insight into the significance of certain decisions made by the men of the Constitutional Convention; (b) ability to evaluate certain clauses of our Constitution; (c) ability to decide whether our forefathers intended to give us a democracy or not?

If you have gained any of these three things, will you try to show that you have acquired them through use of practical illustrations in each of the three cases?

*Figure 3.* Examples from some of the earliest twentieth-century "standard" tests and objective-type classroom tests.

explained that techniques for measuring factual knowledge and skills were well worked out and used in evaluation and in teaching while "understanding," "meaningful learning," and "the higher mental processes" were neglected (Brownell, 1946, p. 2). In a 1967 national survey, Goslin reported that 67% of public secondary teachers and 76% of elementary teachers reported using objective items "frequently," "most of the time," or "always." Many teachers used both types of questions, but "objective" questions were used more often than essays. In recent decades, analysts have documented the reciprocal influence of textbooks on standardized tests and standardized tests on textbooks (Tyson-Bernstein, 1988), which has also served to carry forward a conception of subject matter that is mostly vocabulary, facts, and decontextualized skills.

The dominance of objective tests in classroom practice has affected more than the form of subject matter knowledge. It has also shaped beliefs about the nature of evidence and principles of fairness. In a recent assessment project (Shepard, 1995), for example, where teachers were nominally seeking alternatives to standardized tests, teachers nonetheless worked from a set of beliefs consistent with traditional principles of scientific measurement. As documented by Bliem and Davinroy (1997), assessment was seen as an official event. To ensure fairness, teachers believed that assessments had to be *uniformly* administered; therefore, teachers were reluctant to conduct more intensive individualized assessments with only below-grade-level readers. Because of the belief that assessments had to be targeted to a specific instructional goal, teachers felt more comfortable using two separate assessments for separate goals, a notation system known as "running records" to assess fluency and written summaries to assess comprehension, rather than, say, asking students to retell the gist of a story in conjunction with running records. Most significantly, teachers wanted their assessments to be "objective"; they worried often about the subjectivity involved in making more holistic evaluations of student work and preferred formula-based methods, such as counting miscues, because these techniques were more "impartial."

Any attempt to change the form and purpose of classroom assessment to make it more fundamentally a part of the learning process must acknowledge the power of enduring and hidden beliefs. I have suggested that the present dissonance between instruction and assessment arises because of the misfit between old views of testing and a transformed vision of teaching. However, even reformed versions of instruction have only begun to be implemented. As many studies of teacher change

and attempted curriculum reform have documented, all three parts of the old paradigm—social efficiency, behaviorism, and scientific measurement—continue to provide a mutually reinforcing set of ideas that shape current thinking and practice.

## Conceptual Framework: New Theories of Curriculum, Learning, and Assessment

In order to develop a model of classroom assessment that supports teaching and learning according to a constructivist perspective, it is important to see how a reconceptualization of assessment follows from changes in learning theory and from concomitant changes in epistemology and what it means to know in the disciplines. Figure 4 summarizes key ideas in an emergent, constructivist paradigm. According to constructivist theory, knowledge is neither passively received nor mechanically reinforced; instead learning occurs by an active process of sense making. The three-part figure was developed in parallel to the three-part dominant paradigm to highlight respectively changes in curriculum, learning theory, and assessment. In some cases, principles in the new paradigm are direct antitheses of principles in the old paradigm. The interlocking circles again are intended to show the coherence and interrelatedness of these ideas taken together.

The new paradigm is characterized as emergent because it is not fully developed theoretically and surely not adopted in practice. While there are some shared understandings among cognitivists and constructivists about how learning principles should lead to reform of curriculum and instruction, there are also competing versions of these theories and ideas. In choosing among the different versions, I summarize key ideas that are widely shared and that, for the most part, are compatible with my own view. In the case of constructivist learning theory, for example, I focus on sociocultural theory and a Vygotskian version of constructivism rather than either Piagetian or radical constructivism (von Glasersfeld, 1995). In the case of standards-based curriculum reform, however, I consider the importance of the standards movement in refuting the principles of tracked curricula despite my personal misgivings about the likely harm of standards-based assessments when imposed as part of an external accountability system.

### Cognitive and Social-Constructivist Learning Theories

I began the description of the old paradigm with the tenets of the social efficiency curriculum because zeal for scientific efficiency had led both to the

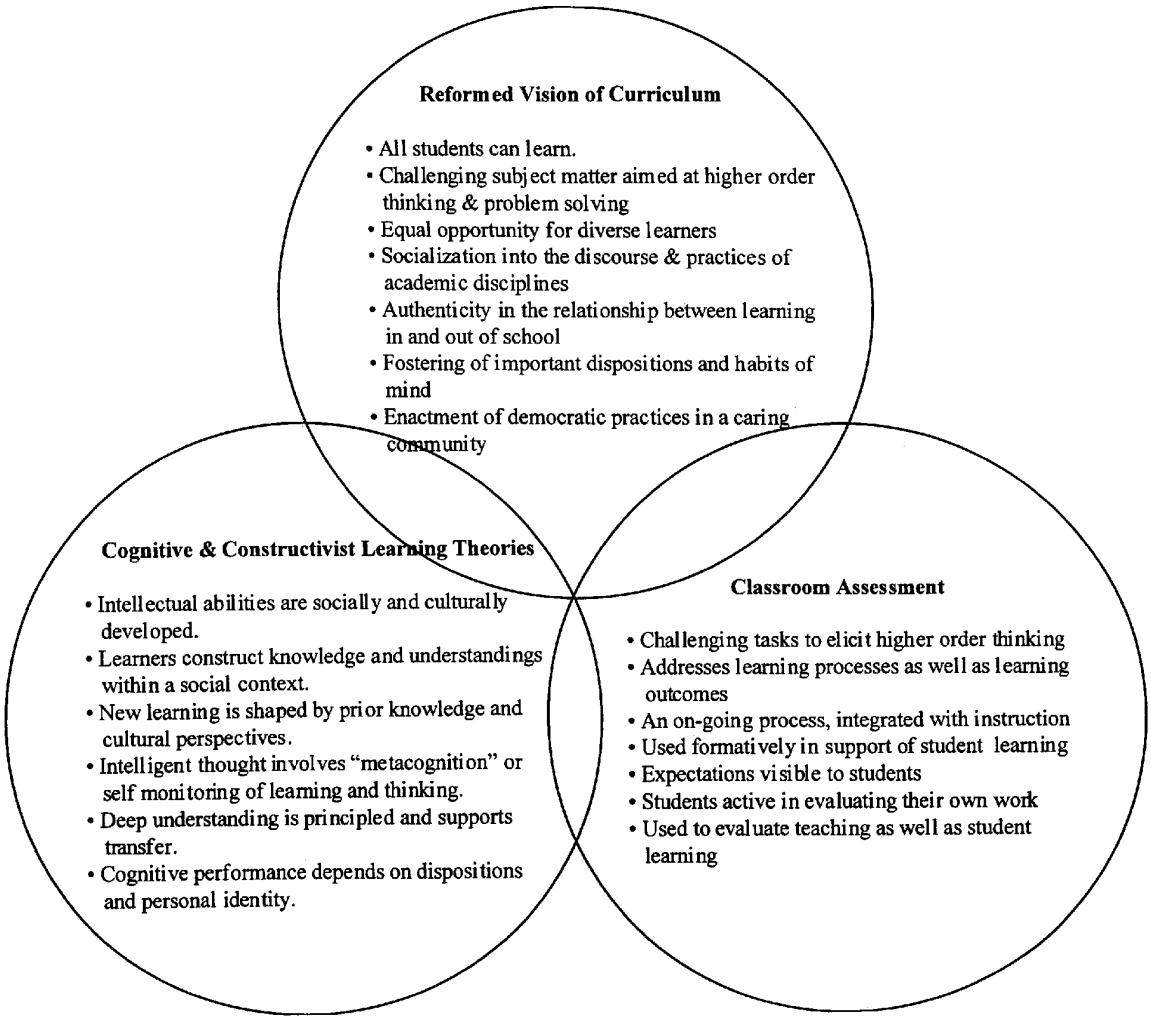



*Figure 4.* Shared principles of curriculum theories, psychological theories and assessment theory characterizing an emergent, constructivist paradigm.

popularity of an atomistic psychology and to enthusiasm for objective measurement formats. Here I treat changes in learning theory as primary and then consider their implications for changes in curriculum and assessment. My summary of new learning theories borrows from similar analyses by Greeno et al. (1996) and by Eisenhart, Finkel, and Marion (1996). However, unlike Greeno et al. (1996), who separate contemporary views into the cognitive and situative perspectives, I list a combined set of propositions that might come to be a shared set of assumptions about learning. Although some of these ideas clearly come from the cognitive tradition, emphasizing mainly what goes on in the mind, and others focus on social

interactions and cultural meanings in the tradition of anthropology, the most important feature of this new paradigm is that it brings together these two perspectives to account for cognitive development in terms of social experience.

The constructivist paradigm takes its name from the fundamental notion that all human knowledge is constructed. As noted by D. C. Phillips (1995), this statement applies both to the construction of public knowledge and modes of inquiry in the disciplines and to the development of cognitive structures in the minds of individual learners. This means that scientists build their theories and understandings rather than merely discovering laws of nature. Similarly individuals make their own interpretations, ways of organizing information, and approaches to problems rather than merely taking in preexisting knowledge structures. For purposes of this framework, I am primarily concerned with constructivist learning theory rather than epistemology. However, an important aspect of individual learning is developing experience with and being inducted into the ways of thinking and working in a discipline or community of practice. Both the building of science and individual learning are social processes. Although the individual must do some private work to internalize what is supported and practiced in the social plane, learning cannot be understood apart from its social context and content.

**Intellectual abilities are socially and culturally developed.** Hereditarian theories of intelligence have been replaced by interactionist theories. It is now understood that cognitive abilities are "developed" through socially mediated learning opportunities (Feuerstein, 1969) as parents or other significant adults interpret and guide children in their interactions with the environment. Interestingly Vygotsky's model of supported learning which has such importance in this volume for the teaching and learning of mathematics, social studies, and so forth, was conceived initially to describe the development of intellectual competence more generally—that is, how one learns to think. Indeed, efforts to study mental processes and how they are developed have blurred the distinctions between learning to think, learning how to solve problems within specific domains and contexts, and developing intelligence. Earlier work in this vein demonstrated the modifiability or instructability of intelligence by working with extreme populations such as educably mentally retarded children (Budoff, 1974) and low-functioning adolescents (Feuerstein, 1980). More complex, present-day intervention programs by Denny Wolf, Ann Brown, and others can be seen as extensions of this same idea, although improving intelligence per se is no longer the aim. Wolf and Reardon (1996), for

example, talk about "developing achievement" by devising a staged curriculum that allows students supported practice with all of the enabling competencies (writing an essay, piecing together historical evidence, or conducting an experiment) that ensure proficient performance of the final, challenging goals. Wolf and Reardon also note that teachers struggling to create such a curriculum must confront "the fundamental difference between raw aptitude and hard-earned achievement" (p. 11).

**Learners construct knowledge and understandings within a social context.** To learn something new the learner must actively teach herself what new information means. How does it fit with what I already know? Does it make sense? If it contradicts what I thought before, how am I going to reconcile the differences? If I substitute this new idea for an old one, do I have to rethink other closely related ideas?

Although earlier, Piagetian versions of constructivism focused on individual developmental stages or processes (Eisenhart et al., 1996), over time, cognitive psychologists have come increasingly to take seriously the influence of social processes. The rediscovery of Vygotsky provided a theoretical model for understanding how social interactions between adult and child could supply both a model of expertise and the opportunity for guided practice so that the child could eventually internalize desired skills and perform them independently. According to Vygotsky, the zone of proximal development (what an individual can learn) "is the distance between the actual developmental level as determined by independent problem solving and the level of potential development as determined through problem solving under adult guidance or in collaboration with more capable peers" (1978, p. 86). Bruner's notion of "scaffolding" elaborated on the kinds of social support that would assist a child in performing a task that would otherwise be out of reach; these supports include engaging interest, simplifying the problem, maintaining direction, marking critical features, reducing frustration, and demonstrating (Wood, Bruner, & Ross, 1976).

Other contemporary perspectives (also borrowing from Vygotsky) go further, suggesting that historical and cultural factors don't merely *influence* learning but *constitute* or form identity, images of possible selves, and the repertoire of knowledge and skills needed to participate in a community of practice (Eisenhart et al., 1996; Lave & Wenger, 1991). A result of this line of thinking, emphasizing socially negotiated meaning, has been to complicate the models of effective learning. In contrast to the decontextualization and decomposition fostered by associationism,

now no aspect of learning can be understood separate from the whole or separate from its social and cultural context. For example, in describing the "thinking curriculum" Resnick and Klopfer (1989) emphasized that thinking skills could not be developed independent of content, nor could cognitive skills be separated from motivation. Apprenticeship models are a natural extension of this reasoning because they provide for role development and support for novice performances as well as the contextualization of skill and knowledge development in a particular community of practice.

**New learning is shaped by prior knowledge and cultural perspectives.** For those eager to throw off the shackles of the old paradigm, the role of content knowledge has posed an interesting dilemma. Because mastery of subject matter knowledge has traditionally implied at least some rote memorization, curriculum reformers have sometimes swung to the other extreme, emphasizing processes over content. Yet, it is a fundamental finding of cognitive research that knowledge enables new learning (Resnick & Klopfer, 1989). Those with existing knowledge stores can reason more profoundly, elaborate as they study, and thereby learn more effectively in that knowledge domain (Glaser, 1984). Knowledge in a domain includes facts, vocabulary, principles, fundamental relationships, familiar analogies and models, rules of thumb, problem-solving approaches, and schema for knowing when to use what. Effective teaching (and assessment) not only begins by eliciting students' prior knowledge and intuitions, it also develops a community of practice where it is customary for students to review and question what they already believe.

Ironically the validity of efforts to assess prior knowledge are themselves affected by a student's knowledge base and by cultural practices. Often prior knowledge is measured using skills checklists or a pretest version of the intended end-of-unit test. Such procedures are likely to underestimate the relevant knowledge of all but the most sophisticated members of the class since most will not be able to make the translation between pretest vocabulary and their own intuitive knowledge gained in other contexts. Open discussions or conversations are more likely to elicit a more coherent version of students' initial conceptual understandings as well as the reasoning behind their explanations (Minstrell, 1989; Yackel, Cobb, & Wood, 1991). It is also essential that teachers become familiar with relevant experiences and discourse patterns in diverse communities so that children entering schools will be able to demonstrate their competence rather than appearing deficient because of unfamiliarity with the teacher's mode of questioning (Heath, 1983).

**Intelligent thought involves "metacognition" or self-monitoring of learning and thinking.** Adept learners are able to take charge of their own learning using a variety of self-monitoring processes (Brown, 1994). This concept of metacognition, or thinking about thinking, is a key contribution of the cognitive revolution. Being able to solve problems within each domain of practice involves what Sternberg (1992) called "executive processes" such as (a) recognizing the existence of a problem, (b) deciding on the nature of the problem, (c) selecting a set of lower-order processes to solve the problem, (d) developing a strategy to combine these components, (e) selecting a mental representation of the problem, (f) allocating one's mental resources, (g) monitoring one's problem solving as it is happening, and (h) evaluating problem solving after it is done.

Metacognitive abilities can be learned through socially mediated processes in the same way that first-order cognitive abilities are learned. For example, Brown, Bransford, Ferrara, and Campione (1983) conducted studies in which children's comprehension of texts could be improved by teaching them specific strategies such as questioning, clarifying, and summarizing—the kinds of strategies that proficient readers use without explicit training. These ideas were then extended to a full-blown reading comprehension intervention, Reciprocal Teaching (Palincsar & Brown, 1984), which blends the ideas of strategies training and cognitive apprenticeship. Although not a scripted lesson, the routines of reciprocal teaching dialogues give students socially supported practice with four metacognitive strategies—predicting, question generating, summarizing, and clarifying—for the purpose of developing a shared understanding of the text. Reciprocal Teaching has been used primarily with learning disabled children and students in remedial reading programs and has shown a median gain of .88 standard deviation for students receiving the intervention compared to controls (Rosenshine & Meister, 1994).

**Deep understanding is principled and supports transfer.** There is a close relationship between truly *understanding* a concept and being able to *transfer* knowledge and use it in new situations. In contrast to memorization—and in contrast to the earlier behaviorist example where students mastered Arabic to Roman numeral translations but couldn't do them in reverse—true understanding is flexible, connected, and generalizable. Not surprisingly, research studies demonstrate that learning is more likely to transfer if students have the opportunity to practice with a variety of applications while learning (Bransford, 1979), and if they are encouraged to attend to general themes or features of problems that imply use of

a particular solution strategy (Brown & Kane, 1988). Learning the rules of transfer is, of course, an example of a metacognitive skill that can be supported instructionally.

In working with pre-service teachers, I have suggested that a goal of teaching should be to help students develop "robust" understandings (Shepard, 1997). The term was prompted by Marilyn Burns's (1993) reference to children's understandings as being "fragile"—they appear to know a concept in one context but not to know it when asked in another way or in another setting. Sometimes this fragility occurs because students are still in the process of learning. All too often, however, mastery appears pat and certain but does not transfer because the student has mastered classroom routines and not the underlying concepts. To support generalization and ensure transfer, that is, to support robust understandings, "good teaching constantly asks about old understandings in new ways, calls for new applications, and draws new connections" (Shepard, 1997, p. 27).

From the situative perspective or the perspective of activity theory (Greeno, Smith, & Moore, 1993; Lave & Wenger, 1991; Rogoff, 1990) the example of children not being able to use their knowledge in new settings might be attributed to their being removed from the original community of practice, which provided both meaning and support for knowledge use. Perhaps. However, a more probable explanation, given the pervasiveness of rote teaching practices, is that children do not really understand even in the initial setting. Although there appears to be disagreement between cognitivists and situativists regarding knowledge generalization (Anderson, Reder, & Simon, 1996), in fact, both groups of researchers acknowledge the importance of transfer. Cognitivists focus more on cognitive structures, abstract representations, and generalized principles that enable knowledge use in new situations, whereas "in situativity, generality depends on learning to participate in interactions in ways that succeed over a broad range of situations" (Greeno, 1996, p. 3). Given Vygotsky's explanation that learning occurs on two planes, first on the social plane between people and then within the individual child, it is likely that a successful program of research will need to consider both.

**Cognitive performance depends on dispositions and personal identity.** Historically, research on motivation was undertaken by social psychologists separate from the work of learning researchers (Resnick & Klopfer, 1989). Only when cognitive researchers began to study metacognition did they come to realize

that students might not employ the strategies they know unless they are motivated to do so.

Traditional classroom practices, especially testing practices, and larger societal norms have created environments in which students may not be motivated to take risks, to try hard, or to demonstrate their intellectual competence. For example, in controlled psychological studies, students are less likely to persist in working on difficult tasks if they know their performance will be evaluated (Hughes, Sullivan, & Mosley, 1985; Maehr & Stallings, 1972). According to motivational researchers, students who believe that academic achievement is determined by fixed ability are more likely to work toward "performance goals," that is, for grades, to please the teacher, and to appear competent. Lave and Wenger (1991) harshly see this "commoditization of learning" to be a pervasive feature of school settings, where the exchange value of learning outcomes is emphasized over the use value of learning. According to this stark portrayal, performance-oriented students tend to pick easy tasks and are less likely to persist once they encounter difficulty (Stipek, 1996). Unfortunately, girls are overrepresented in this category (Dweck, 1986). Students who attribute academic success to their own efforts are more likely to adopt "learning goals," which means they are motivated by an increasing sense of mastery and by the desire to become competent. Not surprisingly, students with a learning orientation are more engaged in school work, use more self-regulation and metacognitive strategies, and develop deeper understanding of subject matter (Wittrock, 1986).

Social psychological research on achievement motivation has produced a list of evaluation practices that are more likely to foster learning goals and intrinsic motivation. For example, motivation is enhanced if errors and mistakes are treated as a normal part of learning, and if substantive, mastery-based feedback is used rather than normative evaluation (Stipek, 1996). Although most of these laboratory-based recommendations make sense and would contribute to a classroom environment where learning and the development of competence are valued, a few points are worrisome. For example, research on intrinsic motivation urges teachers to "de-emphasize external evaluation, especially for challenging tasks" (Stipek, 1996, p. 102) despite the finding elsewhere (Dweck, 1986) that students with a learning orientation "see their teacher as a resource or guide in the learning process, rather than as an evaluator" (Stipek, 1993, p. 15). Moreover, this variable manipulation

approach still leaves the teacher responsible for acting on the student in a way that will induce learning.

In my view, these findings about the negative effects of evaluation on motivation to learn are the product of the beliefs and practices of the old paradigm, which, following Skinner's formulation, provided extrinsic rewards for success on easy tasks. In such an environment, it is not surprising that so many students have developed a performance rather than a learning orientation. It does not follow, however, that evaluation would always stifle motivation if the culture of the classroom were fundamentally altered, and it is dangerous to conclude on the basis of past studies that somehow students need to be protected from critical feedback. Evaluative feedback is essential to learning and presumably can be of the greatest benefit when students are tackling problems that are beyond their level of independent mastery. Thus, the idea of withholding evaluation for challenging tasks is contrary to the idea of supporting students' efforts in the zone of proximal development.

Activity theory and Lave and Wenger 's (1991) concept of legitimate peripheral participation provide a wholly different view of what might "motivate" students to devote their hearts and minds to learning. According to this theory, learning and development of an identity of mastery occur together as a newcomer becomes increasingly adept at participating in a community of practice. If one's identity is tied to group membership, then it is natural to work to become a more competent and full-fledged member of the group. These ideas come from studying learning in the world outside of school. How is it that children learn their native language from parents and community members without the benefit of formal lessons or memorized rules? How do novices learn how to be tailors or Xerox repair technicians or members of Alcoholics Anonymous (Lave & Wenger, 1991)? They are not told explicitly how to participate; instead they are provided with opportunities in the context of practice to see, imitate, and try out increasingly complex skills under the guidance of experts and at the same time practice the role of community member. This is again Vygotsky's notion of socially supported learning applied at once to the development of knowledge, cognitive abilities, and identity. Significantly the beginner is also contributing to and affecting the nature of practice shared with old-timers, which also adds to the worth and meaning of effort. Cognitive apprenticeship programs (Collins, Brown, & Newman, 1989) and Brown and Campione's (1994) community of learners are examples of projects in schools aimed

at developing communities of practice where students' identities as capable learners are constructed as they participate in active inquiry and discussion of challenging problems.

**Reformed Vision of Curriculum**

The elements of a reformed vision of curriculum, summarized in Figure 4, set the direction for the kinds of changes contemporary educational reformers are trying to make in classrooms. Some of these principles are part of the wider public discourse, familiar to policymakers and journalists as well as educators and researchers; others are articulated by a smaller circle of education reformers.

At the political level, present day educational reform is motivated by the poor performance of U.S. students in international comparisons and anxiety about economic competitiveness. In this light, many politicians have accepted the argument from researchers that current problems are in part due to past reforms aimed at minimum competencies and low-level tests. As a result, standards and assessments have been given a central role in reforming public education. The mantra of standards-based reform, "high standards for *all* students," promises the pursuit of both excellence and equity, goals that were held at odds by prior belief systems. The first set of reform principles—all students can learn, challenging standards aimed at higher order thinking and problem solving, and equal opportunity for diverse learners—are widely shared and recur in legislation, various state and national policy reports, and in standards documents for each of the disciplines.

The remaining elements of the agenda are not so familiar in public arenas but are essential to accomplishing the first set. If it has never been true before, how is it that all students can be expected to master challenging subject matter and perform to high standards unless students can be engaged in learning in fundamentally different ways? Socialization into the discourse and practices of academic disciplines, authenticity in the relationship between learning in and out of school, fostering of important dispositions and habits of mind, and enactment of democratic values and practices in a community of learners are elements of the reform agenda that follow from the research on cognition and motivation described previously, as well as the basic empirical work that has documented the dreariness and meaninglessness of traditional practice. Taken together they portray the curriculum

and classroom environment that would be needed to support student learning at a much higher level.

While not diminishing the significance of these research-based "discoveries," it is important to acknowledge that many of the tenets of reform are not new but bear a remarkable resemblance to ideas advanced by John Dewey 100 years ago. Dewey envisioned a school curriculum that would develop intelligence by engaging students' experience, skills, and interests as the necessary first step in teaching more traditional subject matter. He recognized the social nature of learning and the desirability of creating a miniature community so as to initiate the child into effective social membership (Kliebard, 1995). In light of the ambitious claims of the current reforms, it is sobering to recognize how many attempts have been made since then to implement the ideals of progressive education. Successes have been short-lived because, as suggested by Cremin (1961), the complexity of such reforms required "infinitely skilled teachers."

A further caveat is also warranted. The framework in Figure 4 is intended to address how learning theory and curriculum reform come together at the level of the classroom to reshape instruction and assessment. It would be a mistake, however, to imagine that significant changes could occur in classrooms without corresponding changes in the community and at other levels of the educational and political system. McLaughlin and Talbert (1993), for example, identified the multiple, embedded contexts of teachers and classrooms that may constrain or facilitate educational change. These include subject matter cultures, state and local mandates, the parent community and social class culture, the expectations of teachers in the next higher level of schooling, and professional contexts including teachers' associations and university teacher education programs. In later sections of the chapter I address two connections to contexts beyond the classroom: the relationship between classroom and system-level assessments and the professional development needs of teachers. Still the chapter is limited by not being able to treat the concomitant changes that would be needed in these several other contexts to support change in the classroom.

**All students can learn.** The slogan that "all students can learn" is intended to promulgate what the Malcolm Report (Malcolm, 1993) called "a new way of thinking." It is a direct refutation of the long-standing belief that innate ability determines life chances. Although such affirmations by themselves will not be sufficient to provide the necessary learning opportunities, the slogan is important

because it serves to disrupt the self-fulfilling practices of the old paradigm whereby only certain students were smart enough to master difficult content and therefore only an elite group of students was given access to challenging subject matter.

**Challenging standards aimed at higher order thinking and problem solving.** That the common curriculum should address challenging standards aimed at higher order thinking and problem solving is likewise a rejection of past practices and theory. The transmission model of learning based on rote memorization of isolated facts removed learning from contexts that could provide both meaning and application. By watering down curricula and emphasizing minimum competencies, schools have lowered expectations and limited opportunities to learn. By contrast, if children are presented with more challenging and complex problems and given the support to solve them, they will develop deeper understandings and at the same time become more adept at the modes of inquiry and ways of reasoning that will help them solve new problems in the future.

**Equal opportunity for diverse learners.** The commitment to equity as part of standards-based reform implies changing both expectations and resources. Inequality of opportunity pervades the U.S. educational system. Not only do children from poor and minority communities receive less rigorous curricula (a problem that standards are intended to address), but they are taught by teachers with less academic preparation and less experience, have access to fewer books and computers, and often attend schools that are unsafe or where it is uncool to take school work seriously (Fordham & Ogbu, 1986; Kozol, 1991; Oakes, 1985, 1990). It is a belief of standards advocates, such as the National Council on Education Standards and Testing (1992), that public accountability based on standards and assessment will help assure the availability of adequate resources.

Equal access to high-quality instruction implies more than even-handed allocation of fiscal and human resources, however. It also requires a more thoughtful and deeper understanding of the tension between treating everyone the same versus respecting and responding to differences. If prior knowledge enables new learning, then it is essential that children from diverse backgrounds have the opportunity to demonstrate what they "know" about a topic and also that they be able to a participate in the classroom in ways that are consistent with the language and interaction patterns of home and community (Au & Jordan, 1981; Heath, 1983; Tharp & Gallimore, 1988). Brown (1994) talks about providing multiple "ways in" to school learning but also insists on "conformity on the basics," everyone must read, write,

think, reason, etc. Dewey was often misunderstood as being child-centered at the expense of subject matter. His rejection of this false dualism is equally applicable here. One begins with the experience of the child, but the purpose of the course of study was to bring him into the logically organized and disciplined experience of the mature adult (Dewey, 1902).

**Socialization of students into the discourse and practices of academic disciplines.** If psychological studies have demonstrated that both intelligence and expert reasoning in specific knowledge domains are developed through socially mediated cognitive activity, then there still remains the practical question of how to ensure that these kinds of interactions take place in classrooms. And if, as Brown (1994) has suggested, higher thought processes are in part an "internalized dialogue" (p. 10), then how teachers and students *talk* to each other is of paramount concern. Borrowing from learning in informal settings, sociocultural theorists note that development of competencies normally occurs by experts and novices having the opportunity to converse as they work together on a common goal or product (Rogoff, 1991; Tharp & Gallimore, 1988). The point of "instructional conversations" in school is not just to provide information but to develop shared meanings between teacher and students, to connect schooled concepts to everyday concepts, and to allow students to gain experience with the ways of reasoning and thinking in a field (Tharp, 1997). For example, if in classroom exchanges students are routinely asked to explain their thinking or to clarify terms, then eventually these habits are internalized and become a part of their thinking process as well as a social norm in the classroom (Hogan & Pressley, 1997). In mathematics, Schoenfeld discussed the kinds of classroom practices that would foster a "culture of sense making," where "figuring it out" was how students learned to approach mathematical content. The popular writers' workshop (Atwell, 1987; Graves 1983) satisfies the elements of activity theory in its efforts to support the development of young writers. It provides a model of mature practice, engages students in elements of the process (brainstorming ideas, drafting, exchanging critiques, revising, editing) in the context of the whole, and provides the opportunity to try on the role of author.

**Authenticity in the relationship between learning in and out of school.** Whereas the previous principle borrows from models of informal learning in families and communities to change *how* students learn, this principle suggests that the *what* of subject matter should also change to provide better connections with the real context of knowledge use. School learning has traditionally between quite

distinct from learning outside of school. In-school learning is formal and abstract and removed from the use of tools or contexts that would supply meaning (Resnick, 1987). That's why, for example, students often lose track of the problem they are trying to solve or give silly answers, such as "3 buses with remainder 3" are needed to take the class to the zoo. However, school learning is also more reflective, disciplined and general and thereby provides more adaptability to new problem situations than context-specific learning acquired on the job or in the streets. The intention of reformers like Resnick (1987) and Newmann (1996) is to make the boundaries between school and the world more porous, by bringing authentic contexts into classrooms and at the same time developing habits of inquiry in school that will make students good thinkers and problem solvers in the world.

Once again these ideas were anticipated by Dewey. Dewey did not eschew subject matter or discipline-based study but suggested that students could be inducted into more and more formal knowledge by gaining experience with the practical problems that the disciplines had been developed to solve. His intention in "psychologizing" bodies of knowledge, forestalling treatment of them as polished, organized systems, was to connect with children's own understandings and interests and at the same time to reveal the human purposes underlying the disciplines. Today, Newmann (1996) similarly uses authenticity as a key principle of curriculum reform. For Newmann, authentic achievement involves tasks that are significant and meaningful like those undertaken by scientists, musicians, business owners, crafts people, and so forth. Authentic pedagogy is "more likely to motivate and sustain students in the hard work that learning requires" (Newmann, 1996, p. 27) because their intellectual work has meaning and purpose.

**Fostering of important dispositions and habits of mind.** Several of these reform principles, all aimed at changing the nature of classroom interactions as well as curriculum content, are closely interconnected. The goal of fostering important dispositions and habits of minds is largely redundant with the foregoing principles, except that it is worth calling attention to the importance of motivational goals per se. For example, classroom discourse practices that help students develop "a habit of inquiry" (Newmann, 1996; Wiggins, 1993) not only improve academic achievement in the present but increase the likelihood that students will be motivated to adapt and use their knowledge and skills in new situations. Not only will they know how to tackle problems, to ask and persist in trying to answer the right questions (Wiggins, 1993), to use prior knowledge, to strive for in-depth understanding, and to

express their ideas and findings through elaborated communication (Newmann, 1996), but these ways of thinking will have become habits from long practice in a social setting. As suggested earlier, under the learning principle linking motivation and cognitive performance, the goal is not just to motivate students to work hard on challenging problems but to ensure that they develop identities as capable learners.

**Enactment of democratic practices in a caring community.** Preparation for democratic citizenship requires more than literacy skills or knowledge about government; it requires providing students from diverse backgrounds the opportunity to learn to live and work together in the world. A number of curriculum theorists and educational reformers have called for a more personalized educational system (Darling-Hammond, 1996; Martin; 1992; Sizer; 1984). Smaller schools, longer-term relationships between teachers and students, and more nurturing roles can clearly be demonstrated to enhance academic learning, but this is not their only purpose. Joint productive activity around meaningful tasks also develops common understandings and habits of cooperation and mutual respect. Dewey's concept of democracy was based on community, and his intent was to create a miniature society in the school. His interest in bringing practical occupations into schools was to not only to connect with students' understandings but to "cultivate the social spirit" and to "supply the child with motives for working in ways positively useful to the community of which he is a member" (Dewey, 1897). Similarly, Jane Roland Martin (1992) argues for a more inclusionary curriculum that emphasizes caring, concern, and connection as well as academic knowledge. By engaging students in "integrative activities of living," such as a school newspaper, a theater production, farming, or building an historical museum, students could connect thought to action and gain experience as contributing members of society.

## Classroom Assessment

The third circle in the emergent, constructivist framework addresses principles of classroom assessment. What kinds of assessment practices are compatible with and necessary in classrooms guided by social-constructivist views of supported learning? How does assessment fit or intrude, when students are engaged in collaborative conversations and tackle extended real-world problems? If we think of Vygotsky's zone of proximal development, how might assessment insights help extend a student's current level of learning?

The several principles identified in Figure 4 fall into two main categories having to do with transformation of both the substance of assessments and how they are used. Because these principles are elaborated in the subsequent sections of the chapter, I present them here only briefly. First, the substance of classroom assessments must be congruent with important learning goals. In contrast to the reductionistic and decontextualized view of subject matter knowledge produced by the scientific measurement paradigm, this means that the content of assessments must match challenging subject matter standards and be connected to contexts of application. As part of this, assessments must mirror important thinking and learning processes, especially modes of inquiry and discourse, as they are valued and practiced in the classroom.

The purpose of assessment in classrooms must also be changed fundamentally so that it is used to help students learn and to improve instruction rather than being used only to rank students or to certify the end products of learning. The nearly exclusively normative use of tests in the U.S. to compare students to one another and to determine life chances is the key culprit in developing classroom cultures dominated by an exchange value of learning, where students perform to please the teacher or to get good grades rather than to pursue a compelling purpose. By contrast, in classrooms where participation in learning is motivated by its use value, students and teachers would have a shared understanding that finding out what makes sense and what doesn't is a joint and worthwhile project, essential to taking the next steps in learning. To serve this end, more specific principles of classroom assessment require that expectations and intermediate steps for improvement be made visible to students and that students be actively involved in evaluating their own work.

It goes without saying that such a view of assessment is an ideal, rarely observed in practice. In fact, efforts to pursue this vision of assessment practice must contend with the powerful belief system associated with scientific measurement and the dominant paradigm. To be sure, all of the changes called for by the reform agenda and constructivist theory require new knowledge and profound changes in teaching practices. However, I would argue that changing assessment practices is the most difficult because of the continued influence of external standardized tests and because most teachers have had little training beyond objective writing and familiarity with traditional item formats to help them know how to assess their students' understandings (Ellwein & Graue, 1996).

## Relationship of Classroom Assessment to External Assessments

Although this chapter focuses on classroom assessment, it is important to consider how teacher-initiated assessments should relate to external assessments required by district, state, or national mandates, if both were reformed in keeping with the constructivist paradigm. Often assessment reform is promoted without distinguishing among several different assessment purposes, yet it is well known that validity depends on how a test is used. A test designed for one purpose may not be valid if used for a different purpose. Should a statewide literacy test administered to third-graders every April for purposes of school accountability and grade-to-grade promotion also be used instructionally? Gipps (1996), speaking from the context of the British educational system, answers, yes, that "assessment for selection, monitoring and accountability can be assessment to support learning." While it is true that something can be learned from every assessment about one's own teaching as well as students' strengths and weaknesses, I argue that the uniform nature of external assessments and their infrequency means that they will rarely ask the right questions at the right time to be a part of the ongoing learning process.

The distinction between external and classroom assessment is closely related to the familiar distinction between formative and summative evaluation. Scriven (1967) distinguished between the formative role of evaluation feedback, when used internally to improve a program or product, versus the summative role of evaluation data used by outsiders to make final decisions about funding or adopting a program. Typically, external assessments serve summative purposes such as large-scale monitoring of achievement trends, school accountability, school funding, and certification of student proficiency levels. Sometimes external assessments are used formatively at the level of programs, for example, when curriculum revisions are made on the basis of assessment results; but large-scale assessments are rarely used to refocus and improve instruction for individual students. In contrast, as I argue in later sections of the chapter, classroom assessment should be primarily formative in nature, aimed more at helping students take the next steps in learning than at judging the end points of achievement. Still, I also argue that summative evaluation is a natural part of the learning process and should be established as part of classroom routines, especially for older students. Just as students learn the difference between first draft and final versions of their writing, they should also gain experience with making final presentations and reviewing a body of work to

reflect on what has been learned. Summative information is also important for reporting to parents and, if done well, classroom assessments can provide more valuable information about student progress than external measures.

External assessments typically dictate uniformity of content and standardization of procedures, even if they have been reformed to include more open-ended tasks and group problem solving. Standardization is needed for large-scale assessments to ensure that numbers mean the same thing in different contexts, not because of some lingering positivist assumption of pure objectivity but as a basic matter of fairness. For example, if a state assessment is going to be used for school accountability, then content and administration procedures must be standardized to ensure comparability of school results. Everyone takes the same grade-level test at the same, specified time of year. Quite aside from other issues of validity, it would be unfair if some schools were tested in March and others in May, or if some groups had unlimited time to complete the test. Teachers should not give help during the assessment or restate the questions unless it is part of the standard administration. In contrast, for teaching and learning purposes, the timing of assessments makes the most sense if they occur on an ongoing basis as particular skills and content are being learned. Similarly, the level of the test should be focused closely on the student's current level of functioning even if this means using material that is well above or well below a child's nominal grade level. For example, it is well known that the National Assessment of Educational Progress Grade 4 Reading Test could not be used to measure progress for below-grade-level readers because improvement from reading second-grade texts to third-grade texts is off the scale of the fourth-grade test. In the classroom context, teachers may well provide help while assessing to take advantage of the learning opportunity, to gain insight into a child's thinking, and to see what kinds of help make it possible to take the next steps (Shepard, Kagan, & Wurtz, 1998).

Although they are offered as an alternative to standardized procedures, current calls for more interpretive forms of assessment (Gipps, 1999; Moss, 1994, 1996) do not obviate the need for external assessments to be consist and generalizable across sites. Moss (1996) contrasted traditional psychometric emphases on nomological or generalizable explanation with the goal of the interpretive tradition, which is to understand meaning in context. The interpretive model she proposes, in the context of a teacher licensure assessment, eschews independent ratings of various portfolio components and instead engages pairs of judges in weighing evidence from

multiple sources to arrive at a defensible interpretation. This more comprehensive and contextualized look at candidates' performance increases the likelihood of valid assessment results because judges have access to more information. Indeed, they have access, *as part of the assessment,* to the kinds of corroborating evidence that would normally be gathered as part of traditional validity investigations. Although judges interpret evidence from various sources rather than using a standard algorithm to combine scores, Moss is nonetheless concerned about whether different pairs of judges will produce consistent results. For example, to ensure adherence to common standards, she finds it necessary to build in review by a "criterion" reader to verify the evidentiary warrant of interpretive summaries and to resolve disagreements between judges.

The degree of consistency required of external assessments will depend on how they are used and whether they rely on precise meanings of scores or less formal assurances of comparability. Linn (1993) identifies five different levels of precision in "linking" large-scale assessments, ranging from statistical "equating" to more judgmentally-based "social moderation." For example, a report that National Assessment Reading scores improved by 5 points requires strict equivalence of the assessments from year to year. Professional judgment models that result in accreditation of institutions or passing of degree candidates can rely on less precise correspondence between judgments and standards, but fairness and validity still require that results not depend on the idiosyncrasies of individual judges.

Shepard, Kagan, and Wurtz (1998) showed how assessment systems could be designed to serve both external and classroom purposes. However, such multipurpose assessments are costly because the technical and content requirements for each purpose must be satisfied. Although advocates of interpretivist forms of assessment would like to see context and local meanings preserved in what is aggregated for state and national purposes, it is more likely that combining external and classroom purposes will impose standardization in classrooms. In Kentucky, for example, all fifth-grade teachers had to use the same mathematics tasks as portfolio entries so that school comparisons could be made. Using instructionally based assessments for accountability purposes also requires standardization of scoring and external checks, called moderation, to make sure that data being aggregated across classrooms are comparable. The BEAR[4] Assessment System (Wilson & Sloane, in

[4] Developed at the Berkeley Evaluation and Assessment Research (BEAR) Center at the University of California at Berkeley.

press) is a rare example of a curriculum embedded, issues-oriented science assessment that was developed to support classroom-level assessment but also satisfies requirements for comparability across classrooms through scoring moderation and a system of link tests.

Because usually a single articulated system cannot be devised, because of cost constraints, or more likely because of local control of curriculum, the model I propose is one of substantive compatibility between two separate assessment systems. Large-scale assessments should be substantively consistent with high-quality classroom assessments though procedurally separate; that is, they should be guided by the same curriculum standards, engage students in the same kinds of inquiry and demonstrations of proficiency, and be evaluated in terms of shared criteria for judging high-quality work. Given the extensive evidence that external, high-stakes assessments drive instruction, it is essential that external tests reflect more ambitious conceptions of subject matter knowledge than found in traditional tests; they should also elicit thinking and problem-solving skills. Therefore, the kinds of content reforms proposed in the next section, involving more extended and open-ended tasks, are relevant to both large-scale formal assessments and day-to-day classroom assessment. Although I believe that using assessments, even good ones, to drive instructional reform makes the mistake of continuing to de-skill and disempower teachers, I also admit that grassroots, professionally-initiated reforms are unlikely to be successful if teachers continue to feel pressured to drill students in preparation for traditional basic skills tests. So ideally, top-down and bottom-up reforms would be made in concert, with plenty of support for professional development in the middle.

As indicated above, the recommendation that classroom assessments should operate independently from large-scale, external assessments is based on their needs for quite different types of information: immediate and contextualized data, on the one hand, as opposed to rigorously comparable results, on the other. In addition, the two types of assessments differ sharply in the stringency of technical standards they must meet. External tests must demonstrate higher reliability because they are limited, one-time assessments and are often used to make critically important decisions. In contrast, day-to-day evaluations that are made in the context of classroom lessons do not have such high-stakes consequences for students. If a teacher makes an invalid inference on a given day about a student's understanding, that error can be corrected by new information in subsequent days. The purpose of

classroom assessment is not primarily to certify student proficiency levels at a fixed point with precision, but rather to generate hypotheses and guide intervention. Although single-teacher assessments may be significantly less reliable than formal, external tests, it is nonetheless *possible* for teachers operating in a systematic way over time to develop highly accurate assessments of student learning.[5]

## Improving the Content and Form of Assessments

Assessment reform is part of a larger effort to raise standards and improve the quality of education. Standards-based reform envisions a more challenging curriculum for all students focused on higher order thinking skills and depth of understanding. It involves a thoroughgoing reconceptualization of what it means to know in each of the disciplines, as well as fundamental changes in teaching and learning consistent with constructivist theory. Transforming assessment is seen as an essential part of curriculum reform because of widespread beliefs and evidence documenting the distorting effects of high-stakes basic skills tests on teaching and learning (Madaus et al., 1992; Resnick & Resnick, 1992; Romberg, Zarinnia, & Williams, 1989). This belief, that the content of assessments had to be changed to effect other changes, was captured in the slogan "WYTIWYG," "What You Test Is What You Get."

In organizing this section and the next, I follow the logic of assessment reform rhetoric. I consider first the transformation of assessment content, then its form, and finally its use as part of the teaching and learning process. In this section, I review the development of content standards and efforts to redefine important learning goals in each of the disciplines. Sample problems and assessment tasks serve to instantiate the meaning of new curricular goals and at the same time help to illustrate how the form of assessments must change to better represent students' thinking and problem-solving abilities. In the subsequent section, I consider how classroom norms, attitudes, and practices might be changed so that assessment can be used to check on prior knowledge, provide feedback, engage students in self-evaluation, and so forth. Although this sequential arrangement is useful for describing each aspect of assessment reform, in practice these changes are all

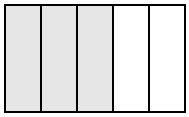
[5] The vision of assessment proposed in this chapter—how classroom assessment should relate to external assessments and how it should be re-formed to reflect social-constructivist learning principles and reformed curriculum standards—is an idealization. No claim is made that teachers will automatically be effective in using assessment in these ways without help in developing extensive subject matter knowledge and expertise in constructivist pedagogy, as well as training in assessment.

entwined. Changing the content and form of assessment are essential in changing its use as part of instruction and, in some cases, helping it become indistinguishable from instruction.

**Reconceptualizing Learning and Achievement in Subject Areas**

Expressly with the intention of changing what it means to know and do mathematics, the National Council of Teachers of Mathematics developed *Curriculum and Evaluation Standards for School Mathematics* (1989). The Standards depart from traditional mathematics instruction, with its focus on computation and rote activities, and instead emphasize sense making in a much broader range of content area topics. In grades K-4, for example, the topics include number sense and numeration, concepts of whole number operations, whole number computation, geometry and spatial sense, measurement, statistics and probability, fractions and decimals, patterns and relationships. In each of the content areas the emphasis is on understanding and on students' ability to investigate and represent relationships. In addition, the standards include what might be termed process goals that emphasize problem-solving, communication, mathematical reasoning, and mathematical connections.

The Standards documents and scores of other mathematics reform projects provide sample problems both to illustrate and to enact the reform. For example, Patrick Thompson (1995) provided the set of questions below to illustrate how non-algorithmic problems can help students "see" a mathematical idea. A traditional fraction question based on the same picture would have asked students only to supply or pick the answer 3/5. In contrast, ongoing experience with a more extended set of questions like these helps students develop their understanding of



a) Can you see 3/5 of something?
b) Can you see 5/3 of something
c) Can you see 5/3 of 3/5?
d) Can you see 2/3 of 3/5?
e) Can you see 1 ÷ 3/5?
f) Can you see 5/4 ÷3/4?

part-whole relationships and multiplicative reasoning applied to fractions. Thus, they can begin to see fractions greater than one and are able to conceptualize a fraction of a fraction as well as a fraction of a whole number.

Additional open-ended tasks, for 4th and 12th graders respectively, are shown in Figure 5. The 4th-grade problem set, for example, asks that children recognize and then generalize a pattern, which is an important precursor to understanding functions. Several features are worth noting. Each mathematics task engages students in thinking and reasoning about important content. Each task could be used interchangeably as an instructional activity, as an assessment, or both. The tasks are complicated and rich enough to involve students in talking about problem solutions so that they can gain experience with explaining and evaluating their own thinking. If students were provided an organized diet of these kinds of activities, as well as more extended projects in the same spirit, there would be no divergence of purpose between the content of assessments and important learning goals.

Similarly in science, several reform documents, especially *Benchmarks for Science Literacy* produced by the American Association for the Advancement of Science's Project 2061 (1993) and the *National Science Education Standards* developed by the National Research Council (1996), have articulated a vision of how curricula should be revitalized to ensure that all students become scientifically literate. The NRC Standards identify fundamental concepts and principles, the "big ideas," in each area of science, as well as inquiry skills needed to conduct investigations and evaluate scientific findings. For example, in Grades K-4 students should know: that plants need air, water, nutrients, and light; that many characteristics of organisms are inherited from their parents but that other characteristics result from interaction with the environment; and that the sun provides light and heat necessary to maintain the temperature of the earth. More importantly, however, the standards emphasize that students should have the opportunity to learn fundamental concepts in depth, to develop subject matter knowledge in the context of inquiry, and to become adept at using scientific knowledge to address societal issues and make personal decisions. Inquiry skills that should be manifest in both instructional activities and assessment tasks include being able to formulate questions, to design and conduct scientific investigations, to use tools for data collection, to formulate and defend a scientific argument, to evaluate alternative explanations based on evidence, and to communicate the results of scientific studies.

**Grade 4 Mathematics Problem Set**
(Mathematical Sciences Education Board, 1993)

All of the bridges in this part are built with yellow rods for spans and red rods for supports, like the one shown here. This is a 2-span bridge like the one you just built. Note that the yellow rods are 5 cm long.


1. Now, build a 3-span bridge.
   a. How many yellow rods did you use? ____
   b. How long is your bridge? ____
   c. How many red rods did you use? ____
   d. How many rods did you use altogether? ____

2. Try to answer these questions without building a 5-span bridge. If you want, build a 5-span bridge to check your answers.
   a. How many yellow rods would you need for a 5-span bridge? ____
   b. How long would your bridge be? ____
   c. How many red rods would you need? ____
   d. How many rods would you need altogether? ____

3. Without building a 12-span bridge, answer the following questions.
   a. How many yellow rods would you need for a 12-span bridge? ____
   b. How long would your bridge be? ____
   c. How many red rods would you need? ____
   d. How many rods would you need altogether? ____

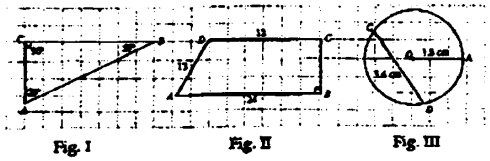
4. How many yellow rods and red rods would you need to build a 28-span bridge? ____ yellow rods and ____ red rods. Explain your answer.

5. Write a rule for figuring out the total number of rods you would need to build a bridge if you knew how many spans the bridge had.

6. How many yellow rods and red rods would you need to build a bridge that is 185 cm long? ____ yellow rods and ____ red rods. Explain your answer.

**Grade 12 Open-ended Mathematics Questions**
(California Assessment Program, 1989)

1. Look at these plane figures, some of which are not drawn to scale. Investigate what might be wrong (if anything) with the given information. Briefly write your findings and justify your ideas on the basis of geometric principles.




2. James knows that half of the students from his school are accepted at the public university nearby. Also, half are accepted at the local private college. James thinks that this adds up to 100 percent, so he will surely be accepted at one or the other institution. Explain why James may be wrong. If possible, use a diagram in your explanation.

**Grade 5 Science Tasks**
(California Learning Assessment System, 1994)

**Fossils**
You are a paleontologist (a scientist who studies past life forms).
You were digging and just discovered a large group of fossils.
Directions:
Open BAG A and spread the fossils on the table.
Use the hand lens to carefully observe each fossil.
Sort your fossils into groups. You may make as many groups as you like.
Write answers to these questions in your journal.
1. Draw your groups. Circle and number each group.
2. How many groups do you have?
3. List the number of each group and tell why you sorted your fossils into these groups.

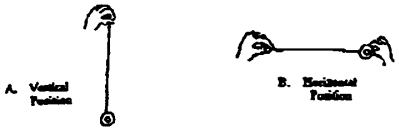
==========

BAG B has a fossil that was found in the area near where you were digging.
Directions:
Open BAG B.
Take out the new fossil and compare it with the other fossils on the table.
4. Does this new fossil fit into one of your groups? If YES, how are they alike?
5. If the new fossil does not fit into any of your groups, describe a new group in which this fossil would fit.
6. Choose one of the fossils and draw a picture of it.
7. In what kind of habitat (environment) do you think this fossil might have once lived? Why?

**Grade 8 Illinois Hands-on Tests for Science**
Pendulum Performance Assessment Task

1. Hypothesis: Write a hypothesis in your journal on how different lengths of string affect the number of swings of the pendulum in 15 seconds.

   Set up materials according to the diagram below:




2. Work with a partner on the pendulum task. Procedure:
   a. Hold the string in one hand with the washer hanging at the other end of the string in the vertical position shown in **Diagram A** above. The length of the string from the top of the washer to your hand should measure 100 cm.
   b. With your other hand raise the washer up until the string is now parallel to the floor in the horizontal position shown in **Diagram B** above.
   c. Release the washer and let it swing back and forth. Do not move your hand. A swing is counted every time the washer makes one complete trip back and forth.
   d. Count how many swings the washer makes in 15 seconds and record this information in your journal. If the final swing is not completed at the end of 15 seconds, count it as one swing.
3. Use the same procedure (steps a-d) to compare how many swings the string and washer make for various lengths of string. Measure the length of each string between the top of the washer and your hand in cm and record this length in a data table in your journal.
4. Graph the data from your investigation in your journal.
5. Use your data and graph to help you answer this question. If the string was longer than 100 cm, what do you predict would happen to the number of swings? Explain your prediction.

*Figure 5.* Examples of open-ended assessment tasks intended to engage students in thinking and reasoning about important content.

Assessment reform has played a key role in giving flesh to the intended science reform just as it has in mathematics. Several examples are provided in Figure 5 to illustrate the alignment of assessment with important content and inquiry skills, the use of extended tasks to elicit student reasoning, and the expectation that students be able to communicate their ideas. The character of these tasks and problems is in marked contrast to earlier item types that emphasized knowledge of scientific facts and terminology.

In English language arts and literacy, the same needs for curricular and instructional reforms are felt as in every other field. If anything, the hostility toward traditional standardized measures has been even greater in this subject area than in others because of the serious ways that tests have misrepresented children's skill development. For example, in their review of formal early literacy tests Stallman and Pearson (1990) document that most measures are based on outmoded theories of early reading development. In fact, readiness tests and first-grade reading tests look very similar to those designed by Gates and Bond (1936) in the 1930s. Moreover, such tests engage young children in a set of activities that are anathema to high-quality reading instruction. Skills are tested out of context, items require that students recognize answers rather than produce an oral or written response, and the activity is dominated by test-taking behavior such as keeping the right place and bubbling answers rather than reading for meaning. The complaints against writing tests are even more severe. Typically, writing tests measure grammar, spelling, and punctuation skills but not the ability to write. When writing tests include an essay component, they still lack the properties of authentic writing situations. The reason for writing is artificial and often unmotivating, the assigned topic may be unfamiliar, the absence of resources such as books and peers is inauthentic, timing constraints are unlike the usual time allotted to develop a piece of writing, and lack of opportunity to revise and edit is inconsistent with good writing practice.

Perhaps it is because existing measures were so at odds with the substance of good instruction, that literacy experts have gone to the greatest lengths to revise the form of assessment as well as its content. For example, over the last two decades, research in emergent literacy has produced increasingly rich descriptions of the typical progressions (and variations) in children's reading and writing development (Sulzby, 1990). This knowledge base could then be used to establish the idea of benchmarking (Au, 1994) not only to document students' progress but also to increase teachers' knowledge about the next steps forward. The writing samples in

Figure 6, excerpted from the North Carolina Grades 1 and 2 assessment materials, illustrates the normal progression in children's increasing writing proficiency. Although there is considerable variation in how children gain command over different conventions of writing and how these conventions are used in different contexts to convey meaning, it was nonetheless possible for North Carolina to construct the following rough, "control of writing" continuum to be used as a framework in analyzing children's writing samples.

Uses invented spelling to convey meaning
- random letters
- letter names
- phonetic spelling

Spells frequently-used words correctly
Uses lower / upper case letters appropriately
Spaces words
Writes complete thoughts / ideas
Ties one thought to another
Sequences events / ideas
Uses details
Moves from a beginning, develops the idea, and concludes
Uses conventional punctuation
Uses conventional spelling

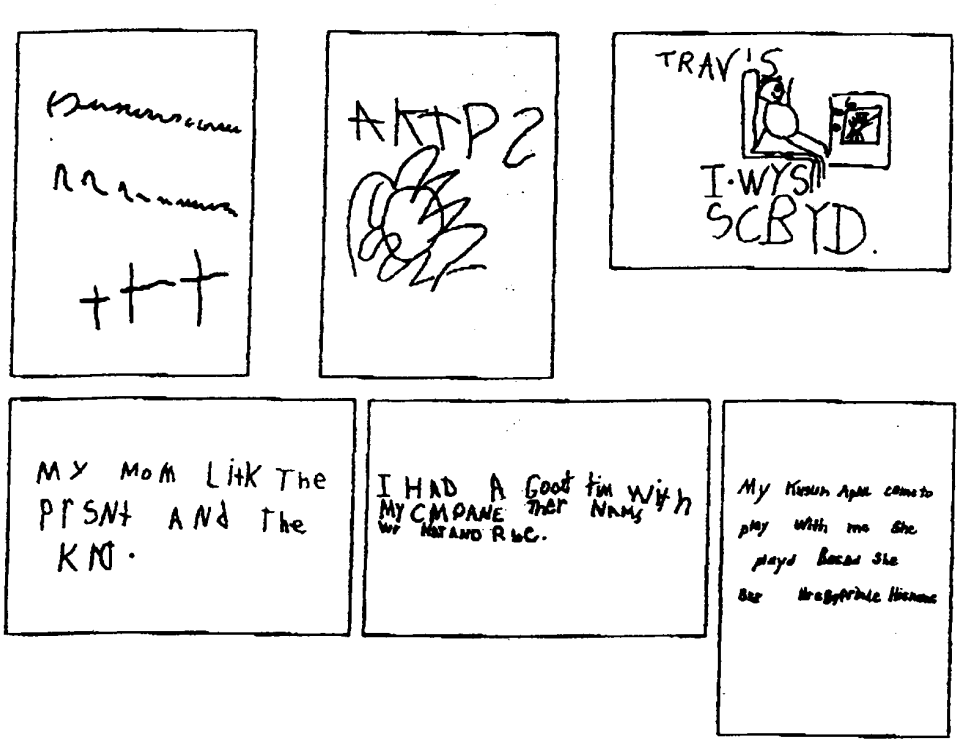



*Figure 6.* Samples of student work illustrating progress on an emergent writing continuum (excerpted from the North Carolina Grades 1 and 2 Assessment).

A shared feature of various literacy assessment practices is that they began in the context of instruction and then made explicit both the process and products indicative of children's increasing proficiency. In the next section, I discuss further how alternative forms of assessments, such as running records and portfolios, do a better job of capturing what is important to measure and at the same time work more smoothly in blending assessment with ongoing instruction.

Rounding out the reforms, curriculum standards have also been developed in history (National Center for History in the Schools, 1996), civics and government (Center for Civic Education, 1994), geography (Geography Education Standards Project, 1994), and social studies (National Council for the Social Studies, 1994). Following the same general outline as the science standards, these documents emphasize development of inquiry skills as well as conceptual understanding of core ideas. For example, in geography, students should know and understand the patterns and networks of economic interdependence on earth's surface. Interestingly, these documents are largely silent about the need for assessment reform. Perhaps because these subject areas were less frequently tested in external accountability programs and because many history and social studies teachers have held on to essay questions and term projects as means of evaluation, assessment was not seen as the driving force for curricular change. Nonetheless, content standards in each area are accompanied by "performance standards" that clearly imply the need for in-depth assessment methods (not short-answer questions) to tap important skills. For example, the "Historical Research" standard for Grades 5-12 is elaborated by six statements about what "the student is able to do," three of which read as follows:

1. Formulate historical questions from encounters with historical documents, eye-witness accounts, letters, diaries, artifacts, photos, historical sites, art, architecture, and other records from the past.

2. Interrogate historical data by uncovering the social, political, and economic context in which it was created; testing the data source for its credibility, authority, authenticity, internal consistency and completeness; and detecting and evaluating bias, distortion, and propaganda by omission, suppression, or invention of facts.

3. Employ quantitative analysis in order to explore such topics as changes in family size and composition, migration patterns, wealth distribution, and changes in the economy. (National Center for History in the Schools, 1996, p. 68)

To develop these abilities, students clearly need supported practice undertaking the very sorts of tasks, that is, instances of the above performances, that will in turn be used to assess mastery at the end of a course of study or in application contexts. Thus, again there is no distinction between instructional activities and authentic assessment tasks.

**Tools and Forms of Assessment**

A broader range of assessment tools is needed to capture important learning goals and to more directly connect assessment to ongoing instruction. As illustrated above, the most obvious reform has been to devise more open-ended performance tasks to ensure that students are able to reason critically, to solve complex problems, and to apply their knowledge in real-world contexts. In addition, if instructional goals include developing students' metacognitive abilities, fostering important dispositions, and socializing students into the discourse and practices of academic disciplines, then it is essential that classroom routines and accompanying assessments reflect these goals as well. Furthermore, if assessment insights are to be used to move learning along rather than merely to keep score on how much learning has occurred so far, then assessment has to occur in the middle of instruction, not just at end points, and must focus on processes of learning—what strategies are children using—not just outcomes. In response to these needs, the armamentarium for data gathering has been expanded to include observations, clinical interviews, reflective journals, oral presentations, work samples, projects, and portfolios. Here I review several of the more prominent alternative forms of assessment. Performance assessments are not considered as a separate category because performance tasks are expected to be a part of ongoing instructional activities and therefore are included in observation-based assessments; they are among the entries in a portfolio assessment system, as well as being used in on-demand, formal tests.

External assessments are necessarily structured and formal to ensure comparability across school settings. Within classrooms, however, it is possible to use both formal and informal assessments, with the balance between the two shifting across the age span. For very young children, assessments should be almost entirely informal. For example, parents and teachers use observations and work samples (children's drawings) to know when scribbling has progressed enough and letter recognition is in place so that demonstration of specific letter shapes would be appropriate. As children grow, it is not only possible for them to participate in more

formal events, designated for assessment purposes, but desirable that they do so, if such events are authentic and consistent with the goal of inducting students into the practices of the discipline. For example, middle-school students might make presentations to report findings from a field project or take a performance-based examination to see if they can use inquiry skills to help conceptualize the class's next project.

Observation-based assessment tools used in early literacy classrooms (Hiebert & Raphael, 1998) illustrate how data about learning can be collected systematically alongside of normal instructional activities. For example, "running records" developed by Marie Clay (1985) are a notation system used during oral reading to keep track of a child's omissions, substitutions, and self-corrected miscues. By close attention to the nature of student errors—called "miscues" (Goodman, 1973) to emphasize that students are responding to cues even when mistaken—teachers can identify students' word recognition skills as well as their ability to make sense of the text. To assess comprehension, teachers might also use story retellings or ask specific questions about the text. Used routinely as a follow-up to reading activity, such assessments provide valuable information but also convey to students the importance of thinking and talking about what they read. Brief assessments during reading time can be used to make immediate instructional decisions, such as focusing on compound words, emphasizing sense making, or changing text level; but informal assessment techniques can also be structured to document children's growth over time, especially if running records and story retells are recorded in relation to graded texts or reading passages of increasing difficulty (Hiebert & Raphael, 1998). Consistent with the idea of socializing students into the discourse and practices of a literacy community, Mervar and Hiebert (1989) documented that children's abilities to choose books can be developed as a goal of instruction and correspondingly are amenable to systematic observation. For example, they noted that students without previous modeling by adults might pick a book without opening it, whereas children in a literature-based classroom were more likely to sample a number of books before choosing by reading segments aloud or looking for specific topics.

Clinical interviews or think-alouds are research techniques that can also be used in classrooms to gain insights about students' learning. One-on-one interactions provide a more extended opportunity to hear and observe students' strategies and to have them explain their reasoning. Individual interviews also make

it possible to conduct "dynamic assessments" that test (and thereby extend) what a student can do with adult support. In these interactions a teacher-clinician is not merely collecting data but is gathering information and acting on it at the same time, thus completely blurring the boundaries between assessment and instruction. Clinical interviews, like good teaching, require that teachers be knowledgeable about underlying developmental continua. In an analysis of a video transcript of Marilyn Burns (1993) conducting an individual assessment (Shepard, 1997), I note that when a student can't answer a question about place value, Burns poses a new problem more meaningful to the child by backing up along an implied developmental progression. She also knows clearly (at a slightly easier point on the imagined continuum) when the child is ready to learn something just out of reach and provides a hint (and indeed the child answers correctly). In other instances, Burns does not attempt to resolve errors that are too far beyond where the child is functioning. Although researchers can provide support for teachers' learning by developing benchmarks, it is unlikely that teachers can develop the kind of detailed knowledge evidenced by Burns except by extensive experience working with children of a specific age and subject-specific curricula. Fortunately, conducting such interviews, possibly with only a few students at any given time, is one way for teachers to develop this knowledge base regarding typical progressions and common errors.

Portfolios are another, highly popularized, new form of assessment. Borrowing from the arts and from professions such as architecture and advertising where individuals collect samples of their best work to demonstrate their talents and skills, the intention of assessment reformers is to use portfolios of student work to provide more authentic documentation of achievement. When considered from the perspective of external, accountability assessments, portfolio-based assessments face a number of serious obstacles including reliability of scoring and fairness questions such as "Whose work is it, really?" However, when used in classrooms solely for teaching and learning purposes, portfolios can provide an organizing structure for teacher-student critiques and student self-reflections, thereby fostering metacognitive goals that might not be attended to if the various assignments in the portfolio were undertaken separately. Within classrooms the relevant comparison is not whether portfolio assessments can be made as reliable and rigorously comparable as standardized measures but whether a portfolio structure can help

teachers and students become more systematic in analyzing and learning from student work than would ordinarily occur as a part of instructional activities.

A number of researchers have written about the unique features of portfolios as a teaching tool. Yancey (1996), for example, argues that reflection is the defining characteristic of writing portfolios. Through construction of portfolios students set goals for learning, review their work and develop criteria for selecting particular pieces over others, learn to evaluate the strengths and weaknesses of their own work, and gain experience in communicating their purposes and judgments to others. Work by Camp (1992) and Hilgers (1986) among others illustrates that students can develop the ability to articulate and apply critical criteria if they are given practice and experience doing so. Klimenkov and LaPick (1996), teachers at Orion Elementary School, combined the use of portfolios with student-led conferences. Their goals included student empowerment as well as helping students understand what steps they needed to take to move ahead. Evaluation of the Orion project found that students indeed took greater responsibility for their own learning; but the drop-off in effort after the midyear conference suggested that the device still relied on external motivation to a large extent. Duschl and Gitomer (1997) sought to create what they call a portfolio culture in classrooms by using portfolio assignments and negotiated criteria to engage in "assessment conversations." Through such conversations teachers find out what students know, students gain experience with processes of scientific explanation, argument, and presentation, and students learn to apply standards of scientific plausibility. Portfolios are the vehicle for conceptualizing and structuring these classroom interactions.

Whether portfolios can be used for both classroom and external purposes is highly controversial. As suggested previously, large-scale assessment purposes bring with them the need for uniform assignments and scoring criteria. Not only will such constraints make it less likely that instructional activities will fit the learning needs of individual students, but the high-stakes, evaluative context may also defeat efforts to engage students in taking responsibility for their own learning. While acknowledging the tension, Au and Valencia (1997) argued that the benefit teachers derived from learning to score portfolios and the improvement in students' writing proficiency, even from a mandated portfolio system, were sufficient to warrant their use. Myers (1996) suggested that there might be a conscious interaction (or articulation) between portfolios constructed for formative and summative purposes rather than assuming that a single portfolio could reasonably

serve both purposes. While I think it will be impractical to implement two full-blown portfolio systems in the same subject area throughout the school year, it would be feasible for classroom portfolios to include evidence from other forms of external assessments, such as on-demand performance tasks or standardized tests, and to address explicitly their relationship to classroom-based evidence.

As a general rule, teachers should use a variety of assessment tools, choosing in each case the mode of data collection that best captures intended knowledge and skills in their context of use. Sometimes this will mean using more traditional-looking quizzes and examinations. As is the case for all assessment modes, there should be an explicit rationale for using conventional assessment techniques both with respect to the format of test questions and for the "on-demand" character of test events. Essay questions, for example, may still be the best means for students to demonstrate their ability to use either historical or scientific evidence to support an argument. Traditional fill-in-the-blank, short-answer, or multiple-choice questions may also be useful in checking for certain kinds of procedural knowledge, so long as these skills are represented in proportion to their substantive importance in the curriculum and not simply because they are the easiest to measure.

Learning goals should also determine whether and in what proportion assessments should be administered "on demand." In my personal experience, this point is often in contention for both teacher education and doctoral students. "If you believe in assessment reform, why give tests or doctoral comprehensive examinations?" The answer should be that classroom participation, extended projects, research papers, and tests each support and reflect different kinds of learning and therefore provide different kinds of evaluation data. Assuming that content has been reformed in the ways described previously, tests demonstrate "walking around knowledge," that is, the conceptual schemes and big ideas that you should have in your head without having to look them up in a book or ask a colleague. For prospective teachers, this means that formal examinations should tap the kinds of knowledge required "on demand" in authentic applications, for example, when asked for your teaching philosophy in a job interview, when making an instructional decision on the fly, when arguing for one choice over another in a district curriculum committee meeting, or when explaining student work and assessment data to a parent. Similarly, for doctoral students, tests can be an authentic measure of professional knowledge if they draw on the expertise one needs to answer questions from school board members, to review manuscripts

submitted to journals, to brainstorm about study designs, to respond off the cuff in a professional debate, and so forth. In a later section on the culture of the classroom, this issue is taken a step further, emphasizing that students should be made aware of the pedagogical rationale for the balance of assessments chosen—how do they as a set represent the learning goals for the class.

**Multiple Modes of Assessment to Ensure Fairness and Transfer**

Variety in assessment techniques is a virtue, not just because different learning goals are amenable to assessment by different devices, but because the mode of assessment interacts in complex ways with the very nature of what is being assessed. For example, the ability to retell a story after reading it might be fundamentally a different learning construct than being able to answer comprehension questions about the story; both might be important instructionally. Therefore, even for the same learning objective, there are compelling reasons to assess in more than one way, both to ensure sound measurement and to support development of flexible and robust understandings.

In the measurement literature, it is well know that assessment formats can have significant effects on performance levels. For example, one of the best known and pervasive effects is the relative advantage of women over men on essay examinations compared to multiple-choice measures of the same content domain (Mazzeo, Schmitt, & Bleistein, 1993). In science, Shavelson, Baxter, and Pine (1992) found that students did not score equivalently on paper-and-pencil, computer simulations, or hands-on versions of the same electric circuit problems. In the Orion portfolio project described previously, teachers worried that students who were shy or suffering from stage fright were at a disadvantage in demonstrating their knowledge in student-led conferences (Klimenkov & LaPick, 1996). When different results occur from different assessment formats, it will depend on the situation whether one result should be treated as more valid than the others. Validity studies of male-female differences on Advanced Placement history exams, for example, suggest that multiple-choice and essay exams are actually measuring different constructs roughly corresponding to historical knowledge and historical argument. Instead of concluding that one format is biased, the evidence suggests that both formats are needed to adequately represent the content domain. By contrast, in the accommodations literature, certain test formats would be deemed biased if irrelevant features of the assessment prevent students from demonstrating their true

level of competence. This occurs, for example, when English language proficiency is confounded with assessment of mathematics or when learning disabled students are unable to demonstrate their knowledge because of excessive writing demands or lengthy examination periods. In the case of nonconstant results for tasks believed to be equivalent, when there is no basis for choosing between bias or multiple-construct interpretations, as in the Shavelson example above, the best strategy is to use multiple data sources for purposes of triangulation without presuming that one assessment mode is more accurate than others.

From the perspective of assessment fairness in classrooms, students should be allowed to demonstrate their competence using the particular conditions that show them to best advantage (at least as *one* of the ways they are assessed). This might mean giving an oral presentation rather than taking a written exam, writing about a topic that is familiar, having access to translated versions of the task, and so forth. From a teaching perspective, however, students should not always rely on the format that is most comfortable. Good instruction focuses on areas of weakness as well as strengths and ensures that students' knowledge becomes increasingly flexible and robust (i.e., transfers) across contexts of application. To do this, teachers must be aware of how variation in assessment or instructional task features affects performance. For many students there will not be reliable patterns of difference across assessment modes, but when consistent patterns occur, they should lead to targeted interventions. For example, English-language learners should have the opportunity to demonstrate their mathematical knowledge without the confounding effects of language proficiency but at the same time should be working to improve mathematical communication. Similarly, in the Orion portfolio example, teachers worked with students who had difficulty presenting at conferences to set goals for public speaking skills and developing public voice.

Using a variety of tasks, for both instruction and assessment, is also important in teaching for understanding and transfer. Teaching-the-test research reminds us that repeated practice with identical instructional and test formats leads to an inflated picture of student achievement (Shepard, 1997), because students can appear to have mastered instructional routines without understanding underlying concepts. Students are more likely to develop understanding and the ability to apply knowledge in new situations if they are presented with a variety of problems and encouraged to draw connections. For example, a 6-year-old may not be troubled if he adds 4 + 6 on paper and gets 11, then counts 4 beans and 6 beans and gets 10

beans altogether (developmentally, the 6-year-old does not see a discrepancy, one is numbers and the other is beans). Obviously the goal of early numeracy instruction is to help children develop the correspondence between numbers and objects. Similarly in third grade, students should be helped to draw the connections between area problems (4 x 7) and number line problems (counting by sevens), and will thereby develop more robust understandings of how multiplication works. The principle of multiple assessment modes does not mean using one set of formats for teaching and another for testing but rather to use a range of activities for both and to make awareness of task features an explicit part of classroom discourse. "How is this problem the same as problems we've done before?" "How is it different?"

**Qualitative Methods of Evaluation and Data Synthesis**

Evaluating open-ended tasks and drawing valid inferences from both formal and informal data sources requires new methods of data analysis and interpretation. Telling where as student "is at" can no longer be calculated as the percent of problems answered correctly. With all of the assessment methods described above, there is a profoundly greater need for teacher judgment and qualitative methods of inquiry.

The most obvious new technique for evaluating open-ended tasks and complex performances has been the development of scoring "rubrics." Rubrics provide a set of ordered categories and accompanying criteria for judging the relative quality of assessment products. However, rubrics and formal scoring schemes are inappropriate for many moment-to-moment uses of instructional assessments and, more generally, in classrooms with young children. Furthermore, there are serious questions about whether assigning a quantitative score and ordering performance on a continuum are compatible with sociocultural and constructivist perspectives. Lave and Wenger's (1991) complaint, for example, that testing contributes to the commoditization of learning is likely to apply to new forms of assessment as well unless they have a very different role in the cultural practices of the classroom. Wile and Tierney (1996) argue against "positivistic" or objectified analytic schemes because such schemes "assume relationships between elements which may not be accurate" (p. 212) and "risk excluding or discounting experiences that do not coincide with curriculum guides or checklist descriptors" (p. 213).

My own view is that good assessment practice should include a combination of both locally-negotiated scoring routines and clinical or interpretivist approaches to

data synthesis. Explicit scoring criteria or qualitative descriptors are essential for giving feedback to students and, as I discuss in the next section, for engaging students in self-assessment. Formal criteria for evaluating student work can become the locus of important negotiations and dialog among teachers and students as they develop a shared understanding of what it means to do excellent work. Although it may be nice for the teacher occasionally to write "good idea" in the margin of a history paper, feedback is much more useful if on every paper the teacher or peer critics address familiar categories such as "quality of ideas," "use of evidence," "historical content," and "clarity of communication" and if, over time, students have ample opportunity to connect the meaning of these criteria to examples in their own work. I agree with Wile and Tierney (1996) that it is more useful to keep these descriptive categories separate rather than subsuming them arbitrarily in one holistic score. Older students, however, whose grades will be extracted by some alchemy from numerous sources of evidence, deserve to know how various elements are being sifted and weighed if not strictly added up, because this aggregation process whether quantitative or qualitative also embodies and communicates what is important to know. Teachers need not share their scoring rules with very young children, but might comment, "Oh, that's great Ramona, I see you're making spaces between your words."

But what about all of the other learning occasions and classroom interactions that do not result in a product amenable to scoring? And given multiple sources of evidence, how should a teacher make sense of the whole (not for purposes of a composite grade but to make instructional decisions)? Like a number of other authors, I see the need for an interpretivist approach to data analysis and synthesis (Gipps, 1999; Graue, 1993; Moss, 1996). In my own case, there is a strong connection between the use of qualitative research methods and my training as a clinician, using observations to form a tentative hypothesis, gathering additional information to confirm or revise, planning an intervention (itself a test of the working hypothesis), and so forth. Indeed, some time ago, Geertz (1973) drew an analogy between clinical inference as used in medicine and the way in which cultural theorists "diagnose" the underlying meaning of social discourse, meaning that they use theory to generate cogent interpretations, or generalizations, that have explanatory power beyond thick descriptions.

In classrooms, making sense of observational and work sample data means looking for patterns, checking for contradictions, and comparing the emerging

description against models of developing competence. Thus teachers need to be adept at methods of data sifting and triangulation and at the same time must have a good command of theory (about subject matter learning) to bring to bear in interpreting evidence. Cambourne and Turbill (1990) have suggested that when teachers attempt to make sense of information collected from a variety of classroom literacy activities they proceed in the same way as classical field anthropologists would, by reading through the information and attempting to categorize it. Cambourne and Turbill go on to suggest that the kinds of strategies suggested by Lincoln and Guba (1986) to ensure the dependability and confirmability of naturalistic data—that is, triangulation, purposive sampling, and audit trails—apply as well to classroom-based interpretations of student performance.[6]

Although many would endorse the use of qualitative methods as more philosophically compatible with constructivist approaches to teaching, not everyone would subscribe to the eclectic use of qualitative and quantitative methods as I propose or even to more systematic qualitative schemes. Wile and Tierney (1996), for example, object to the use of benchmarking and categorical descriptions as merely the reimposition of positivistic requirements for experimental control and objectivity. Theirs is a relatively extreme position more consistent with radical constructivism, which allows learners to invent their own reality, and a particular version of qualitative research known as grounded theory (Glaser & Strauss, 1967), which resists the imposition of prior theory on data shifting and interpretation. From Wile and Tierney's (1996) viewpoint, assessment is either positivistic or highly personal and unique, standardized or divergent, simplistic or complex, deductive or inductive, colonial or empowering. I would argue, however, that the use of benchmarks does not have to mean that categories are rigidly imposed, nor does having an eye on shared curricular goals mean that children's individuality must be stifled. While I agree that creating regularized scoring rules for the purposes of external assessments will necessarily compromise the flexibility and responsiveness of assessments for classroom purposes, the question here has to do with the role of

[6] Admittedly it is a tall order to expect teachers to be able to identify consistent and inconsistent patterns of student performance across different types of assessments and to act as amateur anthropologists. Although many good teachers do this, much more training would be needed for it to become a normal part of teaching practice. Note, however, that professional development activities that help teachers develop a deeper understanding about how competence develops in a discipline, of criteria for judging student work, and about making judgments based on multiple sources of evidence, need not be separate training activities but can be closely tied to efforts to develop teachers' pedagogical content knowledge (of which assessment strategies are a part) and to enhance their subject matter knowledge.

discipline- and developmentally-based expectations when the assessment is entirely under the control of the classroom teacher.

Both social constructivism and Deweyan philosophy suggest that teaching should begin with the child but should move toward the organized and disciplined knowledge of mature practice. Correspondingly, more deductive forms of qualitative research balance what emerges from the data with insights provided by theory. Phillips (1996) for example, takes the position that "a person whose mind is a blank slate cannot do research. A researcher notices things that are of interest or that are pertinent—and interest and pertinence depend on, or are relative to, the prior beliefs or assumptions or expectations that are held by the researcher" (p. 1008). By the same reasoning, how could a teacher as researcher form an opinion about student "growth" without a mental model of effective literacy participation? Indeed, Cambourne and Turbill (1990) found that when teachers tried to make sense of assessment data, "the categories they subsequently devised were inevitably related to their values and beliefs about language and language development" (p. 344). If theory drives data interpretation (whether implicit or explicit), why not make it explicit and amenable to critique? In fact, interrogating the adequacy of one's curricular or instructional theory can be an important aspect of using assessment to improve instruction.

## Using Assessment in the Process of Learning

Improving the content of assessments is important but is not sufficient to ensure that assessment will be used to enhance learning. In this section, I consider the changes in classroom practices that are also needed to make it possible for assessment to be used as part of the learning process. How should the culture of the classroom be changed so that students and teachers look to assessment as a source of insight and help instead of its being the occasion for meting out rewards and punishments? In particular, how is learning helped by assessing prior knowledge and providing feedback as part of instruction? How might assessment-based classroom routines, such as reviewing evaluation criteria and engaging students in self-assessment, be used to develop metacognitive skills and students' responsibility for their own learning? How might these endeavors become so seamlessly a part of classroom discourse that students develop a learning orientation, motivated by the desire to increase their competence instead of performing to get good grades or to

please the teacher? As part of this collaborative bargain, how might teachers explicitly use assessment to revise and adapt instruction?

**Changing the Role of Assessment in the Culture of the Classroom**

In a recent review, Gipps (1999) summarized several of the shifts in assessment at the classroom level that follow from sociocultural and interpretive perspectives. I suggest that these can be seen as changes in the cultural practices of the classroom. First, based on Vygotsky's zone of proximal development, assessment should be interactive and dynamic (Lunt, 1993). By providing assistance as part of assessment, the teacher can gain valuable insights about learning strategies and how understandings might be extended. My own view goes further than information gathering, suggesting that, except when there is a formal requirement to record assisted vs. independent performance, dynamic assessment can be used as the occasion to teach, especially to scaffold next steps. Second, assessments should be conducted in the social setting of the group. Closely tied to the view of learning as enculturation, students are socialized into the discourse of the disciplines and become accustomed to explaining their reasoning and receiving feedback about their developing competence as part of a social group. Third, the traditional relationship between teacher and student should be opened up to recognize the learner's perspective. This does not mean that teachers give up responsibility, since they have expert knowledge, but rather that the process becomes more collaborative. Finally, students are given an understanding of the assessment process and evaluation criteria as a means to develop their capacity as self-monitoring learners.

There will, of course, be resistance to these cultural changes. As Sadler (1998) points out, "the long-term exposure of students to defective patterns of formative assessment and the socialization of students into having to accept a wide variety of practices and teacher dispositions (many of which may appear incoherent or inconsistent), promote accommodating survival habits among students" (p. 77). Consistent with my earlier summary of the motivational literature, in which some students are found to have a learning orientation and others a performance orientation, Perrenoud (1991) notes that there are always certain students in a class who are willing to work harder to learn more and therefore go along with formative assessment. But other children and adolescents are "imprisoned in the identity of a bad pupil and an opponent" (p. 92). Perrenoud's description of students, whose aim is to get through the day and school year without any major disaster, is reminiscent

of Holt's (1965) earlier observation that children hide their lack of understanding and use dysfunctional strategies—like guessing or mumbling so the teacher will answer his own question—because of their fears and need to please grownups. According to Perrenoud, therefore, "every teacher who wants to practice formative assessment must reconstruct the teaching contract so as to counteract the habits acquired by his pupils" (p. 92).

Changing cultural practices will be especially difficult because it requires that teachers change their own habits as well. Tobin and Ulerick (1989) described the changes that occurred when a teacher, whose assessment practices had been built on the metaphor of being a "fair judge," adopted instead the metaphor of "a window into students' minds." The result was greater sharing of responsibility between teacher and students. Especially, students had to decide how to represent what they knew and had to schedule time to meet with the teacher to demonstrate their learning. Efforts to transform assessment routines should not be undertaken, however, as if they were separated from curricular goals; instead, particular assessment processes should be selected to model the habits of inquiry, problem-solving approaches, brainstorming ideas, modes of debate and critique, and other discourse practices associated with each discipline. For example, a study by Cobb, Yackel, Wood, Wheatley, and Merkel (1988) was meant to help teachers develop a problem-solving atmosphere, but several of its strategies would foster collaborative assessment as well. For example, students had to listen to and make sense of explanations given by other children and had to work to evaluate and resolve conflicting solutions when they occurred; at the same time, teachers had to learn to communicate to children that they were genuinely interested in their thinking and that one can learn from errors. Duschl and Gitomer (1997) explicitly have in mind the blending of instructional and assessment goals through the creation of a "portfolio culture" and "assessment conversations." For them, central practices include acknowledging student conceptions through assessment strategies; shared evaluation of knowledge claims through application of scientifically legitimate criteria; emphasis on explanations, models, and experimentation as critical forms of scientific reasoning; and communication as a requisite skill in all science activities.

### Assessing Prior Knowledge

Consistent with the principle that new learning is shaped by prior knowledge and cultural perspectives, classroom practices should include assessment of students' relevant knowledge and experience not only to inform teaching but also to

draw students into the habit of reflecting on their own knowledge resources. The number of studies documenting the effect of prior knowledge on new learning is quite large (e.g., see the special issue of *Educational Psychologist*, Spring 1996, edited by Patricia Alexander), but unfortunately, many of these studies involve contrived examples in nonclassroom settings. Most studies are merely predictive indicating that subjects who start out knowing more end up with greater knowledge. A much smaller number of studies demonstrate how background knowledge might be elicited as a means to adapt, focus, or connect instruction. For example, in Au and Jordan's (1981) study, Hawaiian children were encouraged to tell about experiences in their own lives that related to the stories they were learning about in school. Importantly, Au and Jordan's strategy elicited relevant information and invited children to apply it directly without the need for a separate assessment step. Similarly, Pressley et al. (1992) reviewed studies that used questioning to help students activate prior knowledge and make connections to new content. Again, the purpose of the questioning was not so much for the teacher to gain information about students' knowledge but to engage students in explaining their own understandings as a step in learning.

Although relevant background knowledge is usually a help in learning, researchers, especially in science education, have documented how students' intuitive and often naïve beliefs about scientific phenomena may impede development of scientific understanding. For example, students may hold everyday conceptions of heat and temperature that don't match scientific terminology, they may believe that heavy objects fall faster than light ones, or they may be confused about how atoms act together, having only seen textbook pictures of single atoms. Similarly in learning history, students may have quite fanciful beliefs about historical events or expect the past to be a timeless extension of present-day culture (Wineburg, 1996). Although earlier cognitive studies tried confrontation as a means for overturning students' misconceptions, contemporary approaches are more collaborative, providing students with multiple supports, including investigations and hearing ideas from other students, so as to reformulate their ideas (Smith, diSessa, & Roschelle, 1993/94).

In my own experience working in schools, I note two divergent sets of teaching practices that address students' prior knowledge. First, many teachers rely on a traditional, pretest-posttest design to document student progress. The premeasures are often formal, commercially purchased tests and may bear little resemblance to

instructional materials. Pretest results are used to establish each student's achievement level or location but are typically not used to gain insight into the nature of student's understanding. For example, when a problem is missed, it is not known what partial knowledge or competing conception is at work. Detailed objective-referenced measures may tell that a student "can do 2-digit subtraction" but cannot "subtract across zeros"; but formal survey measures of this type are often filed away as baseline data without using such information for specific interventions. At the same time, a significant number of teachers, especially in reading and language arts, use prior knowledge activation techniques as a part of teaching but without necessarily attending to the assessment information provided. For example, K-W-L is a instructional strategy suggested by Ogle (1986) in which students first brainstorm about what they "Know" about a new topic, then try to make predictions about "What" they want to learn from the text or activity, and finally review what they have "Learned."

It is possible that better prior knowledge assessments could be devised along the same lines as the content reforms of outcome assessments described in the previous section. However, as classroom discourse patterns are changed to help students draw connections and reflect on their own understandings, it is arguable that assessing background knowledge should disappear as a separate pretest step and should instead become a part of scaffolding and ongoing checks for understanding. Nonetheless, as part of our efforts to change the culture of the classroom, I would suggest that prior knowledge activation techniques should be marked and acknowledged as "assessments." What safer time to admit what you don't know than at the start of an instructional activity? What better way to demonstrate to students that assessment (knowing what you know and what you don't know) helps learning? Moreover, to develop students' metacognitive knowledge about what helps in their own learning, there might be explicit discussion of both the facilitating and inhibiting effects of background knowledge. The present research literature does not provide clear guidance on the effectiveness of prior knowledge assessments used both as an engagement and reflective activity for students and as an information source for teachers. But this kind of question will be important in a program of research aimed at changing the role of assessment in instruction.

## The Effect of Feedback on Learning

The idea of feedback comes from electronics where the output of a system is reintroduced as input to moderate the strength of a signal. Correspondingly, it is taken for granted in both behaviorist and constructivist learning theories, that providing information to the learner about performance will lead to self-correction and improvement. Extensive reviews of the effects of feedback on learning are provided by Black and Wiliam (1998) and Kluger and DeNisi (1996). Although on average, feedback does improve learning outcomes, Kluger and DeNisi found that one third of 607 effect sizes were negative. The authors were able to explain some of the variation in study findings using a theoretical hierarchy linked to the motivation literature that distinguished between task-oriented feedback, which tended to enhance learning, and self-oriented evaluation, which was more likely to be ineffective or debilitating.

The self vs. task distinction may well be worth attending to in trying to develop a learning-oriented classroom culture. For the most part, however, meta-analyses of the feedback literature are of limited value in reconceptualizing assessment from a constructivist perspective, because the great majority of existing studies are based on behaviorist assumptions. The outcome measures used in typical feedback studies may be narrowly defined indicators of academic achievement, feedback may consist of simple reporting of right and wrong answers, and the end-of-study test may differ only slightly from the prior measure and from instructional materials. For example, in a meta-analysis of 40 studies on feedback by Bangert-Drowns, Kulik, Kulik, and Morgan (1991), half of the studies were based on programmed instruction, nearly all of the studies involved interventions of only one-week duration, feedback consisted mostly of telling students the right answer to the items they got wrong, and both formative and instructional materials were described as "test-like events" by the authors.

Although different in its content from behavioristic models, giving feedback is also an essential feature of scaffolding. As summarized by Hogan and Pressley (1997):

> A key role of the scaffolder is to summarize the progress that has been made and point out behaviors that led to the successes, expecting that eventually students will learn to monitor their own progress. One type of feedback is pointing out the distinction between the child's performance and the ideal. Another important type of feedback is attributing

> success to effort in order to encourage academically supportive attributions. Explicitly restating the concept that has been learned is another helpful form of feedback. (p. 83)

This portrayal derives mostly from research leading to Wood and Bruner's original conception of scaffolding, from Vygotskian theory, and from naturalistic studies of effective tutoring described next. Relatively few studies have been undertaken in which explicit feedback interventions have been tried in the context of constructivist instructional settings.

In one study by Elawar and Corno (1985), teachers were trained to provide written feedback on mathematics homework based on a cognitive perspective, that is, comments were focused on specific errors and on poor strategy, gave suggestions about how to improve, and emphasized understanding rather than superficial knowledge. Not only did written feedback improve achievement significantly, but it reduced the initial superiority of boys over girls and improved attitudes toward mathematics. A slightly different view of the role of feedback emerges, however, from Lepper, Drake, and O'Donnell-Johnson's (1997) study of selected, highly successful tutors. The most effective tutors appear not to routinely correct student errors directly. Instead they *ignore* errors when they are inconsequential to the solution process and *forestall* errors that the student has made systematically before by offering hints or asking leading questions. Only when the forestalling tactic fails do expert tutors *intervene* with a direct question intended to force the student to self-correct, or they may engage in *debugging* using a series of increasingly direct questions to guide the student through the solution process. According to Lepper et al.'s analysis, the tendency of expert tutors to use indirect forms of feedback when possible was influenced by their desire to maintain student motivation and self-confidence while not ignoring student errors.

These two studies highlight a tension in the literature on constructivist teaching practices about the role of formative assessment and feedback. Some might argue that discourse practices in inquiry-based classroom would allow students to revise their thinking without the need for explicit, corrective feedback, because the evidence gathered in the course of an investigation would naturally challenge their misconceptions. My own view is that, yes, formative assessments should be embedded in ongoing instructional activities. Sometimes this will mean that students will receive feedback from the teacher, classmates, or self-reflections without the interactions being marked explicitly as assessments. At other times, as I suggest in the next sections, it will be important that students consciously

participate in assessment so that they can develop an understanding of the criteria that define good work and take responsibility for monitoring their own learning. As was the case with prior knowledge assessment, the question of using explicit feedback versus indirect means for helping students reexamine their ideas will be an important part of a research agenda on constructivist assessment practices.

**Explicit Criteria and Self-Assessment**

Frederiksen and Collins (1989) used the term *transparency* to express the idea that students must have a clear understanding of the criteria by which their work will be assessed. In fact, the features of excellent performance should be so transparent that students can learn to evaluate their own work in the same way that their teachers would. According to Frederiksen and Collins (1989), "The assessment system (should) provide a basis for developing a metacognitive awareness of what are important characteristics of good problem solving, good writing, good experimentation, good historical analysis, and so on. Moreover, such an assessment can address not only the product one is trying to achieve, but also the process of achieving it, that is, the habits of mind that contribute to successful writing, painting, and problem solving (Wiggins, 1989)" (p. 30). For example, in a more recent study, Frederiksen and White (1997) developed assessment criteria to address the most important attributes that they wanted students to develop and exhibit while conducting investigations in science. These included content-oriented criteria (Understanding the Science, Understanding the Processes of Inquiry, and Making Connections), process-oriented criteria (Being Inventive, Being Systematic, Using the Tools of Science, and Reasoning Carefully), and socially-oriented criteria (Communicating Well and Teamwork). Although access to evaluation criteria satisfies a basic fairness criterion (we should know the rules for how our work will be judged), the more important reasons for helping students develop an understanding of standards in each of the disciplines are to directly improve learning and to develop metacognitive knowledge for monitoring one's own efforts. These cognitive and metacognitive purposes for teaching students explicitly about criteria then speak to a different sense of fairness than merely being even-handed in evaluating students, that is, they provide students with the opportunity to get good at what it is that the standards require.

Wolf and Reardon (1996) have this same sense of fairness and equity in mind when they talk about "making thinking visible," and "making excellence

attainable." The specific classroom strategies they describe from Project PACE (Performance Assessment Collaboratives for Education) blend modeling of important processes—for example, showing students what it means "to have a theory" and "support it with evidence" (p. 12)—and explicit discussion by teacher and students of the evaluative criteria they will use in peer editing. Consistent with learning principles in the conceptual framework, these instructional/assessment strategies are examples of socially mediated learning opportunities that help to develop cognitive abilities.

As Wolf and Reardon (1996) anticipate, there is a tension regarding the prescriptive nature of scoring rubrics. Claxton (1995) cautions that students could learn to apply prespecified criteria and thereby raise their achievement but without improving learning acumen, if they become dependent on others for clarification and correction. He argues that "quality" in a particular domain is "*in principle* incapable of complete explication." "Self-evaluation, the ability to recognize good work as such, and to correct one's performance so that better work is produced, grows in the doing as much as in the reflecting, and is irreducibly intuitive" (Claxton, 1995, p. 341). In other words, the ability to self-evaluate is developed in the same way as, and indeed is indistinguishable from, intelligence and discipline-related cognitive abilities. Although Wolf and Reardon (1996) describe a context in which evaluation criteria were negotiated and made a part of the learning process, Claxton's point is well taken. The mere provision of explicit criteria will not enable learning in all the ways desired if they are imposed autocratically and mechanically applied. For the intended benefits to occur, self-assessment has to be a part of more pervasive cultural shifts in the classroom. Students have to have the opportunity to learn what criteria mean (surely not memorize them as a list), be able to apply them to their own work, and even be able to challenge the rules when they chafe.

In its ideal form, self-assessment serves social and motivational purposes as well as improving cognitive performance. Engaging students in debates about standards and in reflecting on their own work can increase students' responsibility for their own learning and redistribute power, making the relationship between teacher and students more collaborative. As stated previously, the teacher does not give over responsibility but by sharing it gains greater student ownership, less distrust, and more appreciation that standards are not capricious or arbitrary. In case studies of student self-evaluation practices in two Australian and English sites, Klenowski (1995) found that students participating in self-evaluation became more

interested in the criteria and substantive feedback than in their grade per se. Students also reported that they had to be more honest about their own work as well as being fair with other students, and they had to be prepared to defend their opinions in terms of the evidence. Klenowski's (1995) data support Wiggins's (1992) earlier assertion that involving students in analyzing their own work builds ownership of the evaluation process and "makes it possible to hold students to higher standards because the criteria are clear and reasonable" (p. 30).

Although claims about the expected benefits of explicit criteria and self-assessment follow logically from the research literature on motivation and cognitive and metacognitive development, there are only a few studies that directly examine the effects of these practices on student learning. The Frederiksen and White (1997) study described previously provided criteria and also engaged students in a set of activities to foster "reflective assessment." At several stages in the Inquiry Cycle curriculum, students evaluated their own work in terms of the criteria. Each time they not only applied the criteria but also wrote a brief rationale pointing to the features of their work that supported their rating. In addition, students in the reflective assessment classrooms used the criteria to give feedback to classmates when projects were presented orally in class. Compared to control classrooms, where students evaluated the curriculum rather than their own learning, students who participated in reflective assessment produced projects that were much more highly rated by their teachers. Importantly, these positive gains were greatest for low-achieving students. On a follow-up test of conceptual understanding in physics, less directly tied to the inquiry criteria, there was no difference between high-achieving students in reflective assessment and control classrooms but heretofore low-achieving students showed dramatic gains in conceptual understanding as a result of reflective self-assessment.

**Evaluating and Improving Teaching**

In addition to using assessment to monitor and promote individual students' learning, classroom assessment should also be used to examine and improve teaching practices. Although a number of authoritative sources (National Council of Teachers of Mathematics, 1995; National Forum on Assessment, 1995; National Research Council, 1996; Shepard, Kagan, & Wurtz, 1998) have acknowledged the importance of using assessment data as a tool for systematic reflection and teacher learning, there has been much less empirical research or formal theorizing about this

collateral use of student assessment. How is assessment used to learn about one's own pedagogy different from use of assessment data to promote individual student growth? Is one just the aggregation of data from the other? (The whole class is struggling with this concept versus three students need extra help.) Although reform rhetoric makes it seem as if there is a shared understand about what it means to use assessment data to improve instruction, examples offered suggest considerable ambiguity. On the one hand "using assessment to improve instruction" might mean using assessment data to select the best technique from one's repertoire to address the problem at hand; or it could imply much more critical inquiry and a transformative purpose.

The authors of the NCTM (1995) *Assessment Standards* offer a conception of classroom assessment that depends on the close intertwining of student growth and instructional improvement purposes. "Although evidence of progress originates with individual students, as indicated in the 'Purpose: Monitoring Students' Progress' section, teachers also sample and collect such evidence to provide information about the progress of the groups of students they teach" (p. 45). Evidence about what students are understanding leads to instructional decisions about both individuals and groups. The NCTM Assessment Standards go on to elaborate three types of instructional decisions informed by assessment data: moment-by-moment decisions, short-term planning, and long-term planning. During instruction, informal observation and questioning help teachers know when to clarify directions, when to redirect instruction to address misconceptions, when to capitalize on student insights to extend a lesson, and so forth. As part of planning for the next day, to ensure the close integration of instruction and assessment, teachers should not only review goals but should also consider what questions or samplings of student work will be used to check on understanding. This process is recursive such that insights from one day's questioning help in shaping the direction of subsequent lessons. Longer term planning requires that teachers consider not only the broader set of learning goals to be addressed but also how students' learning will be assessed across a variety of modes and contexts and in a way that is responsive to students' cultural experiences.

The *National Science Education Standards* (National Research Council, 1996) go further than the NCTM in laying out a continuum of teaching-oriented assessment uses ranging from instructional decision making to critical analysis of teaching effectiveness. At one end of the continuum, the NRC Science Standards, like the

NCTM Assessment Standards, propose that ongoing assessment of students' understanding be used to adjust lessons and teaching plans. At a midpoint on the continuum, assessment data are also used to plan curricula, especially by helping to evaluate "the developmental appropriateness of the science content, student interest in the content, the effectiveness of activities in producing the desired learning outcomes, the effectiveness of the selected examples, and the understandings and abilities students must have to benefit from the selected activities and examples" (p. 87). Finally, at the other end of the continuum, the NRC Science Standards suggest how assessment might be used in "researching" teaching practices. "Engaging in classroom research means that teachers develop assessment plans that involve collecting data about students' opportunities to learn as well as their achievement" (p. 89). Although for each of these purposes teachers follow procedures of systematic inquiry, the "instructional adjustment" end of the continuum is much less critical and seeks to make the best decisions (efficiently within the flow of instruction) without seeking root causes. In contrast, the "teacher as researcher" or "critical inquiry" end of the continuum requires more formal problem identification, involves more systematic data collection, and seeks better understanding and explanation about *why* certain teaching strategies work better. The critical end of the continuum is more consistent with seminal theories of action research (e.g., Corey, 1953; Lewin, 1948).

The NCTM and NRC visions are idealizations based on beliefs about constructivist pedagogy and reflective practice. Although both are supported by examples of individual teachers who use assessment to improve their teaching, little is known about what kinds of support would be required to help large numbers of teachers develop these strategies or to ensure that teacher education programs prepared teachers to use assessment in these ways. Research is needed to address these basic implementation questions, but there are serious theoretical questions as well. To what extent are models of action research applicable to the systematic use of assessment data to improve teaching? There are a number of different definitions of action research, some of which emphasize formal reporting of results to give teachers voice outside the classroom. Even those that focus within the classroom require more formal procedures than could be applied to all areas of instruction all of the time. To be feasible, then, how do master teachers learn to balance ongoing uses of assessment to revise instruction with action research studies reserved for deeper and more systematic investigation of specific instructional practices? To what

extent and in what ways should teachers make their investigations of teaching visible to students? This question seems to me to be fundamentally important to the issue of transforming the culture of the classroom. If we want the cultural practices in the classroom to support development of students' identities as learners—where students naturally seek feedback and critique their own work as part of learning—then it is reasonable that teachers would model this same commitment to using data systematically as it applies to their own role in the teaching and learning process. Finally, how are idealizations about reflective practices affected when external assessment mandates are used to leverage instructional changes? Although aggregate classroom assessment data may indeed be useful when teachers are attempting to make major changes in their instruction, assessment-driven reform may distort the intended curriculum (Koretz & Barron, 1998) and undermine the role of teacher as researcher.

## Conclusions

In this chapter I considered how classroom assessment practices might be reconceptualized to be more effective in moving forward the teaching and learning process. To develop a "social-constructivist" conceptual framework, I borrowed from cognitive, constructivist, and socio-cultural theories. (To be sure these camps are warring with each other, but I predict that it will be something like this merged, middle-ground theory that will eventually be accepted as common wisdom and carried into practice.) Key ideas are recapitulated here briefly, again emphasizing the close interconnections among new theories of learning, reformed curricula, and new ideas about assessment. Then, in closing, I turn to the implications of this vision of classroom assessment for future research.

### Summary

The cognitive revolution reintroduced the concept of mind. In contrast to past, mechanistic theories of knowledge acquisition, we now understand that learning is an active process of mental construction and sense making. From cognitive theory we have also learned that existing knowledge structures and beliefs work to enable or impede new learning, that intelligent thought involves self-monitoring and awareness about when and how to use skills, and that "expertise" develops in a field of study as a principled and coherent way of thinking and representing problems not just as an accumulation of information. At the same time, rediscovery of

Vygotsky and the work of other Soviet psychologists led to the realization that what is taken into the mind is socially and culturally determined. Fixed, largely hereditarian theories of intelligence have been replaced with a new understanding that cognitive abilities are "developed" through socially supported interactions. Although Vygotsky was initially interested in how children learn to think, over time the ideas of social mediation have been applied equally to the development of intelligence, to development of expertise in academic disciplines, to development of metacognitive skills, and to the formation of identity. Indeed, a singularly important idea in this new paradigm is that development and learning are primarily social processes.

These insights from learning theory then lead to a set of principles for curriculum reform. The slogan that "all students can learn" is intended to refute past beliefs that only an elite group of students could master challenging subject matter. A commitment to equal opportunity for diverse learners means providing genuine opportunities for high-quality instruction and "ways into" academic curricula that are consistent with language and interaction patterns of home and community (Au & Jordan, 1981; Heath, 1983; Tharp & Gallimore, 1988). Classroom routines and the ways that teachers and students talk with each other should help students gain experience with the ways of thinking and speaking in academic disciplines. School learning should be authentic and connected to the world outside of school not only to make learning more interesting and motivating to students but also to develop the ability to use knowledge in real-world settings. In addition to the development of cognitive abilities, classroom expectations and social norms should foster the development of important dispositions, such as students' willingness to persist in trying to solve difficult problems and their identities as capable learners.

To be compatible with and to support this social-constructivist model of teaching and learning, classroom assessment must change in two fundamentally important ways. First, its form and content must be changed to better represent important thinking and problem-solving skills in each of the disciplines. This means assessing learning based on observations, oral questioning, significant tasks, projects, demonstrations, collections of student work, and students' self-evaluations, and it means that teachers must engage in systematic analysis of the available evidence. Second, the way that assessment is used in classrooms and how it is regarded by teachers and students must change. This literally calls for a change in the culture of classrooms so that students no longer try to feign competence or work

to perform well on the test as an end separate from real learning. Instead, students and teachers should collaborate in assessing prior knowledge, probing apparent misconceptions, and resolving areas of confusion because it is agreed that such assessments will help students understand better. Students should engage in self-assessment not only to take responsibility for their own learning but to develop metacognitive skills by learning to apply the standards that define quality work in a field to their own work. Similarly, teachers should demonstrate their own willingness to learn by explicitly using assessment data to evaluate and improve instruction.

## Implications for Research

This social-constructivist view of classroom assessment is an idealization. The new ideas and perspectives underlying it have a basis in theory and empirical studies, but how they will work in practice and on a larger scale is not known. The chapter's framework is offered as a conceptual framework for the ambitious program of research and development that will be needed to make the idealization real. In the following paragraphs, I suggest important questions to be addressed from this perspective in three broad areas of investigation: (1) the reliability and validity of classroom assessments, (2) the effects of social-constructivist uses of assessment on learning and motivation, and (3) the professional development of teachers.

**Reliability and validity of classroom assessments.** I argued previously that classroom assessments do not have to meet the same standard of reliability as external, accountability assessments primarily because no one assessment has as much importance as a one-time accountability test and because there are opportunities for correcting erroneous decisions in the classroom context. Still, it is important to have some level of consistency in classroom assessments both for the accuracy of information and to ensure fairness. None of the aforementioned benefits will accrue if students perceived assessment to be erratic or unfair. Teachers are generally accurate in ranking students in their class though not with the same precision as standardized tests. For example, in a recent study teachers' standards-based ratings of students' mathematics achievement showed a strong correlation with test results (.58) (Shepard, Taylor, & Betebenner, 1998; this degree of agreement is impressive given that the rating scale had only four categories and teachers received no special training). However, it is also known that teacher-based

evaluations are prone to certain biases such as the use of idiosyncratic criteria, halo effects, and the tendency to persist with initial judgments of ability rather than adjusting in response to evidence (Shavelson & Stern, 1981). We know also that specific training improves the consistency of teachers' judgments; that is, evaluations can become both more self-consistent and congruent with shared criteria and exemplar papers. An important practical question then will be to decide when such specialized training makes sense for classroom uses of assessment. To what extent might common training in the development and use of classroom assessment scoring criteria lead to greater teacher understanding of curriculum reform as well as enhanced reliability? In contexts where external assessments are aligned with curriculum reform, what connections should be drawn between criteria for evaluating classroom work and external performance standards? For classroom purposes only, what kind of training do teachers need to be able to develop and evaluate hypotheses about students' understandings, so as to gain insights from assessment and not just produce a reliable score?

Validity has ostensibly received the most attention in the assessment reform literature to date because of the emphasis on representing more meaningful content and processes in assessment tasks. Questions still remain, however, as to whether new forms of assessment are measuring as intended. Are students developing and using advanced thinking and problem-solving abilities? Are they able to show what they know? Or do artifacts of assessment format interfere with students demonstrating proficiency? In particular, how should emphases on communication skills and shared academic discourse patterns be mediated for special needs and language-minority students without implicitly setting lower standards for these groups? Are open-ended forms of assessment vulnerable to the same sorts of corruption from teaching-to-the test as traditional closed forms of assessment? Can students "pretend to know" by repeating formulaic routines? Returning to the points raised earlier about the close correspondence between validity across modes of assessment and teaching for transfer, what kinds of studies can be undertaken to address this relationship explicitly? If students can appear to be proficient if asked to perform in one way (e.g., paper-and-pencil circuit problems) but not if asked in another way (e.g., hands-on versions of the same electric circuit problems) (Shavelson, Baxter, & Pine, 1992), then is this a measurement problem or a learning problem? How might teachers use multiple modes of assessment to support development of flexible and robust understandings?

**Effects of assessment on learning and motivation.** Contemporary validity theory asks not only the question "Does the test measure what it purports to measure?" but "Does its use produce effects as intended?" (This concept is sometimes referred to as "consequential validity"; Messick, 1989). If formative use of assessments in classrooms is claimed to improve student learning, is this claim warranted? To find out how things work based on constructivist perspectives, it will be important to conduct studies in classrooms where instruction and assessment strategies are consonant with this model. In many cases this will mean "starting over again" and not assuming that findings from previous research studies can be generalized across paradigms. This will be especially important, for example, when conducting studies on topics such as feedback and motivation. The concept of "feedback" derives from the behaviorist model of learning and, as suggested previously, the great majority of studies available on feedback conform to behaviorist assumptions; instruction is of short duration, posttests closely resemble pretests and instructional materials, feedback is in the form of being told the correct answers, and so forth. New studies will be needed to evaluate the effect of feedback provided in ways that reflect constructivist principles, for example, as part of instructional scaffolding, assessment conversations, and other interactive means of helping students self-correct and improve. Similarly, the research literature on motivation makes sweeping claims about the risks of evaluating students, especially when they are tackling difficult problems. Yet, these findings are based on students' experiences with traditional, inauthentic and normative forms of assessment, where students took little responsibility for their own learning, and criteria remained mysterious. If the classroom culture were to be shifted dramatically, consistent with social-constructivist learning perspectives, then the effects of assessing students on difficult problems will have to be reexamined. The same is true for many other research areas as well. Likewise, when conducting comprehensive reviews or meta-analyses it will be important to consider the perspective represented and not aggregate studies across paradigms.

Although I have worked to merge cognitive and sociocultural perspectives, and taken together, to distinguish them from behaviorally oriented studies, there are nonetheless some important questions and controversies separating these two perspectives that should be addressed by a serious program of research. In this chapter, for example, I have taken the position that teachers should act as clinicians, using interpretive forms of data analysis as well as formal assessments, and I

emphasized social, motivational, and identity-producing aspects of classroom discourse practices, self-assessment, and the like. By contrast, more cognitively oriented approaches to assessment tend to emphasize the use of computer modeling to help diagnosis student thinking (Pellegrino, Baxter, & Glaser, in press). Although these two approaches hold most theoretical principles in common, as outlined in the conceptual framework, they disagree about the extent to which most of student learning in various domains can be formalized, that is, modeled by computer algorithms such that feedback from the machine could be as good as interacting with the teacher. Even if certain domains can be adequately specified to account for most student developmental pathways, I would argue that there are too many domains for the teaching day to be captured by the sum of a set of models. These are, of course, points of debate that should be addressed empirically.

To what extent are computer-delivered curricula effective in helping students learn challenging subject matter and develop habits of inquiry? What are the positive and negative side effects of using technology-based curricula? Do boys and girls participate equally? Or more or less equally than they do with hands-on, nontechnology-based curricula? Do students with less technology sophistication engage in the same science or history learning as students who are adept at using technology? Do students generalize inquiry skills and discourse practices, such as self-assessment and principled peer critique, to non-technology parts of the school day? Importantly, what are the effects of such embedded assessment projects on teacher learning? Do teachers develop richer understandings of student development because of the benchmarking provided by computerized assessments? Or do teachers learn less about students' understandings because the machine is doing the thinking? Are teachers marginalized as non-experts if branches of computer and Internet resources go beyond their own knowledge? What support do teachers need to model the role of learner in contexts where they are not expert? Of course, parallel questions should be asked regarding more clinical approaches to assessment, as I suggest below. My own view is that complex, new, cognitive and psychometric models are unlikely to be successful in creating an entire diagnostic and prescriptive system independent of teacher judgment; nonetheless, projects such as those described by Minstrell (1999) and the Cognition and Technology Group at Vanderbilt (1998) could serve as powerful professional development aids to help teachers become more insightful about techniques that provide access to students' thinking.

**Professional development of teachers.** Clearly, the abilities needed to implement a reformed vision of curriculum and classroom assessment are daunting—reminiscent of Cremin's (1961) earlier observation that progressive education required "infinitely skilled teachers." Being able to ask the right questions at the right time, anticipate conceptual pitfalls, and have at the ready a repertoire of tasks that will help student take the next steps requires deep knowledge of subject matter. Teachers will also need help in learning to use assessment in new ways. They will need a theory of motivation and a sense of how to develop a classroom culture with learning at its center. Given that new ideas about the role of assessment are likely to be at odds with prevailing beliefs, teachers will need assistance to reflect on their own beliefs as well as those of students, colleagues, parents, and school administrators. Because teachers' beliefs, knowledge and skills are pivotal in bringing about change in assessment practices, teachers' knowledge and beliefs should be a primary site for research.

In studies such as Frederiksen and White's (1997), where students have clearly benefited from inquiry-based curricula and reflective assessment practices, what have been the corollary effects on teachers' beliefs and practices? What supports have led to enactment of the vision, what impediments have subverted change? In Wilson and Sloane's (in press) study of the BEAR Assessment System, for example, improvements like those found by Frederiksen and White (1997) were again obtained in students' learning over and above the benefits from curricular change and teacher professional development. In addition, as a result of using the BEAR assessments and participating in scoring moderation sessions, teachers exhibited greater collegiality and used open-ended questions more than teachers in a reform-oriented comparison group, "which retained their rosy perceptions of alternative assessment strategies, but never really used them" (Roberts, Wilson, & Draney, 1997). While in theory all aspects of the reform are conceptually interrelated, practically speaking, how can teachers try out manageable segments of the reform (one subject area, or one instructional unit) so as to gain experience with these ideas in the context of their own practice? Although incremental change seems the most practical, what happens when conceptually incompatible systems are overlaid, as might occur when self-assessment is used alongside of traditional grading practices?

This chapter began with a portrayal of ideas from the past—about inherited ability, tracked curricula, atomistic conceptions of knowledge, and "scientific" measurement—that continue to shape educational practice and popular beliefs.

Against this backdrop, a reformed vision for classroom assessment was offered consistent with social constructivist principles. This vision may seem overly idealistic and optimistic given the demands it makes on teachers' knowledge and insight. Nonetheless, this vision should be pursued because it holds the most promise for using assessment to improve teaching and learning. To do otherwise means that day-to-day classroom practices will continue to reinforce and reproduce the status quo. Each time that teachers hold conferences with students, grade papers, ask students to explain their answers, or use results from a quiz to reorganize instruction, they are either following in the rut of existing practices and beliefs or participating in transforming the culture of the classroom. The task of implementing new assessment practices can be made easier if specific innovations are chosen to support and complement concomitant changes in curriculum and instruction. Indeed, attempts to improve instruction without corresponding changes in assessment are likely to be thwarted by powerful assumptions underlying assessment practices.

## References


Alexander, P. A. (Ed.). (1996). Special issue: The role of knowledge in learning and instruction. *Educational Psychologist, 31,* 89-145.

Allington, R. L. (1991). Children who find learning to read difficult: School responses to diversity. In E. H. Hiebert (Ed.), *Literacy for a diverse society: Perspectives, practices, and policies* (pp. 237-252). New York: Teachers College Press.

American Association for the Advancement of Science. (1993). *Benchmarks for science literacy*. New York: Oxford University Press.

Anderson, J. R., Reder, L. M., & Simon, H. A. (1996). Situated learning and education. *Educational Researcher, 25*(4), 5-11.

Atwell, N. (1987). *In the middle: Writing, reading, and learning with adolescents*. Portsmouth, NH: Heinemann.

Au, K. H. (1994). Portfolio assessment: Experiences at the Kamehameha Elementary Education Program. In S. W. Valencia, E. H. Hiebert, & P. Afflerbach (Eds.), *Authentic reading assessment: Practices and possibilities* (pp. 103-126). Newark, DE: International Reading Association.

Au, K. H., & Jordan, C. (1981). Teaching reading to Hawaiian children: Finding a culturally appropriate solution. In H. Trueba, G. P. Guthrie, & K. H. Au (Eds.), *Culture in the bilingual classroom: Studies in classroom ethnography* (pp. 139-152). Rowley, MA: Newbury House.

Au, K. H., & Valencia, S. W. (1997). The complexities of portfolio assessment. In N. C. Burbules & D. T. Hansen (Eds.), *Teaching and its predicaments* (pp. 123-144). Boulder, CO: Westview Press.

Ayers, L. P. (1918). History and present status of educational measurements. *Seventeenth Yearbook of the National Society for the Study of Education, Part II,* 9-15. Bloomington, IL: Public School Pub. Co.

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

Binet, A. (1909, 1973 ed.). *Les idees modernes sur les enfants.* Paris: Flammarion. (As cited in S. J. Gould, *The mismeasure of man*. New York: W. W. Norton, 1981.)

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

Bliem, C. L., & Davinroy, K. H. (1997). *Teachers' beliefs about assessment and instruction in literacy.* Unpublished manuscript, University of Colorado at Boulder, School of Education.

Bloom, B. S. (Ed.). (1956). *Taxonomy of educational objectives: The classification of educational goals*. New York: David McKay Company.

Bobbitt, F. (1912). The elimination of waste in education. *The Elementary School Teacher, 12,* 259-71.

Bransford, J. D. (1979). *Human cognition: Learning, understanding, and remembering.* Belmont, CA: Wadsworth.

Brown, A. L. (1994). The advancement of learning. *Educational Researcher, 23*(8), 4-12.

Brown, A. L., Bransford, J. D., Ferrara, R. A., & Campione, J. C. (1983). Learning, remembering, and understanding. In J. H. Flavell & E. M. Markman (Eds.), *Handbook of child psychology: Vol. 3. Child development* (4th ed., pp. 77-166). New York: Wiley.

Brown, A. L., & Campione, J. C. (1994). Guided discovery in a community of learners. In K. McGilly (Ed.), *Classroom lessons: Integrating cognitive theory and classroom practice* (pp. 229-270). Cambridge, MA: MIT Press / Bradford Books.

Brown, A. L., & Kane, M. J. (1988). Preschool children can learn to transfer: Learning to learn and learning from example. *Cognitive Psychology, 20,* 493-523.

Brownell, W. A. (1946). Introduction: Purpose and scope of the yearbook. *The forty-fifth yearbook of the National Society for the Study of Education, Part I, The measurement of understanding.* Chicago, IL: University of Chicago Press.

Budoff, M. (1974). *Learning potential and educability among the educable mentally retarded* (Final Report, Project No. 312312). Cambridge, MA: Research Institute for Educational Problems, Cambridge Mental Health Association.

Burns, M. (1993). *Mathematics: Assessing understanding.* White Plains, NY: Cuisenaire Company of America.

California Assessment Program. (1989). *A question of thinking: A first look at students' performance on open-ended questions in mathematics.* Sacramento, CA: California Department of Education.

California Learning Assessment System. (1994). *A sampler of science assessment —elementary*. Sacramento, CA: California Department of Education.

Callahan, R. E. (1962). *Education and the cult of efficiency: A study of the social forces that have shaped the administration of the public schools.* Chicago, IL: University of Chicago Press.

Cambourne, B., & Turbill, J. (1990). Assessment in whole-language classrooms: Theory into practice. *The Elementary School Journal, 90,* 337-349.

Camp, R. (1992). Portfolio reflections in middle and secondary school classrooms. In K. B. Yancey (Ed.), *Portfolios in the writing classroom* (pp. 61-79). Urbana, IL: National Council of Teachers of English.

Center for Civic Education. (1994). *National standards for civics and government*. Calabasas, CA: Author.

Claxton, G. (1995). What kind of learning does self-assessment drive? Developing a 'nose' for quality: Comments on Klenowski. *Assessment in Education, 2,* 339-343

Clay, M. M. (1985). *The early detection of reading difficulties* (3rd ed.). Portsmouth, NH: Heinemann.

Cobb, P., Yackel, E., Wood, T., Wheatley, G., & Merkel, G. (1988). Research into practice: Creating a problem solving atmosphere. *Arithmetic Teacher, 36,* 46-47.

Cognition and Technology Group at Vanderbilt. (1998). Designing environments to reveal, support, and expand our children's potentials. In S. A. Soraci & W. McIlvane (Eds.), *Perspectives on fundamental processes in intellectual functioning, Vol. 1* (pp. 313-350). Stamford, CT: Ablex.

Cohen, S. A. (1987). Instructional alignment: Searching for a magic bullet. *Educational Researcher*, *16*(8), 16-20.

Collins, A., Brown, J. S., & Newman, S. E. (1989). Cognitive apprenticeship: Teaching the crafts of reading, writing, and mathematics. In L. B. Resnick (Ed.), *Knowing, learning, and instruction: Essays in honor of Robert Glaser* (pp. 453-494). Hillsdale, NJ: Lawrence Erlbaum.

Corey, S. M. (1953). *Action research to improve school practices.* New York: Columbia University, Teachers College Bureau of Publications.

Cremin, L. (1961). *The transformation of the school: Progressivism in American education, 1876-1957*. New York: Vintage Books.

Cronbach, L. J. (1975). Five decades of public controversy over mental testing. *American Psychologist, 30,* 1-14.

Darling-Hammond, L. (1996). The right to learn and the advancement of teaching: Research, policy, and practice for democratic education. *Educational Researcher, 25*(6), 5-17.

Darling-Hammond, L., & Wise, A. E. (1985). Beyond standardization: State standards and school improvement. *The Elementary School Journal, 85,* 315-336.

Dewey, J. (1897). The university elementary school: History and character. *University Record, 2,* 72-5.

Dewey, J. (1902). *The child and the curriculum*. Chicago, IL: University of Chicago Press.

Duschl, R. A., & Gitomer, D. H. (1997). Strategies and challenges to changing the focus of assessment and instruction in science classrooms. *Educational Assessment, 4*(1), 37-73.

Dweck, C. (1986). Motivational processes affecting learning. *American Psychologist, 41,* 1040-1048.

Education U. S. A. (1968). *Individually prescribed instruction*. Washington, DC: Author.

Eisenhart, M., Finkel, E., & Marion, S. F. (1996). Creating the conditions for scientific literacy: A re-examination. *American Educational Research Journal, 33,* 261-295.

Elawar, M. C., & Corno, L. (1985). A factorial experiment in teachers' written feedback on student homework: Changing teacher behavior a little rather than a lot. *Journal of Educational Psychology, 77,* 162-173.

Ellwein, M. C., & Graue, M. E. (1996). Assessment as a way of knowing children. In C. A. Grant & M. L. Gomez (Eds.), *Making schooling multicultural: Campus and classroom*. Englewood Cliffs, NJ: Merrill.

Feuerstein, R. (1969). *The instrumental enrichment method: An outline of theory and technique*. Jerusalem: Hadassah-Wizo-Canada Research Institute.

Feuerstein, R. (1980). *Instrumental enrichment: An intervention program for cognitive modifiability*. Baltimore, MD: University Park Press.

Fleming, M., & Chambers, B. (1983). Teacher-made tests: Windows on the classroom. In W. E. Hathaway (Ed.), *Testing in the schools. New directions for testing and measurement* (No. 19, pp. 29-38). San Francisco, CA: Jossey-Bass.

Fordham, S., & Ogbu, J. U. (1986). Black students' school success: Coping with the burden of acting white. *Urban Review, 18,* 176-206.

Frederiksen, J. R., & Collins, A. (1989). A systems approach to educational testing. *Educational Researcher, 18*(9), 27-32.

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

Gagne, R. M. (1965). *The conditions of learning.* New York: Rinehard & Winston.

Gates, A. I., & Bond, G. L. (1936). Reading readiness: A study of factors determining success and failure in beginning reading. *Teachers College Record, 37,* 679-685.

Geertz, C. (1973). *The interpretation of cultures.* New York: Basic Books.

Geography Education Standards Project. (1994). *Geography for life: National geography standards 1994*. Washington, DC: National Geographic Research & Exploration.

Gipps, C. V. (1996). Assessment for learning. In A. Little & A. Wolf (Eds.), *Assessment in transition: Learning, monitoring and selection in international perspective.* Oxford: Pergamon Press.

Gipps, C. V. (1999). Socio-cultural aspects of assessment. In A. Iran-Nejad & P. D. Pearson (Eds.), *Review of research in education* (No. 24, pp. 355-392). Washington, DC: American Educational Research Association.

Glaser, B. G., & Strauss, A. L. (1967). *The discovery of grounded theory: Strategies for qualitative research.* New York: Aldine de Gruyter.

Glaser, R. (1984). Education and thinking: The role of knowledge. *American Psychologist, 39,* 93-104.

Goddard, H. H. (1920). *Human efficiency and levels of intelligence.* Princeton, NJ: Princeton University Press.

Goodman, K. S. (1973). Miscues: Windows on the reading process. In K. Goodman (Ed.), *Miscue analysis: Applications to reading instruction* (pp. 3-14). Urbana, IL: National Council of Teachers of English.

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

Gould, S. J. (1981). *The mismeasure of man.* New York: W. W. Norton.

Graue, M. E. (1993). Integrating theory and practice through instructional assessment. *Educational Assessment, 1,* 293-309.

Graves, D. (1983). *Writing: Teachers and children at work.* Portsmouth, NH: Heinemann.

Greeno, J. G. (1996, July). *On claims that answer the wrong questions.* Stanford, CA: Institute for Research on Learning.

Greeno, J. G., Collins, A. M., & Resnick, L. B. (1996). Cognition and learning. In D. C. Berliner & R. C. Calfee (Eds.), *Handbook of educational psychology* (pp. 15-46). New York: Simon & Schuster Macmillan.

Greeno, J. G., Smith, D. R., & Moore, J. L. (1993). Transfer of situated learning. In D. K. Detterman & R. J. Sternberg (Eds.), *Transfer of trial: Intelligence, cognition, and instruction* (pp. 99-167). Norwood, NJ: Ablex.

Heath, S. B. (1983). *Ways with words: Language, life, and work in communities and classrooms.* Cambridge, England: Cambridge University Press.

Heller, K. A., Holtzman, W. H., & Messick, S. (Eds.). (1982). *Placing children in special education: A strategy for equity.* Washington, DC: National Academy Press.

Herrnstein, R. J., & Murray, C. (1994). *The bell curve: Intelligence and class structure in American life.* New York: The Free Press.

Hiebert, E. H., & Raphael, T. E. (1998). *Early literacy instruction*. Fort Worth, TX: Harcourt Brace College Publishers.

Hilgers, T. (1986). How children change as critical evaluators of writing: Four three-year case studies. *Research in the Teaching of English, 20,* 36-55.

Hobbs, N. (Ed.). (1975). *Issues in the classification of children* (Vol. I). San Francisco, CA: Jossey-Bass.

Hogan, K., & Pressley, M. (1997). Scaffolding scientific competencies within classroom communities of inquiry. In K. Hogan & M. Pressley (Eds.), *Scaffolding student learning: Instructional approaches and issues.* Cambridge, MA: Brookline Books.

Holt, J. (1965). *How children fail.* New York: Pitman Publishing Company.

Hughes, B., Sullivan, H., & Mosley, M. (1985). External evaluation, task difficulty, and continuing motivation. *Journal of Educational Research, 78,* 210-215.

Hull, C. L. (1943). *Principles of behavior: An introduction to behavior theory.* New York: Appleton-Century.

International Reading Association / National Council of Teachers of English Joint Task Force on Assessment. (1994). *Standards for the assessment of reading and writing.* Urbana, IL: National Council of Teachers of English.

Jensen, A. R. (1969). How much can we boost IQ and scholastic achievement? *Harvard Educational Review, 39,* 1-123.

Klenowski, V. (1995). Student self-evaluation process in student-centered teaching and learning contexts of Australia and England. *Assessment in Education, 2,* 145-163.

Kliebard, H. M. (1995). *The struggle for the American curriculum: 1893-1958* (2nd ed.). New York: Routledge.

Klimenkov, M., & LaPick, N. (1996). Promoting student self-assessment through portfolios, student-facilitated conferences, and cross-age interaction. In R. Calfee & P. Perfumo (Eds.), *Writing portfolios in the classroom: Policy and practice, promise and peril.* Mahwah, NJ: Lawrence Erlbaum Associates.

Kluger, A. N., & DeNisi, A. (1996). The effects 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.

Koretz, D. M., & Barron, S. I. (1998). *The validity of gains in scores on the Kentucky Instructional Results Information System (KIRIS).* Washington, DC: RAND.

Koretz, D., Linn, R. L., Dunbar, S. B., & Shepard, L. A. (1991). The *effects of high-stakes testing on achievement: Preliminary findings about generalization across tests.* Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Kozol, J. (1991). *Savage inequalities: Children in America's schools.* New York: Crown Publishers.

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

Lepper, M. R., Drake, M. F., O'Donnell-Johnson, T. (1997). Scaffolding techniques of expert human tutors. In K. Hogan & M. Pressley (Eds.), *Scaffolding student learning: Instructional approaches and issues* (pp. 108-144). Cambridge, MA: Brookline Books.

Lewin, K. (1948). *Resolving social conflicts.* New York: Harper & Brothers.

Lincoln, Y., & Guba, E. (1986). *Naturalistic inquiry*. Beverly Hills, CA: Sage.

Linn, R. L. (1993). Linking results of distinct assessments. *Applied Measurement in Education, 6,* 83-102.

Lunt, I. (1993). The practice of assessment. In H. Daniels (Ed.), *Charting the agenda: Educational activity after Vygotsky*. New York: Routledge.

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.* Chestnut Hill, MA: Boston College, Center for the Study of Testing, Evaluation, and Educational Policy.

Maehr, M., & Stallings, W. (1972). Freedom from external evaluation. *Child Development, 43,* 117-185.

Malcolm, S. M. (Ch.). (1993). *Promises to keep: Creating high standards for American students.* Washington, DC: National Education Goals Panel.

Martin, J. R. (1992). *The schoolhome*. Cambridge, MA: Harvard University Press.

Mathematical Sciences Education Board. (1993). *Measuring up: Prototypes for mathematics assessment*. Washington, DC: National Academy Press.

Mazzeo, J., Schmitt, A. P., & Bleistein, C. A. (1993). *Sex-related performance differences on constructed-response and multiple-choice sections of Advanced Placement examinations* (CB Rep. No. 92-7; ETS RR-93-5). New York: College Entrance Examination Board.

McLaughlin, M., & Talbert, J. E. (1993). *Contexts that matter for teaching and learning: Strategic opportunities for meeting the nation's education goals*. Stanford, CA: Stanford University, Center for Research on the Context of Secondary School Teaching.

Mervar, K., & Hiebert, E. H. (1989). Literature-selection strategies and amount of reading in two literacy approaches. In S. McCormick & J. Zutell (Eds.), *Cognitive and social perspectives for literacy research and instruction, 38th Yearbook of the National Reading Conference* (pp. 529-535). Chicago, IL: National Reading Conference.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational* measurement (3rd ed., pp. 13-103). New York: American Council on Education, Macmillan Publishing Company.

Minstrell, J. (1989). Teaching science for understanding. In L. B. Resnick & L. E. Klopfer (Eds.), *Toward the thinking curriculum: Current cognitive research* (pp. 129-149). Alexandria, VA: Association for Supervision and Curriculum Development.

Minstrell, J. (1999). Student thinking, instruction, and assessment in a facet-based learning environment. In J. W. Pellegrino, L. R. Jones, & K. J. Mitchell (Eds.), *Grading the nation's report card: Research from the evaluation of NAEP*. Washington, DC: National Academy Press.

Moss, P. A. (1994). Can there be validity without reliability? *Educational Researcher, 23*(2), 5-12.

Moss, P. A. (1996). Enlarging the dialogue in educational measurement: Voices from interpretive research traditions. *Educational Researcher, 25*(1), 20-28, 43.

Myers, M. (1996). Sailing ships: A framework for portfolios in formative and summative systems. In R. Calfee & P. Perfumo (Eds.), *Writing portfolios in the classroom: Policy and practice, promise and peril*. Mahwah, NJ: Lawrence Erlbaum Associates.

National Center for History in the Schools. (1996). *National standards for history*. Los Angeles: Author.

National Council for the Social Studies. (1991). Testing and evaluation of social studies students. *Social Education, 55*, 284-286.

National Council for the Social Studies. (1994). *Curriculum standards for social studies*. Washington, DC: Author.

National Council of Teachers of Mathematics. (1989). *Curriculum and evaluation Standards for school mathematics*. Reston, VA: Author.

National Council of Teachers of Mathematics. (1995). *Assessment standards for school mathematics*. Reston, VA: Author.

National Council on Education Standards and Testing. (1992, January 24). *Raising standards for American education*. Washington, DC: Author.

National Forum on Assessment. (1995). *Principles and indicators for student assessment systems*. Cambridge, MA: National Center for Fair and Open Testing.

National Research Council. (1996). *National science education standards.* Washington, DC: National Academy of Sciences.

Newmann, F. M., & Associates. (1996). *Authentic achievement: Restructuring schools for intellectual quality*. San Francisco, CA: Jossey-Bass.

Oakes, J. (1985). *Keeping track: How schools structure inequality*. New Haven, CT: Yale University Press.

Oakes, J. (1990). *Multiplying inequalities: The effects of race, social class, and tracking on opportunities to learn math and science.* Santa Monica, CA: RAND.

Ogle, D. M. (1986). K-W-L: A teaching model that develops active reading of expository text. *The Reading Teacher, 39,* 564-570.

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

Pellegrino, J. W., Baxter, G. P., & Glaser, R. (in press). Addressing the "two disciplines" problem: Linking theories of cognition and learning with assessment and instructional practice. In A. Iran-Nejad & P. D. Pearson (Eds.), *Review of Research in Education*. Washington, DC: American Educational Research Association.

Perrenoud, P. (1991). Towards a pragmatic approach to formative evaluation. In P. Weston (Ed.), *Assessment of pupils' achievement: Motivation and school success* (pp. 77-101). Amsterdam: Swets and Zeitlinger.

Phillips, D. C. (1995). The good, the bad, and the ugly: The many faces of constructivism. *Educational Researcher, 24*(7), 5-12.

Phillips, D. C. (1996). Philosophical perspectives. In D. C. Berliner & R. C. Calfee (Eds.), *Handbook of educational psychology* (pp. 1005-1019). New York: Simon & Schuster Macmillan.

Popham, W. J., Cruse, K. L., Rankin, S. C., Sandifer, P. D., & Williams, P. L. (1985). Measurement-driven instruction: It's on the road. *Phi Delta Kappan, 66,* 628-634.

Pressley, M., Wood, E., Woloshyn, V. E., Martin, V., King, A., & Menke, D. (1992). Encouraging mindful use of prior knowledge: Attempting to construct explanatory answers facilitates learning. *Educational Psychologist, 27,* 91-109.

Resnick, L. B. (1987). Learning in school and out. *Educational Researcher, 16*(9), 13-20.

Resnick, L. B., & Klopfer, L. E., (Eds.). (1989). Toward the thinking curriculum: An overview. In L. B. Resnick & L. E. Klopfer (Eds.), *Toward the thinking curriculum: Current cognitive research* (pp. 1-18). Alexandria, VA: Association for Supervision and Curriculum Development.

Resnick, L. B., & Resnick, D. P. (1992). Assessing the thinking curriculum: New tools for educational reform. In B. R. Gifford & M. C. O'Connor (Eds.), *Changing assessments: Alternative views of aptitude, achievement, and instruction* (pp. 37-75). Boston: Kluwer Academic Publishers.

Roberts, L., & Wilson, M., & Draney, K. (1997). *The SEPUP assessment system: An overview* (BEAR Report Series SA-97-1). Berkeley: University of California, Berkeley Evaluation and Assessment Research Center.

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

Rogoff, B. (1991). Social interaction as apprenticeship in thinking: Guidance and participation in spatial planning. In L. B. Resnick, J. M. Levine, & S. Teasley (Eds.), *Perspectives on socially shared cognition* (pp. 349-364). Washington, DC: American Psychological Association.

Romberg, T. A., Zarinnia, E. A., & Williams, S. (1989). *The influence of mandated testing on mathematics instruction: Grade 8 teachers' perceptions*. Madison: University of Wisconsin, National Center for Research in Mathematical Science Education.

Rosenshine, B., & Meister, C. (1994). Reciprocal teaching: A review of the research. *Review of Educational Research, 64*, 479-530.

Ruch, G. M. (1929). *The objective or new-type examination: An introduction to educational measurement*. Chicago, IL: Scott Foresman.

Sadler, D. R. (1998). Formative assessment: Revisiting the territory. *Assessment in Education: Principles, Policy and Practice, 5*, 77-84.

Scriven, M. (1967). The methodology of evaluation. *AERA Monograph Series on Curriculum Evaluation, No. 1*, 39-83.

Shavelson, R. J., Baxter, G. P., & Pine, J. (1992). Performance assessments: Political rhetoric and measurement reality. *Educational Researcher, 21*(4), 22-27.

Shavelson, R., J., & Stern, P. (1981). Research on teachers' pedagogical thoughts, judgments, decisions, and behavior. *Review of Educational Research, 51*, 455-498.

Shepard, L. A. (1991a). Negative policies for dealing with diversity: When does assessment and diagnosis turn into sorting and segregation? In E. H. Hiebert (Ed.), *Literacy for a diverse society: Perspectives, practices, and policies*. New York: Teachers College Press.

Shepard, L. A. (1991b). Psychometricians' beliefs about learning. *Educational Researcher, 20*(6), 2-16.

Shepard, L. A. (1995). Using assessment to improve learning. *Educational Leadership, 52*(5), 38-43.

Shepard, L. A. (1997). *Measuring achievement: What does it mean to test for robust understanding*? Princeton, NJ: Educational Testing Service, Policy Information Center.

Shepard, L. A., & Dougherty, K. (1991, April). *Effects of high-stakes testing on instruction.* Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Shepard, L. A., Kagan, S. L., & Wurtz, E. (Eds.). (1998). *Principles and recommendations for early childhood assessments*. Washington, DC: National Education Goals Panel.

Shepard, L., Taylor, G., & Betebenner, D. (1998). *Inclusion of limited-English-proficient students in Rhode Island's grade 4 Mathematics Performance Assessment* (CSE Tech. Rep. No. 486). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Shulman, L. S., & Quinlan, K. M. (1996). The comparative psychology of school subjects. In D. C. Berliner & R. C. Calfee (Eds.), *Handbook of educational psychology* (pp. 399-422). New York: Simon & Schuster Macmillan.

Sizer, T. R. (1984). *Horace's compromise: The dilemma of the American high school.* Boston: Houghton Mifflin.

Skinner, B. F. (1938). *The behavior of organisms: An experimental analysis*. New York: Appleton-Century-Crofts.

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

Smith, J. P., diSessa, A. A., & Roschelle, J. (1993/94). Misconceptions reconceived: A constructivist analysis of knowledge in transition. *Journal of the Learning Sciences, 3,* 115-163.

Smith, M. L., Edelsky, C., Draper, K., Rottenburg, C., & Cherland, M. (1990). *The role of testing in elementary* schools (CSE Tech. Rep. No. 321). Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.

Stallman, A. C., & Pearson, P. D. (1990). Formal measures of early literacy. In L. M. Morrow & J. K. Smith (Eds.), *Assessment for instruction in early literacy* (pp. 7-44). Englewood Cliffs, NJ: Prentice Hall.

Starch, D., & Elliott, E. C. (1913). The reliability of grading high-school work in mathematics. *School Review, 21,* 254-259.

Sternberg, R. J. (1992). CAT: A program of Comprehensive Abilities Testing. In B. R. Gifford & M. C. O'Connor (Eds.), *Changing assessments: Alternative views of aptitude, achievement, and instruction* (pp. 213-274). Boston: Kluwer Academic Publishers.

Stipek, D. J. (1993). *Motivation to learn: From theory to practice* (2nd ed.). Boston: Allyn & Bacon.

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.

Sulzby, E. (1990). Assessment of emergent writing and children's language while writing. In L. M. Morrow & J. K. Smith (Eds.), *Assessment for instruction in early literacy* (pp. 83-108). Englewood Cliffs, NJ: Prentice Hall.

Terman, L. M. (1906). Genius and stupidity. A study of some of the intellectual processes of seven "bright" and seven "stupid" boys. *Pedagogical Seminary, 13*, 307-373.

Terman, L. M. (1916). *The measurement of intelligence*. Boston: Houghton Mifflin.

Tharp, R. G. (1997). *From at-risk to excellence: Research, theory, and principles for practice*. Santa Cruz: University of California, Center for Research on Education, Diversity and Excellence.

Tharp, R. G., & Gallimore, R. (1988). *Rousing minds to life: Teaching, learning, and schooling in social context*. New York: Cambridge University Press.

Thompson, P. W. (1995). Notation, convention, and quantity in elementary mathematics. In J. T. Sowder & B. P. Schappelle (Eds.), *Providing a foundation for teaching mathematics in the middle grades.* Albany: State University of New York Press.

Thorndike, E. L. (1922). *The psychology of arithmetic.* New York: Macmillan.

Tobin, K., & Ulerick, S. (1989, March). *An interpretation of high school science teaching based on metaphors and beliefs for specific roles*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Tyack, D. (1974). *The one best system: A history of American urban education*. Cambridge, MA: Harvard University Press.

Tyler, R. (1938). *Thirty-seventh yearbook of the National Society for the Study of Education, Part II*. Bloomington, IL: Public School Publishing Company.

Tyson-Bernstein, H. (1988). A conspiracy of good intentions: The textbook fiasco. *American Educator, 12*, 20, 23-27, 39.

Valencia, R. R. (1997). *The evolution of deficit thinking: Educational thought and practice.* London: The Falmer Press.

von Glasersfeld, E. (1995). *Radical constructivism: A way of knowing and learning.* London: Falmer Press.

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

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

Wiggins, G. (1992). Creating tests worth taking. *Educational Leadership, 49*(8), 26-33.

Wiggins, G. (1993). Assessment: Authenticity, context, and validity. *Phi Delta Kappan, 74,* 200-214.

Wile, J. M., & Tierney, R. J. (1996). Tensions in assessment: The battle over portfolios, curriculum, and control. In R. Calfee & P. Perfumo (Eds.), *Writing portfolios in the classroom: Policy and practice, promise and peril.* Mahwah, NJ: Lawrence Erlbaum Associates.

Wilson, M., & Sloane, K. (in press). From principles to practice: An embedded assessment system. *Applied Measurement in Education.*

Wineburg, S. S. (1996). The psychology of learning and teaching history. In D. C. Berliner & R. C. Calfee (Eds.), *Handbook of educational psychology* (pp. 423-437). New York: Simon & Schuster Macmillan.

Wittrock, M. (1986). Students' thought processes. In M. Wittrock (Ed.), *Handbook of research on teaching* (pp. 297-327). New York: Macmillan.

Wolf, D. P., & Reardon, S. F. (1996). Access to excellence through new forms of student assessment. In J. B. Baron & Wolf, D. P. (Eds.), *Performance-based student assessment: Challenges and possibilities* (pp. 1-31). Chicago, IL: University of Chicago Press.

Wood, D., Bruner, J., & Ross, G. (1976). The role of tutoring in problem solving. *Journal of Child Psychology and Psychiatry, 17,* 89-100.

Yackel, E., Cobb, P., & Wood, T. (1991). Small-group interactions as a source of learning opportunities in second-grade mathematics. *Journal for Research in Mathematics Education, 22,* 390-408.

Yancey, K. B. (1996). Dialogue, interplay, and discovery: Mapping the role and the rhetoric of reflection in portfolio assessment. In R. Calfee & P. Perfumo (Eds.), *Writing portfolios in the classroom: Policy and practice, promise and peril.* Mahwah, NJ: Lawrence Erlbaum Associates.