DIMENSIONS OF THINKING: IMPLICATIONS FOR TESTING

CSE Technical Report 282

Robert L. Linn

University of Colorado

UCLA Center for Research on Evaluation,
Standards, and Student Testing

September, 1988

The research reported herein was conducted with partial support from the U.S. Department of Education, Office of Educational Research and Improvement, pursuant to Grant No. G0086-003. However, the opinions expressed do not necessarily reflect the position or policy of this agency and no official endorsement by this agency should be inferred.

This report will appear as a chapter in *Dimensions of Thinking and Cognitive Instruction* (B.F. Jones and L. Idol, editors), to be published by Lawrence Erlbaum Associates.

Please address inquiries to: CSE Dissemination Office, UCLA Graduate School of Education, 405 Hilgard Avenue, Los Angeles, California, 90024-1521

This report and the associated efforts by the Association for Supervision and Curriculum Development and the North Central Educational Laboratory to encourage the teaching of thinking attest to the fact that there is great interest in the topic of thinking, especially in those skills and processes that have come to be called "higher order thinking skills." Pleas for greater emphasis on higher order thinking skills are plentiful. This is apparent not only in the rash of reports on the status of education that have appeared in the past five years (e.g., The National Commission on Excellence in Education, 1983; National Assessment of Educational Progress, 1985), but can be seen in an array of educational journals and periodicals. Partly as a result, a wide variety of instructional materials and teaching approaches devoted to improving the teaching of problem solving, comprehension, critical thinking, metacognitive skills, and other higher level thinking processes have been developed.

Testing and assessment have not been ignored in the recent emphasis on thinking. The Alexander-James (1987) report of the Study Group on the Nation's Report Card, for example, strongly urged that the National Assessment of Educational Progress place greater emphasis on the assessment of higher order thinking skills. It is apparent from even a cursory review of the brochures of test publishers or a visit to their exhibit booths at a professional meeting that the message that testing needs to include higher order thinking skills has not been missed by the test publishers.

Concerns About Current Approaches

When testing or assessment is mentioned in discussions of ways to improve the teaching of thinking skills, however, three themes are most common. First, dissatisfaction with current tests is expressed frequently because they are seen as placing too much emphasis on simple factual knowledge. Second, suggestions are made that the approaches to assessment need to be expanded beyond paper-and-pencil tests, in general, or—more forcefully and specifically—beyond standardized multiple-choice tests, in particular. Finally, the notion is often expressed that tests, or more broadly conceived assessment systems, signal what is held to be important to teachers, parents, and students. Thus, for the teaching of thinking to be recognized as important and given enough emphasis, it is necessary to develop assessment procedures that do justice to the goals.

This last concern, that tests are devices that direct instruction, is critical to the interpretation of criticisms of tests and to discussions about needed changes. The degree to which tests influence what teachers do and what students learn depends heavily on the uses that are made of test results. Regardless of one's position in the long-standing debate about whether tests should follow or lead instruction, it is clear that tests can be an important factor in shaping instruction and learning.

"Will this be on the test?" is a question that is all too familiar to teachers. Although not a question that most teachers like to hear, it is easily understood in terms of Walker's (1983) observation that "the things that are *really* important, as every student knows, are the things that appear on tests and are used in grading" (p. 173).

Traditionally, of course, it is teacher-made tests, not standardized tests, that determine student grades. In the past, most standardized tests had little influence on learning and instruction because they did not influence student grades and, because most schools made relatively little use of the results, they had little impact on teachers. Certainly, there are exceptions such as the New York State Regents Examinations and the College Board Advanced Placement Tests; however, before the start of the minimum-competency test movement and the great expansion of the use of tests as a means of gaining greater accountability, most uses of standardized achievement tests were relatively benign.

Today, however, there are examples throughout the country of testing programs that have been introduced for the explicit purpose of directing education. Tests have become the major tool of policymakers for implementing educational reform (see, for example, Linn, 1987; Madaus, 1985; Pipho, 1985). By their actions, policymakers have taken the affirmative position on the role of tests in directing teaching and learning.

This changed context is important for considering the concern that tests overemphasize factual knowledge and the concern that the multiple-choice format is inadequate for testing more important, higher order thinking skills. For example, research that shows that multiple-choice tests correlate as well or better with essay tests as the latter do with each other or that shows that multiple-choice tests are better predictors of grades than are essay tests is relevant when the goal is prediction. Such evidence is not convincing, however, when the purpose of the test is to direct instruction and learning. Similarly, the content validity evidence that shows the match between the questions on a test and the content and exercises in, say, science textbooks is not sufficient for the person who is dissatisfied with the textbooks and argues for the importance of hands-on experiences.

In the following sections of this report, the concerns about the overemphasis on factual knowledge, the constraints of multiple-choice tests, and the inadequacy of current tests as targets that will enhance thinking are discussed. That discussion is followed by a consideration of alternatives to current approaches to assessment and changes in testing that are needed in order for tests and other assessment procedures to contribute more to the teaching of thinking.

Factual Knowledge

Dissatisfaction with current tests has a variety of sources. For the focus of this discussion, however, a primary concern is the emphasis that teacher-constructed tests and standardized tests place on the simple recall of factual knowledge. Analyses of teacher-constructed tests (e.g., Fleming & Chambers, 1983; Stiggins, Griswald, & Green, 1988) indicate that questions that require the use of higher order thinking skills are the exception rather than the rule. Questions that ask students to recall information are much more common than questions that require comparisons, inferences, or evaluation.

As has been highlighted in several reviews, the emphasis on factual knowledge is not limited to teacher-constructed tests. The classification of questions on standardized tests by the level of skill required is equally discouraging. Frank (1978), for example, classified 765 items from 12 standardized science tests using the first four developmental levels in Bloom's (1956) taxonomy. Only 2% of the items were placed in the two higher categories: application and analysis. In contrast, 78% of the items were placed in the lowest category: the simple recall of factual information. In another review of science achievement tests, Morgenstern and Renner (1984) found that 90% of the items on the tests that they reviewed required only the recall of factual information. Others, such as Bowman and Peng (1972) and Levine, McGuire, and Nattress (1970), have reported similarly discouraging results even for standardized tests used at the graduate and professional school level.

The perceived overemphasis on low-level skills is not limited to science tests. Similar concerns have been expressed, for example, by critics of standardized reading tests. According to the report of the National Academy of Education's Commission on Reading (1985), "reading is the process of constructing meaning from written words" (p. 7). Skilled reading requires the reader to think and interrelate information from the printed page with prior knowledge in order to construct meaning. It requires the reader to monitor his understanding. Component skills of decoding and fluent word recognition are certainly important, but they are insufficient.

Reviews of standardized reading tests (e.g., Cross & Paris, 1987; Linn & Valencia, 1986; Valencia & Pearson, 1986), however, reveal a heavy emphasis on the component skills. Items that require the recognition of word meanings and literal comprehension make up a majority of the questions on many standardized tests. Items that attempt to assess the test taker's ability to derive meaning from a passage and to make inferences are often limited to questions such as the following: What is the main idea in this story? What is this story mostly about? What is the best title for this story? How did the character probably feel? These are not bad questions. However, a close inspection often reveals that such questions can be answered using information that is explicitly stated in the text.

For survey tests that are not used to make decisions about students, teachers, or programs, and that are not used to rank order schools, there is less reason to be concerned about the number of items placed at various levels of Bloom's taxonomy, the limited degree of integration of information across different segments of a text, or the limited nature of inferences that a student must make to answer reading comprehension questions. Scores based on such items are apt to correlate quite highly with scores based on tests that critics would judge to require greater amounts of integration of information, more complicated inferences, and the solution of novel problems. Thus, if the test is used only as a proxy for the more difficult to measure, albeit more important, outcomes of instruction, then the high correlations between the two types of measures are reassuring.

When schools can, and do, purchase materials specifically designed to increase scores on specific standardized achievement tests because of the increased importance that policymakers, the press, and the general public are placing on the test scores, then predictive validity is not satisfying. As Brown (1987) has argued:

> Insecure teachers and schools under pressure to raise achievement levels blatantly teach to the tests. Since the tests require little thoughtfulness, the instruction and curriculum that revolve around them remain stuck at a very basic level. Under such circumstances, the demands for accountability intrude on teaching and learning time and warp instruction in ways that may often raise test scores even as they lower the quality of the education being delivered. (p. 51)

Constraints of Standardized, Paper-and-Pencil Tests

Even under the best of circumstances, it is difficult to write good questions to assess deeper understanding. However, the difficulty is exacerbated by the constraints of standardized testing. Among the more important constraints are the need for efficiency, the desire to represent the test results by a single number that places each test taker on a common scale, and the emphasis on the use of test results for purposes of accountability. The emphasis on accountability limits the domain of the test, and, I will argue, also reduces the instructional utility. The desire for a single score contributes to the emphasis on well-structured problems with a single right answer. The need for efficiency limits tests to machine-scorable formats, which usually means multiple-choice test items.

Multiple-choice questions that require much more than simple recognition can be written. Consider the following illustrative item from Ebel and Frisbie (1986):

> If the radius of the earth were increased by 3 feet, its circumference at the equator would be increased by about how much?
> a. 9 feet
> b. 12 feet
> c. 19 feet
> d. 28 feet. (p. 171)

Ebel and Frisbie argue that items such as the above present "novel problem situations [that] reward the critical-minded student who has sought to understand what he/she was taught and penalize the superficial learner" (p. 171). The key is that the problem is novel.

Students obviously could be given a formula for calculating the increase in the circumference as a function of an increase in the radius. Memorization of the formula and practice with similar problems would destroy the novelty of Ebel and Frisbie's problem and turn what was once an item that was sensitive to the degree to which students understand fundamental principles and can use those principles to solve problems into a problem that simply requires the recognition of a problem type and the application of a memorized formula with or without understanding the reasons why the formula works.

Problem structure. Norman Frederiksen (1984) provided an excellent discussion of these constraints in an article titled "The Real Test Bias: Influences of Testing on Teaching and Learning." Frederiksen argued that items on standardized tests typically present well-structured problems, that is, ones that are clearly stated: "All the information needed to solve the problem is available in the problem or (presumably in the head of the student) and an algorithm exists that guarantees a correct solution if properly applied" (p. 199). He went on to argue that most important problems, both in and out of school, are ill-structured in the sense defined by Simon (1978). They may not have a single correct answer, much less provide all the information needed to solve the problem or be soluble by applying a previously learned algorithm. Ill-structured problems often contain ambiguities. Problem identification and representation, the generation of hypotheses, planning, and the exploration of solution strategies are of central importance.

Although multiple-choice items can tap higher order thinking and problem solving skills, as the above example item from Ebel and Frisbie (1986) illustrates, such items are difficult to write and usually involve well-structured problems. Multiple-choice items that get at the thinking processes that are crucial in the solution of ill-structured problems are even rarer. Even if multiple-choice items that are effective at measuring the thinking processes required to solve ill-structured problems can be obtained, the problems themselves are apt to be considered to be unacceptable on a test that is used to hold students accountable. As Frederiksen (1984) noted, such problems could almost certainly be attacked as unfair. Challenges to minimum competency tests support this conclusion. In the Debra P. case, for example, the Fifth Circuit Court of Appeals ruled that "fundamental fairness requires that the state be put to test on the issue of whether the students were tested on material they were or were not taught" (*Debra P. v. Turlington,* 1981). Although thinking skills and strategies can be taught, teaching a correct solution path for a particular ill-structured problem that might be included on a test would destroy the purpose by converting the problem into a simple recall question.

The report of the National Academy of Education's Committee that reviewed the recommendations of the Alexander- James study group on the *Nation's Report Card* (National Academy of Education, 1987) provided the following important caution in this regard:

> It is all too easy to think of higher-order skills as involving only difficult subject matter as, for example, learning calculus. Yet one can memorize the formulas for derivatives just as easily as those for computing areas of various geometric shapes, while remaining equally confused about the overall goals of both activities. All subjects have a basic knowledge component that can be taught by drill and practice. This basic knowledge, while prerequisite to competence, is also distinct from the intellectual skills of gathering relevant information, evaluating evidence, weighing alternative courses of action, and articulating reasoned arguments. (p. 54)

The obvious implication of this caution is that it is important not to confuse the apparent difficulty of subject matter with the demands that a test problem makes for the exercise of thinking skills as conceived by the National Academy of Education Committee.

Efficiency. Efficiency looms as another major obstacle to the development of assessment procedures that will better serve the goals of teaching thinking skills. As Frederiksen (1984) noted, "efficient tests tend to drive out less efficient tests, leaving many important abilities untested— and untaught" (p. 201). Multiple-choice tests are certainly efficient, but as has already been stated, they tend to emphasize factual knowledge at the expense of inference, analytical thinking, problem solving, and the application of knowledge to novel situations.

The area of writing provides an excellent example of the tendency for efficiency to govern the nature of testing. The most recent National Assessment results on the performance of students in writing suggested that all is not well (Applebee, Langer, & Mullis, 1986). This should be no surprise, given the amount of writing most students are required to do in school and the nature of the tests that are most often used to assess writing skills. Linn and Palmer (1985) discussed the role of testing in relationship to the teaching of writing—a skill that clearly requires the application of thinking skills—along the following lines.

Chall's (1977) analysis suggested that textbooks, even those in grammar and composition, give little emphasis to writing. Assignments are more apt to require "underlining, circling, and filling in single words" (p. 64). Multiple-choice tests do little to emphasize the importance of writing. Only one of the two College Board Achievement Tests in English composition requires any actual writing, and that one allocates only 2 minutes to writing an essay and the remaining 40 minutes to multiple-choice items (paraphrased from Linn & Palmer, 1985).

There are, of course, many reasons for the scarcity of tests that require actual writing rather than choosing among alternative sentences or the identification of grammatical errors. Essays are expensive to score and yield lower reliabilities than can be achieved in a fixed period of time using multiple-choice items. Research has shown that multiple-choice tests can be constructed to correlate relatively well with essay tests; indeed, the correlations sometimes approach the limits set by the reliabilities of the tests. Essay tests are also unlikely to enhance the predictive validity that is provided by a multiple-choice test. Thus, it is difficult to justify such tests in terms of traditional criteria of reliability and predictive validity, particularly if the substantial difference in efficiency is given any weight. However, as has been noted previously, those should not be the only, or even the primary, considerations for tests that are being used to direct instruction and learning.

Adequacy of Instructional Targets

Writing, like other applications of thinking to solving problems, requires practice. It deserves to be emphasized by students and teachers. But tests send messages to students and teachers that stress the importance of mastering particular skills. To convey the message that writing is of great importance, more emphasis needs to given to tests that actually require students to write, even at the expense of some reduction in reliability (Linn & Palmer, 1985).

In the case of writing, some progress has been made in recent years. Writing samples are included in a number of state and district assessment systems. The scores that are derived, however, are often of less interest than the examples of the essays that students produce. Certainly, from an instructional perspective the comments and suggestions that a teacher makes on a student's written work are apt to be more useful in helping students improve their writing than is the summary grade assigned. That is also apt to be true of other assessments of students' thinking skills.

Consider, for example, the relatively simple case of story problems in arithmetic. I say relatively simple, because the problems are well-structured and yield a single correct answer. Furthermore, an effective job of testing a student's ability to solve such problems can be done using multiple-choice items. However, the instructional utility of the global score provided by a standardized test is limited. It can provide a reasonable indication of a student's overall proficiency level, but a low score doesn't reveal where a student's difficulty lies. The arithmetic operations needed to solve most story problems are less likely to be the major source of difficulty than are other aspects of the problem.

Often the most difficult step in solving a story problem is the formation of a coherent representation of the problem. As Mayer (1985; see also Snow & Lohman, in press) suggests, problem representation requires translation—that is, the use of linguistic and factual knowledge to understand the problem statement—and integration, or use of schema knowledge to create a representation of the problem. The distinction is illustrated by Riley, Greeno, and Heller's (1983) review of research using story problems. Consider, for example, the following two versions of what Riley et al. refer to as a combine problem:

Version 1: "Joe and Tom have eight marbles altogether.
Joe has five marbles.
How many marbles does Tom have?"

Version 2: "Together, Tom and Joe have eight marbles.
Three of these marbles belong to Tom.
How many of them belong to Joe?"

Although the two versions of the problem require the same formal operations, many students who are unable to solve problems like the first version give the correct answer when the problems are stated in the form of the second version. These children clearly have the procedural knowledge necessary to perform the subtraction, but they apparently have difficulty in being able to "represent the relationships among the quantities described in the problem situations in a way that relates to available solution procedures" (Riley, et al., 1983, p. 173). Understanding the reasoning processes used by the student would be of much greater value to a teacher than simply knowing that the student got a low score on the problem solving section of a standardized test containing story problems.

According to Glaser (1986) "novices recognize the surface features of a problem or task situation and more proficient individuals go beyond surface features and identify inferences or principles that subsume the surface structure" (p.55). Thus, in the assessment of a student's problem-solving skills it may be more important to attempt to get at the student's representation of problems or mental models than it is to determine whether the student selected a correct answer on a multiple-choice test.

Alternative Approaches to Assessment

Standardized achievement tests are quite efficient at measuring a student's accumulation of declarative knowledge. They also can efficiently measure a student's success in applying established procedures to solve problems with specified characteristics. For purposes of comparing the general level of knowledge in a given subject domain or predicting future performance, current standardized tests are quite effective. The assessment of a student's general achievement level is useful, but there are serious doubts about the degree to which these tests assess higher order thinking skills such as creative thinking, skills of planning, problem solving, and metacognition. Moreover, current standardized tests do not provide information about the thinking processes that a student is using or help in the diagnosis of misconceptions or particular sources of difficulty. Alternative assessment procedures are needed that focus on specific higher order thinking skills and that will provide information about thinking processes such as the formation of problem representation, the construction of a mental model, the generation of hypotheses, the planning of steps to solve problems, and the self-monitoring of the application of a solution strategy. Alternative approaches that yield diagnostic information that can help guide instruction are also needed.

There are a number of efforts that recently have been undertaken that attempt to overcome some of the limitations that have been discussed. The remainder of this paper is focused on a few such efforts that appear particularly promising.

Teacher-Directed Assessment

Although the focus in the following sections of this paper is on assessment procedures that are intended for large-scale use by districts or states, it should be recognized that externally imposed assessment represents only a fraction of the total amount of time and effort that goes into student assessment. More time and effort is devoted to teacher-constructed and teacher-selected assessment (e.g., tests and exercises accompanying curriculum materials) than to assessments required by districts and states. More importantly, it is the teacher-directed assessments "that most strongly influence student learning and academic self-concept" (Stiggins, 1988, p. 368). Thus, it is essential that attention be given not only to ways of improving the assessment of thinking skills by commercial publishers, districts, and states, but to ways of helping teachers develop and select assessment procedures that encourage thinking.

Many of the concepts and ideas that are discussed in the following sections could be useful as general principles to consider when teachers develop or select their own assessments. However, it is clear that much more than these general ideas are needed. As Stiggins (1988) has convincingly argued, there is a need to give assessment issues much higher priority in pre-service and in-service teacher training. "The available research suggests that pre-service instruction/curriculum in educational assessment is not adequate to develop the desired skills" (Gullickson & Hopkins, 1987, p. 15).

Materials and workshops provided by researchers at the Northwest Regional Educational Laboratory (Stiggins, 1987; Stiggins, Rubel, & Quellmalz, 1986) represent one significant effort to help teachers improve their assessment procedures, in general, and their assessment of thinking skills, in particular. The materials and workshops provide a framework for planning oral, paper-and-pencil, and performance forms of assessment that include five defined levels of thinking skills: recall, analysis, comparison, inference, and evaluation (Quellmalz, 1985). The simple creation of a chart with three rows for the forms of assessment and five columns for the levels of thinking skills encourages the planning of assessments that do more than cover the "recall" row of the matrix. The materials provide good examples of questions that require the higher level skills of analysis, comparison, inference, and evaluation (Stiggins, Rubel, & Quellmalz, 1986). These materials, together with the practice provided by the workshop, have been found to have a positive impact on teacher attitudes toward and self-reports of the use of procedures of assessing higher order thinking skills (Stiggins, no date).

Reading Assessment

There has been a renewed interest in alternative approaches to the measurement of reading comprehension during the past few years. This interest has been stimulated by research that has led to changes in the conception of the reading process. As stated by Wixson and Peters (1987):

> Current research suggests that reading is a process of constructing meaning through the dynamic interaction of the reader, the text, and the context of the reading situation (Wixson & Peters, 1984). This view of reading focuses on how the reader builds meaning from print; what the reader brings to the reading situation in terms of experience, knowledge, skills, and motivation; how information is presented in written text; and the effects of context on reading performance. (pp. 333-334)

This conception of reading, which is consistent with that expressed by the National Academy of Education's Commission on Reading (1985) and a number of other authors (e.g., Curtis & Glaser, 1983; Valencia & Pearson, 1986), places heavy emphasis on the reader as an active agent and stresses the importance of prior knowledge (Bransford & Johnson, 1973; Johnston, 1983; Pearson & Spiro, 1980). It also focuses attention on the "coordination of a number of interrelated sources of information" (National Academy of Education, 1987, p. 7) and the use of higher level integrative and metacognitive skills. Consistent with this view of reading, the recent efforts to develop alternative approaches to assessing reading place more emphasis on the holistic process of constructing meaning and less emphasis on interrelated component skills.

Two important examples of current efforts to develop reading assessments that are consistent with this emerging view of the reading process are the ongoing efforts for the state assessment programs in Illinois and in Michigan. These two efforts, together with an earlier effort in New York that led to the development of the Degrees of Reading Power (DRP) test (College Board, 1986; Koslin, Zeno, & Koslin, 1987), provide illustrations of alternative approaches to the measurement of reading comprehension that emphasize understanding and the integration of information in written text.

The DRP tests are intended to provide "holistic measures of how well the messages within a text are understood" (College Board, 1986, p. 1) rather than a measure of isolated skills. On the surface, the test appears similar to a simple cloze test because, as is true of the latter, words are deleted from a passage, and the test taker's task is to choose from the available options the word that has been deleted. There are, as Linn and Palmer (1985) have noted previously, a number of important differences between the DRP and the traditional cloze test, however. The choice of words to delete is based on a theoretical conception that dictates that the difficulty of identifying the missing word should depend on the difficulty of understanding the surrounding text and not on the difficulty of the word that is deleted. Thus, only a small number of deletions are made per passage (usually only one or two words per paragraph) and the deletions are selected from words that have a higher frequency of occurrence in written text and a greater familiarity than much of the vocabulary in the passage that must be understood to comprehend the surrounding text.

The design of the DRP is intended to assure that "processing the surrounding prose is both necessary and sufficient for choosing the right answer" (Koslin, Koslin, & Zeno, 1979, p. 316). To assure that processing involves integration of information across sentences and not simply the processing of the individual sentence from which a word is deleted, distractors are carefully selected to be appropriate responses when the sentence is considered in isolation. Thus, in order to distinguish between the right answer and the distractors, the test taker "is required to integrate the meaning across several sentences" (Linn & Palmer, 1985, p.94).

Although much shorter and less demanding in terms of the amount of text that a test taker must be able to integrate on the actual test, a sample item from Form PX-1 of the DRP can be used to illustrate the need to rely on more than the single sentence from which a word is deleted:

> It had been sunny and hot for days.
> Then the s-1 changed.
> It turned cloudy and cool.

The five options for this sample item ("price, road, job, weather, and size") are all acceptable if only the second sentence is considered. "When integrated with the information in the preceding and subsequent sentence, however, only 'weather' leads to a coherent set" (Linn & Valencia, 1986, pp. 27-28).

Although the DRP is consistent with the concept of reading as an integrative process, it does not attempt to measure more than "the ability to make inferences concerning the surface meaning of messages in English prose" (Koslin, Zeno, & Koslin, 1987, p. 13). It does not attempt to measure, for example, the ability of a test taker to integrate information from the text with prior experience or background knowledge in order to draw inferences that go beyond the text. Nor is an attempt made to assess metacognitive skills or the degree to which a test taker is able to form "a coherent cognitive model of the text meaning" (Johnston, 1984, p. 236).

The more recent efforts to develop measures of reading comprehension that are currently being undertaken for the Illinois and the Michigan State Departments of Education seek to bring the measurement of reading performance closer to the current theoretical conceptions of the reading process. Although there are a number of differences between the Illinois and Michigan efforts, they share a common conceptual framework and the assessments have many parallels. An excellent description of the work related to the Michigan Educational Assessment Program is provided by Wixson and Peters (1987; see also, Cross & Paris, 1987; Wixson & Peters, 1984; and Wixson, Peters, Weber, & Roeber, 1987). Partial descriptions of the Illinois assessment work are provided by Valencia (1988) and by Kerins (1988) and a comprehensive report by the four principal developers of the assessment (Sheila Valencia, P. David Pearson, Robert Reeve, and Timothy Shanahan) is in preparation.

The Illinois and Michigan assessments use passages that are several times as long as those those found on typical standardized reading tests. The passages consist of full-length text of the sort that children might encounter in the classroom or read outside of school. In both assessment programs, the narrative passages are short children's stories that were selected from such sources as "children's magazines and grade-appropriate literature anthologies" (Wixson and Peters, 1987, p. 340). The expository passages were selected from textbooks and provide information and explanations of the type that students are expected to read and comprehend as part of their school work.

The longer, more coherent passages were selected because it was expected that they would provide a better basis for constructing questions that require test takers to integrate information from various parts of the passage and to draw inferences. Of course, length alone assures neither that the text lends itself to questions requiring integration and the drawing of inferences nor that the questions will require these abilities. Systematic procedures are needed to analyze the text and guide the construction of questions. The Illinois and Michigan efforts rely on the construction of story maps that display the structure of the text and serve as guides to the construction of questions (see Wixson & Peters, 1987, for a description and illustration of the story mapping used in Michigan).

The Illinois and Michigan assessments emphasize integrative processing and the role of the reader in constructing meaning from the text. Questions for the assessments are even referred to as "constructing meaning" questions to highlight this perspective. Test takers are required to use information from various parts of the passage to draw inferences. They also are expected to use the information in situations or contexts other than that presented in the particular passage. These characteristics are illustrated by the three types of processing that Wixson and Peters (1987) describe as required by the constructing meaning questions. The three types are:

> ...*intersentence*, which requires the reader to construct meaning from text of one to three contiguous sentences in length; *text*, which requires the reader to integrate information within sections of text larger than several sentences as well as across the entire text; and *beyond text*, which requires readers to rely heavily on information from their own experiences in addition to information in the text. (p. 346)

In addition to the constructing meaning scores, the two assessment systems attempt to assess two other components: (a) the test taker's prior knowledge that is specific to the passage (referred to as "topic familiarity" in both programs), and (b) his metacognitive skills (referred to as "reading strategies" in Illinois and "knowledge about reading" in Michigan). There are differences in the approaches to the measurement of these two components, but the similarity of intent is clear.

One notable difference in the constructing meaning measures is that the Illinois assessment uses a multiple-right-answer question format rather than the more traditional single-right-answer multiple-choice format that is used in Michigan. As Snow and Lohman (in press) have noted, "most comprehension problems in school and the world involve multiple or alternative or optional correct answers, depending on contextual circumstances" (p. 96 of typescript). The use of questions that have one, two, or three of the options keyed as correct in the Illinois assessment reflects this perspective and encourages test takers to consider more than one interpretation.

Although the Illinois and Michigan assessment procedures have many appealing features, it is important to recognize that many questions about the procedures remain to be answered and that the instruments are designed for group assessment rather than for making decisions about individual students. The use of long passages means that the sampling of topics for a given student is more limited than it is on a test with several short passages. For group assessment different passages can be administered to different subsamples of students so that the coverage is broader for the group as a whole, but for individual student measurement the generalizability of the results across passages is an open question, and it should not be concluded that what works well for group assessment will necessarily meet the needs of individual student assessment.

The topic familiarity and study strategy (or knowledge about reading) sections of the assessments break new ground. Until a good deal more research has been completed that leads to a better understanding of the properties of these measures and their construct validity, however, they are best viewed as promising experimental approaches.

Critical Thinking Skills: History-Social Science

The assessment of critical thinking skills in various content areas has received increasing attention in the past few years. Programs such as the California state-wide assessment in history-social science illustrate the importance that is being attached to the use of critical thinking skills in a subject matter domain. This assessment was developed in response to a mandate from the California State Board of Education in 1982. A state-wide committee of teachers and curriculum leaders was formed to identify areas for assessment and establish assessment priorities; this committe "placed a high priority on critical thinking—a priority supported by teachers and history-social science curriculum specialists throughout the state" (Kneedler, 1985, p. v).

In keeping with the high priority, 40% of the Grade 8 assessment that began in 1985 is devoted to the measurement of critical thinking skills in history-social science. Three general measurement approaches (multiple-choice questions, knowledge of vocabulary that is associated with critical thinking, student writing) are used in the assessment. Together, these three measurement approaches are intended to assess 12 critical thinking skills that are classified in three broad areas: (a) problem definition and clarification (e.g., determination of central issues and formulation of appropriate questions); (b) evaluation of information related to the problem (e.g., distinguishing fact, opinion, and reasoned judgment; recognizing stereotypes; and identifying unstated assumptions); and (c) problem solution and the formation of conclusions (e.g., determining the adequacy of data and predicting probable consequences) (Kneedler, 1985).

A description of each of the critical thinking skills along with illustrative items is provided by Kneedler (1985). Although not unlike items that can be found on some of the better standardized tests, the illustrative multiple-choice items demonstrate that this format can be used to do more than measure recall of historical or social science facts. Consider, for example, the following illustrative item that is intended to assess a student's ability to check the consistency of information in a political argument:

> Women should have the same job opportunities as men. Whether it be cook, fire fighter, executive, marine, or astronaut, no position should be denied a

woman, except, of course, those positions requiring considerable physical strength.

Is the speaker being consistent?

A. No, because he fails to mention the Constitution, which guarantees equal rights to all men and women.
B. Yes, because the speaker does not stray from the topic of equal rights.
C. Yes, because the E.R.A. guarantees equal rights for all men and women.
D. No, because he speaks of equal opportunity at the beginning and excludes women on the basis of physical strength at the end. (p. 17)

The limitations of multiple-choice questions for assessing student ability to think critically and solve complex problems is recognized by the California program. Certain skills, such as the ability to "identify reasonable alternatives" are not addressed by the multiple-choice portion of the assessment. Written responses are needed not only to assess these skills, but to probe the depth of student understanding of arguments and their ability to think critically about more complex messages than are used for stems of multiple-choice items.

Constructed Response

Although research has generally shown that item format makes little, if any, difference when "existing multiple-choice tests [are compared] with their free-answer counterparts," the converse is not true. That is, "when we begin with existing free-response tests designed to measure more complex cognitive problem-solving skills, different results are found" (Frederiksen, 1984, p. 199). The formulating hypotheses test developed by Frederiksen and Ward (1978; Ward, Frederiksen, & Carlson, 1980) demonstrates that a paper-and-pencil test with open-ended responses can be used effectively to assess certain abilities that are crucial to scientific reasoning and which are not readily tapped in a multiple-choice testing format.

Ennis (1987) has provided some compelling examples of the difficulty of constructing fixed-response tests of critical thinking. Using illustrative items from the *Watson-Glaser Critical Thinking Appraisal* (Watson & Glaser, 1980) that are intended to measure a test takers ability to identify assumptions needed to support a statement, Ennis (1987) provides clear explanations of how test takers "can get an item wrong when thinking critically" (p. 416). As Ennis indicates, a major difficulty in defending the keyed answer and assuring that the person who is thinking critically should select the keyed answer is that "different test takers bring different background beliefs to bear on an item" (p. 416). Thus, whether a response should be considered right or wrong may depend on the reasons for the choice.

Consider, for example, the following illustrative item from Kneedler's (1985) description of the assessment of the critical thinking skill "predict probable consequences" as part of the California Grade 8 assessment of history-social science:

When strawberries first appear on the market, their price is quite high. At the height of the season, their price goes down. As the season nears its end, their price goes up again.

The most probable reason for these changes in the price of strawberries is that

A. strawberries are more plentiful at the middle of the season.
B. strawberries are lower in quality at the middle of the season.
C. the first and last crops of the season cost more to grow.
D. strawberry pickers are paid more at the beginning and end of the season. (p. 26)

Although "A", the keyed answer, is what would be expected from anyone who used a supply and demand notion to support their reasoning, it is not inconceivable that some of the 29% of the 15,000 students in the field test sample who chose "C" or some of the 29% who chose "B" or "D" brought different beliefs or assumptions to bear on the item and got the wrong answer while thinking critically.

Ennis (1987) suggests several possible ways of dealing with this problem, including the use of interviews, the use of essay tests of critical thinking (e.g., *The Ennis-Weir Critical Thinking Essay Test,* Ennis & Weir, 1985), and the use of open-ended questions that ask why a given multiple-choice option was selected. He concludes that "open-endedness in critical thinking testing seems to be a good way to deal with the problems, but open-endedness in testing is expensive" (p. 419).

Essay Tests of Subject-Matter Understanding

Research that is currently being conducted at the UCLA Center for Research on Evaluation, Standards, and Student Testing (CRESST) as part of a project co-directed by Eva Baker and Joan Herman is investigating the use of essays and other constructed response formats as ways of assessing the depth of a student's understanding in the area of social studies. A series of interviews with political scientists, historians, social studies educators, and high school social studies teachers were conducted as part of the overall project to explore ways of identifying key concepts that could be used to define the content domain for social studies tests (House, 1988). Although the interviews did not suggest that there was a consensus regarding the content that was most critical, there was agreement on two issues. First, there was considerable agreement that "one should not teach facts alone or primarily" (House, 1988, p. 78). Rather, the focus should be on concepts and the development of thinking critically about the content. Second, there was strong agreement that essay tests rather than multiple-choice tests are needed to assess depth of student understanding.

CRESST's studies (Baker & Herman, 1988) of measuring deep understanding of history content are in keeping with the opinions expressed in the interviews conducted by House. The research is exploring ways of using essay tests of history content that "move beyond molecular multiple choice assessments or global content quality judgments of students' writing" (Baker & Herman, 1988, p. 1). The approach that is being explored uses primary source materials rather than textbook or specially written materials as the test stimuli. For example, in one of the initial experimental versions of the measure, students were given speeches from the Lincoln-Douglas debates and were asked to analyze how Lincoln and Douglas interpreted Lincoln's statement, "A house divided against itself cannot stand...It will become all one thing or all the other." The students were instructed that their analyses should: (a) clearly identify the issue(s) or problems that threatened to 'divide the house,' (b) summarize how each politician believed that the issue(s) would best be resolved, (c) explain the most convincing evidence each man used to support his position, and (d) indicate the greatest strength(s) and the greatest weakness(es) of each man's argument.

A central issue in this research program is the development of scoring procedures that will yield reliable scores that reflect the depth of a student's understanding rather than a global measure of their writing. The approach being investigated follows the lead of work in cognitive psychology that compares expert and novice performance (e.g., Voss, Green, Post, & Penner, 1983; Voss, 1986). According to Glaser (1987), past work comparing expert and novice performance suggests that the features of the essays that are apt to distinguish superior performance from average performance include coherence, the degree to which information is interrelated, the use of underlying principles rather than only the surface features of the speeches, and the use of background knowledge about that period of history and about the two men.

Although still in its early stages, the work on assessment in history builds on a substantial experience with research on writing assessment (e.g., Baker, 1987). The demands of the task are more in line with the judgments of what is required to evaluate the important goals of social studies that were expressed by the subject matter experts and educators in the interviews conducted by House (1988).

Diagnostic Testing

Regardless of its format, a test should contribute to student learning. The nature of the contribution could take a variety of forms, but one that is often proposed is to provide teachers with diagnostic information that will lead to improved instructional decisions about individual students.

There are a number of standardized achievement tests that have been labeled "diagnostic." However, there is relatively little research evidence to support the diagnostic interpretation of such tests (Bejar, 1984; Linn, 1986). That is, there is little evidence that students with a particular pattern of scores on a set of skills measured by a given "diagnostic" test battery would benefit most from a particular series of instructional experiences. Educational tests currently do a much better job of prediction of future achievement than of diagnosis and prescription. According to Glaser (1986), current "tests (with the exception of the important informal assessment of the good classroom teacher) typically are not designed to guide the specifics of instruction" (p. 45).

There are a variety of reasons that diagnostic achievement testing has not been more successful. One reason is that diagnosis often requires quite detailed information about what a student can and cannot do and, more importantly, about the types of misunderstandings and the nature of his or her mental model. In a one-on-one teaching situation a skilled tutor may be able to probe the areas of strength and weakness and the ways in which the student conceptualizes and attacks problems, and come to an understanding of the student's misconceptions and the types of errors that the student is likely to make. From this information the tutor may also be able to prescribe a set of instructional activities that will be especially effective for that student at that particular stage of development. When faced with a classroom of, say, 30 students, however, a teacher does not have the time to obtain such detailed information about each individual student. The tasks of collecting, analyzing, and prescribing a set of instructional activities suited to each individual student's needs are just too labor intensive.

Computers have the potential for providing the support needed for keeping track of the types of errors that each student makes and searching for consistencies that suggest particular misunderstandings. Researchers such as Brown and Burton (1978) and Tatsuoka (1983) have demonstrated that the errors that students make in solving problems are often systematic. The identification of systematic errors and the use of that information to draw conclusions about the nature of a student's misconception, however, requires detailed analyses of patterns of response to numerous problems.

The research of Brown and Burton (1978) and Tatsuoka (1983) has demonstrated that computerized tests can be designed that do the labor intensive task of keeping track of student responses. Computers can also test hypotheses that suggest that errors are the result of a particular misconception or the consistent application of a faulty algorithm. Furthermore, computerized tests can obtain the needed information more efficiently than can paper-and-pencil tests by tailoring the choice of problems that are presented to the previous performance of the student and the current hypotheses about the student's current source of difficulty. Once a systematic error or misconception is identified, the computer can also suggest instructional activities that are designed to remedy the specific difficulty.

Although computers can reduce labor by keeping track of detailed information about student responses to problems, the development of an effective diagnostic testing technology will require much more than a delivery system. As Ward (1984) has noted, diagnostic testing needs to be based on a theory of knowledge and a theory of instruction as well. Diagnostic testing is also likely to require a different type of psychometric theory. "Neither classical test theory nor item response theory is well suited for diagnostic testing problems. Both approaches rely on an assumption of unidimensionality and treat deviations as noise. It is precisely those deviations from a single dominant dimension, however, that are of central concern in diagnostic testing" (Linn, 1986, p. 159).

These are challenging demands. Considerable progress toward these ends has been made in the past decade, however. Research in cognitive psychology and artificial intelligence has made great progress toward providing a theoretical foundation for diagnostic testing. This work suggests, as was indicated above, that the mental models or theories that students have and the ways in which they represent problems are central to student learning. As Glaser (1986) has noted, with the right instruction "students test, evaluate, and modify their current theories on the basis of new information, and as a result, develop new schema that facilitate more advanced thinking" (p. 55).

A challenge for effective instruction, according to this conceptualization, is to decide on the information that will be most useful at a given stage of learning in helping a student "test, evaluate, and modify" his or her current mental model. This suggests that effective diagnostic testing needs not only to identify student errors, but to assess their mental models as well.

Computer Simulations

Such assessments are incorporated as a necessary part of intelligent tutor systems. Simulations such as the patient management problems that have been used in training and certification of physicians for a number of years incorporate many of these features. Patient management problems simulate the interaction between a physician and a patient. The test taker initially is presented with a limited set of information about a patient, such as a verbal description of symptoms of the type that a patient might provide at the start of a visit. The test taker then has a variety of options, such as getting a patient history, ordering laboratory tests, or deciding on a course of treatment. Requested information is provided and new options can be followed by the test taker until a diagnosis is made and a course of treatment is prescribed.

As Ward (1984) has indicated, patient management problems "are attractive because they offer greater realism than do standard 'one-shot' examination questions, and because they provide samples of performance that can be scored in example, how efficiently did the examinee attack the problem, how many serious errors were made, what was the cost to the patient in dollars and in pain and suffering, and so on" (p. 19). A computer can contribute to the realism of the simulations and can facilitate the scoring along different dimensions such as the ones suggested by Ward.

Computer-administered problem simulations have a number of potential advantages over current paper-and-pencil tests in all areas of the curriculum. They can provide a means of going beyond the sort of factual recall questions that too often dominate paper-and-pencil tests. Instead, they focus attention on a student's ability to use information to solve problems. Unlike multiple-choice questions that only reveal the product of a student's thinking, computerized simulations can assess the process that a student uses to solve a problem, including the way in which the problem is attacked, the efficiency of the solution, and the number of hints that may be needed to solve the problem. The process information may also be used to assess a student's mental representation of a problem, which, as was indicated above, may be more important to identify than whether a correct answer is produced.

Simulations are an important part of a number of computer-based instructional programs, especially ones that are called "intelligent tutors." As Collins (in press) has pointed out, such systems have great promise as intelligent testers. He identified three important benefits of developing intelligent tutors as intelligent testers: "(a) testing would be focused on problem solving and planning skills, (b) [students'] ability to learn in a domain as well as their prior knowledge could be tested, and (c) the tests could be adaptive to the student's prior knowledge, and would test their generative abilities instead of their recognition abilities" (p. 11 of typescript).

Conclusion

Constructing valid assessment procedures to tap thinking processes is certainly not an easy task, but the difficulty of the task is not the major barrier. Practical concerns about cost and efficiency, the seemingly insatiable demand to boil everything down to a single number, and the overreliance on standard psychometric criteria to judge test reliability and validity present much more formidable barriers. As in the instance of writing assessment, the case will have to be made that the form of the assessment represents an important part of the specification of educational goals, which in turn influence what is taught and what is learned. The case must also be made that thinking skills and processes are essential educational goals that go well beyond the accumulation of factual knowledge. Furthermore, the reliance solely on multiple-choice test items distorts the goals and frustrates their achievement. Hence, it is worth the added expense and complexity that such assessment will require.

References

Alexander, L., & James, H.T. (1987). *The nation's report card: Improving the assessment of student achievement.* Cambridge, MA: National Academy of Education.

Applebee, A.N., Langer, J.A., & Mullis, I.V.S. (1986). *Writing: Trends across the decade, 1974-84.* Princeton, NJ: Educational Testing Service.

Baker, E.L. (1987, September). *Time to write: Report of the US-IEA study of written composition.* Paper presented at The IEA General Assembly, New York.

Baker, E.L., & Herman, J.L. (1988). *Content assessment: Assessing deep understanding of social studies* (Proposal to the Office of Educational Research and Improvement). Los Angeles: UCLA, Center for the Study of Evaluation.

Bejar, I.I. (1984). Educational diagnostic assessment. *Journal of Educational Measurement, 21,* 175-189.

Bloom, B.S. (1956). *Taxonomy of educational objectives.* Ann Arbor, MI: Edwards Brothers.

Bowman, C.M., & Peng, S.S. (1972). *A preliminary investigation of recent advanced psychology tests: An application of a cognitive classification system.* Princeton, NJ: Educational Testing Service.

Bransford, J.D., & Johnson, M.K. (1973). Consideration of some problems in comprehension. In W. G. Chase (Ed.), *Visual information processing.* New York: Academic Press.

Brown, J.A., & Burton, R.R. (1978). Diagnostic models for procedural bugs in basic mathematical skills. *Cognitive Science, 2,* 155-192.

Brown, R. (1987). Who is accountable for 'thoughtfulness'? *Phi Delta Kappan, 69*(1), 49-52.

Chall, J.S. (1977). *An analysis of textbooks in relation to declining SAT scores.* New York: College Board.

College Board. (1986). *DRP handbook.* New York: College Entrance Examination Board.

Collins, A. (in press). Reformulating testing to measure learning and thinking. In N. Frederiksen, R. Glaser, A. Lesgold, & M. Shafto (Eds.), *Diagnostic monitoring of skills and knowledge acquisition.* Hillsdale, NJ: Erlbaum.

Commission on Reading, National Academy of Education. (1985). *Becoming a nation of readers: The report of the commission on reading.* Washington, DC: National Institute of Education.

Cross, E.R., & Paris, S.G. (1987). Assessment of reading comprehension: Matching test purposes and test properties. *Educational Psychologist, 22,* 313-332.

Curtis, M., & Glaser, R. (1983). Reading theory and assessment of reading achievement. *Journal of Educational Measurement, 20,* 133-147.

Debra P. v. Turlington. 644 F. 2d 397, 404 (5th Cir. 1981).

Ebel, R.L., & Frisbie, D.A. (1986). *Essentials of educational measurement* (4th ed.). Englewood Cliffs, NJ: Prentice-Hall.

Ennis, R.H. (1987). Testing teachers' competence, including their critical thinking. *Proceedings of the Forty-third Annual Meeting of the Philosophy of Education Society* (pp. 413-420). Cambridge, MA: Philosophy of Education Society.

Ennis, R.H., & Weir, E. (1985). *The Ennis-Weir critical thinking essay test.* Pacific Grove: Midwest Publications.

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

Frank, H.J. (1978). An examination of the levels of questions on standardized tests of elementary science. *Science and Children, 14,* 30-32.

Frederiksen, N. (1984). The real test bias: Influences of testing on teaching and learning. *American Psychologist, 39,* 193-202.

Frederiksen, N., & Ward, W.C. (1978). Measures for the study of creativity in scientific problem solving. *Applied Psychological Measurement, 2,* 1-24.

Glaser, R. (1986). The integration of instruction and testing. *The redesign of testing for the 21st century: Proceedings of the 1985 ETS Invitational Conference* (pp. 45-58). Princeton, NJ: Educational Testing Service.

Glaser, R. (1987, December). *Expertise and assessment.* Paper presented at the UCLA/CRESST Approaches to Subject Matter Assessment Conference, Los Angeles.

Gullickson, A.R., & Hopkins, K.D. (1987). Perspectives on educational measurement instruction for preservice teachers. *Educational Measurement: Issues and Practice, 6*(3), 12-16.

House, E.R. (1988). *Definition of content in social studies testing.* (Report to OERI, Grant No. G-86-0003). Los Angeles: UCLA Center for Evaluation, Standards, and Student Testing.

Johnston, P.H. (1983). *Reading comprehension*: A cognitive basis. Newark, DE: International Reading Association.

Johnston, P.H. (1984). Prior knowledge and reading comprehension test bias. *Reading Research Quarterly, 19,* 219-239.

Kerins, T. (1988, April). *Changing the assessment program to respond to research on reading: The Illinois experience.* Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans.

Kneedler, P.E. (1985). *Assessment of critical thinking skills in history-social science.* Sacramento, CA: California State Department of Education.

Koslin, B.L., Koslin, S., & Zeno, S. (1979). Towards an effectiveness measure in reading. In R.W. Tyler & S.H. White (Eds.), *Testing, teaching, and learning: Report of a conference on research on testing.* Washington, DC: National Institute of Education.

Koslin, B.L., Zeno, S., & Koslin, S. (1987). *The DRP: An effectiveness measure of reading.* New York: College Entrance Examination Board.

Levine, A.G., McGuire, C.H., & Nattress, L.W. (1970). The validity of multiple-choice achievement tests as measures of competence in medicine. *American Educational Research Journal, 7,* 69-82.

Linn, R.L. (1986). Barriers to new test design. *The redesign of testing for the 21st century: Proceedings of the 1985 ETS Invitational Conference* (pp. 69-79). Princeton, NJ: Educational Testing Service

Linn, R.L. (1987). Accountability: The comparison of educational systems and the quality of test results. *Educational Policy, 1,* 181-198.

Linn, R.L., & Palmer, C.N. (1985). Standards and expectations: The role of testing. *Excellence in our schools: Making it happen.* New York: College Board.

Linn, R.L., & Valencia, S.W. (1986). *Reading Assessment: Practice and theoretical perspectives* (Report to OERI, Grant No. G-86-0003). Los Angeles: UCLA Center for the Study of Evaluation.

Madaus, G.F. (1985). Public policy and the testing profession: You've never had it so good. *Educational Measurement: Issues and Practices, 4*(4), 5-11.

Mayer, R.E. (1985). *Thinking, problem solving, and cognition.* San Francisco: Freeman.

Morgenstern, C.F., & Renner, J.W. (1984). Measuring thinking with standardized tests. *Journal of Research in Science Teaching, 21,* 639-648.

National Academy of Education. (1987). *The Nation's Report Card: Improving the assessment of student achievements.* Cambridge, MA: National Academy of Education.

National Assessment of Educational Progress. (1985). *The reading report card: Progress toward excellence in our schools.* Princeton NJ: Educational Testing Service.

National Commission on Excellence in Education. (1983). *A nation at risk: The imperative for educational reform.* Washington, DC: U.S. Printing Office.

Pearson, P.D., & Spiro, R.J. (1980). Toward a theory of reading comprehension instruction. *Topics in Language Disorders, 1,* 71-88.

Pipho, C. (1985, May 22). Tracking the reforms, Part 5: Testing—Can it measure the success of the reform movement? *Education Week,* p. 19.

Quellmalz, E. (1985). Needed: Better methods for testing higher-order thinking skills. *Educational Leadership, 43*(2).

Riley, M.S., Greeno, J.G., & Heller, J.I. (1983). Development of children's problem-solving ability in arithmetic. In H.P. Ginsburg (Ed.), *The development of mathematical thinking.* New York: Academic Press.

Simon, H.A. (1978). Information-processing theory of human problem solving. In W.K. Estes (Ed.), *Handbook of learning and cognitive process: Vol. 5, Human information processing* (pp. 271-295). Hillsdale, NJ: Erlbaum.

Snow, R.E., & Lohman, D.F. (in press). Implications of cognitive psychology for educational measurement. In R.L. Linn (Ed.), *Educational Measurement* (3rd ed.). New York: Macmillan.

Stiggins, R.J. (1987). Design and development of performance assessment. *Educational Measurement: Issues and Practice, 6*(3), 33-42.

Stiggins, R.J. (1988). Revitalizing classroom assessment: The highest instructional priority. *Phi Delta Kappan, 69*(5), 363-368.

Stiggins, R.J. (no date). *High impact teacher training in classroom assessment.* Portland, OR: Northwest Regional Educational Laboratory.

Stiggins, R.J., Griswald, M., & Green, K.R. (1988, April). *Measuring thinking skills through classroom assessment.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans.

Stiggins, R.J., Rubel, E., & Quellmalz, E. (1986). *Measuring thinking skills in the classroom.* Washington, DC: National Education Association.

Tatsuoka, K.K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20,* 221-20.3.

Valencia, S. (1988, April). *Research for reforming the assessment of reading comprehension.* Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans.

Valencia, S., & Pearson, P.D. (1986). *Reading assessment: Time for change.* Unpublished manuscript, University of Illinois, Center for the Study of Reading, Champaign.

Voss, J.F. (1986). Social studies. In R.F. Dillon & R.J. Sternberg (Eds.), *Cognition and instruction.* Orlando: Academic Press.

Voss, J.F., Greene, T.R., Post, T.A., & Penner, B.C. (1983). Problem solving skills in the social sciences. In G. Bower (Ed.), *The psychology of leaning and motivation: Advances in research and theory* (Vol. 17). New York: Academic Press.

Walker, D.F. (1983). What constitutes curricular validity in a high school leaving examination? In G.F. Madaus (Ed.), *The courts, validity and minimum competency testing* (pp. 171-181). Hingman, MA: Kluwer Nijhoff.

Ward, W.C. (1984). Using microcomputers to administer tests. *Educational Measurement: Issues and Practice, 3,* 16-20.

Ward, W.C., Frederiksen, N., & Carlson, S. (1980). Construct validity of free-response and multiple-choice versions of a test. *Journal of Educational Measurement, 17,* 11-29.

Watson, G., & Glaser, W.M. (1980). *The Watson-Glaser Critical Thinking Appraisal.* San Antonio, TX: The Psychological Corporation.

Wixson, K.K., & Peters, C.W. (1984). Reading redefined: A Michigan Reading Association position paper. *Michigan Reading Journal, 17,* 4-7.

Wixson, K.K., & Peters, C.W. (1987). Comprehension assessment: Implementing an integrative view of reading. *Educational Psychologist, 22,* 333-356.

Wixson, K.K., Peters, C.W., Weber, E.M., & Roeber, E.D. (1987). New directions in statewide reading assessment. *The Reading Teacher, 40,* 749-754.