Chapter 9

Addressing the "Two Disciplines" Problem: Linking Theories of Cognition and Learning With Assessment and Instructional Practice

JAMES W. PELLEGRINO
Vanderbilt University

GAIL P. BAXTER
Educational Testing Service

ROBERT GLASER
University of Pittsburgh

On September 2, 1957, Lee Cronbach delivered his visionary presidential address to the American Psychological Association (APA), calling for the unification of differential and experimental psychology, the two disciplines of scientific psychology. He described the essential features of each approach to asking questions about human nature, and he strongly hinted at the benefits to be gained by unification. Cronbach was calling for linking theories and research on learning and instruction, especially the instructional treatments that logically and psychologically followed from such research, with the tradition of assessing individual differences in cognitive abilities. In his opinion, such work would probably yield information of profound educational relevance. In describing some illustrative examples, he stated, quite boldly, "Such findings ... when replicated and explained, will carry us into an educational psychology which measures readiness for different types of teaching and which invents teaching methods to fit different types of readiness" (p. 681). He subsequently went even further and argued that this work had broader theoretical impact and meaning. "Constructs originating in differential psychology are now being tied to experimental variables. As a result, the whole theoretical picture in such an area as human abilities is changing" (p. 682).

Preparation of this chapter was supported by the Learning Technology Center at Peabody College, Vanderbilt University; the K-12 Learning Consortium Project; the Learning Research and Development Center at the University of Pittsburgh; the National Center for Research on Evaluation, Standards, and Student Testing (CRESST); and the Educational Testing Service. The views expressed are solely those of the authors. We would like to express our appreciation to Dave Lohman, Lorrie Shepard, and Ali Iran-Nejad for their constructive comments on drafts of this chapter. We would also like to thank Dave Law for his assistance in assembling and reviewing some of the material discussed in this chapter, as well as his reactions to earlier versions. We dedicate this chapter to our departed colleague and friend Richard Snow, whose career was dedicated to seeking out answers to difficult questions about the nature of human abilities, constantly probing how such knowledge could be profitably used to enhance education. Many of our thoughts and insights draw heavily from his seminal works addressing the two disciplines issues highlighted in this chapter.

No doubt, Cronbach was overly optimistic about what would and could be accomplished at the practical instructional level and at the theoretical level. Much, however, has changed in the ensuing 40-plus years. Developments in cognitive psychology, together with changes in the goals and standards for instructional practice and heightened demands for educational assessments to both reform and inform education, collectively establish a new context for considering the confluence of cognitive science and psychometrics. Advances in cognitive science, learning, and human development provide new perspectives for the design, administration, and use of assessments of academic achievement. It is our contention that these perspectives are critical if contemporary education is to achieve the goal of helping all students succeed academically. To accomplish such a goal, assessment practices must be based on contemporary understandings and appropriate standards regarding the acquisition of proficiency and expertise in specific academic content domains. Assessment could then have a significant positive influence on what is learned in classrooms and on how knowledge and competence are demonstrated in various contexts of educational consequence, ranging from the classroom to state and national assessments of educational attainment.

In this chapter, we probe several issues of historical and conceptual significance that provide a foundation for negotiating a necessary and fruitful coordination of cognitive psychology and psychometrics, as originally envisioned by Cronbach (1957) and as further articulated over time by others (e.g., Anastasi, 1967; Glaser, 1981). The 40-year odyssey that we recount includes significant theoretical developments in cognitive psychology and changes in assessment practice that are coincident with the changing sociopolitical context of education. Explicit in our discussion are the consequent challenges to the research community to harness theoretical developments in ways that are responsive to demands for educational equity and accountability. A contemporary cognitive perspective on assessment can inform the larger educational debate about what types of learning matter most, and it can lead to improved educational practice at the place where learning and instruction occur. In presenting our case, we call attention to progress that has been made together with unresolved conceptual issues and divergent theoretical perspectives that affect any prospects for a truly successful integration of cognition, instruction, and assessment.

We begin with a historical account of how empirical and theoretical efforts in cognitive psychology were linked to and supported the study of individual differences in performance and learning. Our discussion starts with the 1960s and aptitude-treatment interaction (ATI) research of the type originally mentioned by Cronbach, and it continues through the 1970s and 1980s with a brief review of the substantial body of work on cognitive process analyses of aptitude and achievement. Despite intense activity in these lines of research and theory, neither one obtained the hoped-for objective of creating a measurement milieu that substantially contributed to the enabling conditions of a psychology of instruction. Subsequent efforts during the 1980s and 1990s to study competent performance and expertise were more relevant to this purpose. From such work a theory of cognition has developed that includes models of learning and performance in knowledge-rich domains, the type of theory needed to accomplish the goals of affecting instructional practice and improving educational attainment, including the ''what'' and ''how'' of educational assessment.

As we entered the 1990s, national standards in multiple curriculum areas, together with the formulation and adoption of national education goals, were a dominant force in shaping American educational policy, including assessment practices. We consider efforts to respond to these sociopolitical forces through changes in achievement testing that took into account developments in both cognitive and curricular theory. Multiple-choice tests were put under scrutiny, and their shortcomings forced the development of alternative forms of assessment. Performance-based assessments and multilevel scoring procedures were offered alongside traditional multiple-choice tests as part of a response to the expanded role of assessment in addressing issues of educational policy and practice. As progress was made in the development of alternative methods of assessment for various summative and formative purposes, the conceptual discrepancies between contemporary cognitive theory and extant psychometric models and methods became especially apparent. Resolution, of necessity, requires reconceptualizing psychometric theory to harness the theoretical developments in cognition in ways that are responsive to educational needs at multiple levels of practice and policy.

The implications of all that has gone before and the prospects for the future with regard to fully developing a cognitive approach to assessment are the subject of the final two sections of this chapter. We consider contemporary attempts to bridge the conceptual divide and tightly couple cognitive theories with psychometric models, including separate developments in the areas of aptitude and achievement assessment. It appears that more progress has been made in linking cognitive constructs with the design and assessment of cognitive aptitudes and that much remains to be done in the assessment of domains of academic achievement. We also consider the possibilities of bringing together cognitive theories of aptitude and achievement in the context of interactions with instructional treatments. This brings us full circle relative to Cronbach's original concerns. In the final section of the chapter, we reflect on the current state of knowledge with regard to unifying cognition and assessment and consider what still needs to be done for a rapprochement of the ''two disciplines'' that effectively serves the educational needs of an increasingly diverse population of students.

DEVELOPMENTS LINKING INDIVIDUAL DIFFERENCES ASSESSMENT WITH COGNITIVE THEORY AND RESEARCH

Despite the elegance of his logic, Cronbach's arguments and exhortations for unification of the two disciplines largely fell on deaf ears. Some might argue that this was because he was wrong about the utility of such an enterprise. We would argue instead that his fundamental premises were not flawed and that the rapprochement he sought remains important and desirable. Unfortunately, Cronbach's problem was one of feasibility. What he could not foresee was that the necessary empirical and theoretical foundation on which to develop such a unified approach to understanding and assessing cognition was unavailable, nor could he have known that this state of affairs would remain so for quite some time.

In the intervening years, there has been a remarkable transformation in the study of learning and cognition (see, e.g., Bransford, Brown, & Cocking, 1999). As a result, the nature of the interaction between research and theory on cognition and instruction and research and theory on individual differences has changed significantly over the decades since Cronbach's address. The resultant state of knowledge and understanding renders it conceivable that some of what Cronbach wished for in 1957 might be attainable in the foreseeable future. To see how far we have come, as well as to appreciate what we have yet to do, we provide a recapitulation and analysis of some of the historical developments connecting research on learning and cognition with research on individual differences. We do so by selectively focusing on research in two broad areas: aptitude-treatment interactions and cognitive analyses of aptitude and expertise.

The Pursuit of Aptitude-Treatment Interactions

Cronbach's call to integrate differential and experimental psychology described the conflict between scientific traditions that originated in different conceptions of the natural world. Differential psychology, a "conservative" tradition, originated in the works of Darwin, Spencer, and Galton, whereas experimental psychology, a "liberal" tradition, originated in the works of Ward, James, and Dewey. Differential psychology attempted to identify the individual who would perform best in the environment, whereas experimental psychology tried to identify the environment that would work best for all individuals. In considering the future of these somewhat parallel lines of work for improving instruction, Cronbach (1957) concluded: "The greatest social benefit will come from applied psychology if we can find for each individual the treatment to which he can most easily adapt. This calls for the joint application of experimental and correlational methods" (p. 679).

Cronbach asked the two schools to pool their efforts in the combined evaluation of aptitudes and learning treatments with the expectation that some aptitudes would show strong interactions with educational treatments. For example, it was thought that using diagrams and figural materials would promote learning among individuals with high spatial ability. The effects of general aptitudes such as Spearman's *g*, however, were believed to effectively span all treatments and thus were expected to be less involved in ATIs. It was hoped that the combined study of aptitudes and educational treatments would form the basis of a new science benefiting both individuals and society as a whole (e.g., Cronbach & Gleser, 1957). By capitalizing on knowledge about specific ATIs, it would be possible for students lower in general aptitude to reach higher levels of achievement than were customary in the one-size-fits-all world of traditional education.

Almost 20 years later, Glaser framed the issues of assessment and learning in terms of a new ''linking science.'' This psychology of instruction, argued Glaser (1976), would function prescriptively, turning knowledge gleaned from learning laboratories into instructional designs. As part of the prescription, individual learners would be matched by their initial states to optimal instructional treatments. Thus, the overall goal of the new science, and especially the combination of two of its components—describing the initial state of the learner and creating conditions that foster competence—was essentially an ATI design. To make this all possible would require the adoption of a new attitude toward assessment, namely, that it could be instrumental in providing information to improve instruction. This in turn would be facilitated by the development of new assessments of the specific cognitive processes and structures involved in academic learning and performance.

By 1975, two major research endeavors had begun that, at least in part, were based on the expectation that the study of ATIs would be useful in improving educational achievement. Nevertheless, in 1975, when Cronbach (p. 116) once again addressed the APA, he said, ''the line of investigation that I advocated in 1957 no longer seems sufficient . . . complexity forces us to ask once again, Should social science aspire to reduce behavior to laws?'' Essentially, evaluating ATIs turned out to be more complex than originally expected, as the effects of educational treatments and their interactions with student aptitudes were found to be highly contextualized. This created higher order interactions that were less generalizable than the first-order interactions envisioned in the original ATI proposal. Also, the confounding variables in the higher order interactions were often not part of the experimental design and were detected only post hoc in the context of a number of experiments providing similarly enigmatic results. Instructional implementation, teachers' perspectives and attitudes, and students' attitudes, maturation, and socioeconomic status were just a few among many variables that complicated the ATI picture. This is particularly problematic because higher order interactions, aside from being more difficult to interpret, are more difficult to show statistically as a result of the decrease in statistical power as the order of an interaction increases. Thus, statistically nonsignificant interactions became uninterpreted interactions, which in turn supported contradictory findings and precluded any general statement that could serve as a guide in instructional decisions.

Cronbach (1975) suggested that dealing with the problems of ATI research would require a more sophisticated approach that dealt with the combined challenges of complexity, statistical significance testing, and the ever-present factor of change. In particular, he was concerned with the practices of accepting the null hypothesis and reporting *F* statistics and associated significance levels without concern for practical significance. He suggested deemphasizing statistical significance testing in favor of using components of variance, effect sizes, and confidence intervals to evaluate the practical significance of variables and their interactions. Regarding complexity and change, Cronbach stressed that empirical relationships change with changes in time and place and the overall context in which research is conducted. Accordingly, researchers should attempt to observe and report the boundary conditions that limit generalizations of their findings. Apparently, this is what Cronbach meant when he questioned the utility of seeking psychological laws, which presumably would cover all persons at all times.

In 1977, Cronbach and Snow presented a book-length review of more than a decade's worth of ATI research. Among the conclusions they arrived at were the following: (a) ATIs do exist; (b) as yet, no ATI hypothesis had been sufficiently confirmed to serve as a basis for instructional practice; and (c) contrary to earlier predictions, general abilities enter into interactions more often than specific abilities. Snow (1989a, p. 21) noted it as an interesting aspect of the sociology of science that many cited the review as involving a negative conclusion, when "a thorough examination of it should lead to quite the opposite conclusion." Although they realized that there might be decades of research preceding an ultimate resolution, Cronbach and Snow remained generally optimistic regarding the potential utility of ATIs. Others, however, saw things somewhat differently:

> Even if one includes a general mental ability, which appears to interact relatively more often with certain instructional treatments than more specialized abilities, the overall conclusion regarding mental aptitude measures is that "no interactions are so well confirmed that they can be used as guides to instruction." In those occasional instances when positive results were obtained, no general principles emerged because the findings were rarely sustained when new subject matter tasks were used or when similar studies on the same tasks were undertaken. (Bond & Glaser, 1979, p. 139)

As implied in the preceding comments, the most consistent findings in ATI research are that general aptitudes interact with instruction (Cronbach & Snow, 1977; Snow, 1989a, 1994). This was surprising, because general aptitudes were not originally expected to be involved in ATIs but were expected to affect all treatments equally. Furthermore, general aptitudes were thought of as supporting a single rank ordering of individuals, and the ultimate goal of ATI research is to disrupt such rankings. However, in hindsight it becomes apparent that if one wants to disrupt a rank ordering, look first at what supports the ranking. In most cases, aptitude-treatment regressions involving general aptitudes are steeper in low-structure treatments, showing rapidly increasing benefit as aptitude increases, and shallower in high-structure treatments, providing potentially better outcomes for low-aptitude students. For example, elaborated materials tend to benefit low-aptitude students, whereas treatments that put the burden of organization on the learner tend to benefit high-aptitude students. Similarly, discovery learning tends to benefit high-aptitude students, whereas direct instruction tends to benefit low-aptitude students. Reversing the optimal treatment level for either group of students can also have detrimental effects. According to Snow (1989a, 1994), typical teaching is generally between the extremes of high and low structure but is probably closer to low structure than high. A number of personality traits or conative aptitudes also appear regularly in ATIs (see Snow, 1989b).

The failure of ATI research to prescribe assistance to instruction as was originally intended suggested, among other things, that standardized tests may be inappropriate measures for this purpose. The assumption that the label of a particular aptitude measure had direct implications for instructional practice was generally false (e.g., pairing a spatial aptitude test with treatment procedures that deemphasized verbal content in instruction). The mere absence of words (diagrams, for example) by no means implies the presence of abilities required in such tests. When "off the shelf" aptitude measures were not available, investigators developed their own based on an analysis of what was sensible to measure, and, not surprisingly, the results were generally more informative. Although established tests of aptitude or general ability were reliable, they could not compete with information about learning processes afforded by tests specially constructed for experimental work (Bond & Glaser, 1979; Pellegrino & Glaser, 1979).

In short, there was a mismatch between the aptitude measures derived from a psychometric selection-oriented tradition and the processes of learning and performance under investigation in experimental and developmental psychology. The use of traditional psychometric instruments for fruitful ATI research requires a careful analysis of processes that relate aptitude, treatment, and the knowledge or skills being learned. Testable theories are required that describe competencies measured in the pretest, competencies required for task performance, and treatment procedures that connect the two (Snow, 1980). However, at the time of the initial ATI work, sufficient theories were not available to support those efforts.

For various reasons, including the developments considered next, enthusiasm for ATIs had waned considerably by the beginning of the 1980s, and this line of inquiry never emerged as a significant theoretical and empirical enterprise. Within the last decade, however, a small number of researchers have conducted what might be termed "second-generation" ATI research. This line of work carefully addresses some of the theoretical issues that plagued earlier work and has yielded some potentially interesting outcomes (e.g., Shute, 1993; Swanson, 1990). Further discussion of such work and its implications is presented in a later section.

Cognitive Analyses of Aptitude and Expertise

By the end of the 1960s and at the beginning of the 1970s, an emerging psychology of human cognition was focused on explicating the mental structures and processes underlying various simple and complex performances. The terminology of information processing, accompanied by the methodology of reaction time and protocol analyses, came to dominate the empirical and theoretical landscape. During the 1970s, it occurred to some resolute educational, cognitive, and developmental psychologists that it might be possible to accomplish part of what Cronbach originally suggested, that is, link individual differences in aptitude to constructs emanating from laboratory research on cognition. Because intelligence and aptitude tests are essentially measures of scholastic ability, it was also thought that one way to increase our understanding of academic achievement, and perhaps even improve it, would be to develop process-based performance models for intelligence and aptitude test items. From such knowledge, basic cognitive abilities might be identified that could then be influenced by instructional programs. Carroll (1978) described the situation as follows:

> The performances required on many types of mental ability tests—tests of language competence, of ability to manipulate abstract concepts and relationships, of ability to apply knowledge to the solution of problems, and even of the ability to make simple and rapid comparisons of stimuli (as in a test of perceptual speed)—have great and obvious resemblances to performances required in school learning, and indeed in many other fields of human activity. If these performances are seen as based on learned, developed abilities of a rather generalized character, it would frequently be useful to assess the extent to which an individual has acquired these abilities. This could be for the purpose of determining the extent to which these abilities would need to be improved to prepare the individual for further experiences or learning activities, or determining what kinds and amounts of intervention might be required to effect such improvements. These determinations, however, would have to be based on more exact information than we now have concerning the effects of different types of learning experiences . . . on the improvements of these abilities. (pp. 93–94)

Extensive programs of research on the cognitive analysis of aptitude and intelligence began in the 1970s. Researchers pursued two complementary approaches to studying individual differences in cognitive abilities, termed the cognitive correlates approach and the cognitive components approach (Pellegrino & Glaser, 1979). The cognitive correlates approach consisted of assessing relationships between performances on psychometric ability tests with parameters derived from standard laboratory information-processing tasks. Typically, subjects were divided into high- and low-ability groups based on external aptitude test scores (e.g., Scholastic Aptitude Test [SAT] scores), and between-group analyses were conducted in which various parameters for basic information-processing operations, such as stimulus encoding and matching, memory search, and response execution, were used as dependent variables. Thus, the cognitive correlates approach attempted to explain differential performance on standard aptitude measures by using theory-based information-processing constructs. A significant mean difference between high- and low-ability groups suggested that the structure or process represented by the cognitive task parameter was instrumental in ability test performance. Overall, the cognitive correlates approach is best exemplified by the work of Hunt and his colleagues on verbal ability (Hunt, 1978; Hunt, Frost, & Lunnenborg, 1973; Hunt & Lansman, 1975) and is typified in the title of the article "What Does It Mean to Be High Verbal?" (Hunt, Lunnenborg, & Lewis, 1975). This approach was also used successfully to elucidate the varying nature of reading ability differences as represented in the extensive work done by Perfetti and his colleagues (see, e.g., Perfetti & Goldman, 1976; Perfetti & Hoagaboam, 1975; Perfetti & Lesgold, 1982).

The cognitive components (or task analytical) approach attempted to directly understand the components of performance underlying individuals' solution of items used to assess intelligence and aptitude. Working from detailed task analyses, the objective was to develop models of task performance and use these process models for analyzing individual differences. Information-processing parameters were then derived and used as the basis for explaining high and low scores. The cognitive components approach assessed performance strategies, executive routines, and also declarative and procedural knowledge that interacted with processing capabilities varying across individuals (Snow & Lohman, 1989). The componential approach has been highly influential in modeling performance in domains such as spatial transformation (e.g., Kyllonen, Lohman, & Woltz, 1984; Mumaw & Pellegrino, 1984; Pellegrino & Kail, 1982; Shepard & Cooper, 1983) and inductive and deductive reasoning (e.g., Goldman & Pellegrino, 1984; Kotovsky & Simon, 1973; Pellegrino & Glaser, 1982; Sternberg, 1977), domains that are primarily process based rather than knowledge based. The latter distinction became increasingly important as researchers pursued more complex domains of performance and more direct measures of learning within such domains.

Collectively, the cognitive correlates and cognitive components approaches provided substantial information regarding individual differences in performance on cognitive ability tests. Much of the work that has been done is both theoretically and methodologically elegant and, from a scientific standpoint, served to demonstrate how cognitive psychology could be effectively applied to analyzing important domains of human performance. Significant work in this vein continues, most notably in the study of individual differences in attentional processes, dynamic spatial reasoning (e.g., Law, Pellegrino, & Hunt, 1993; Pellegrino & Hunt, 1989), information coordination (Law, Morrin, & Pellegrino, 1995; Morrin, Law, & Pellegrino, 1994), and related instances of basic information processing (e.g., Kyllonen, 1993). Characteristic of this work is a focus on aspects of cognition that are largely structure and process oriented, and individual differences tend to be closely linked to limitations of the information-processing architecture such as working memory capacity and processing speed (e.g., Woltz, 1988; Woltz & Shute, 1993).

Despite the contributions of this line of cognitive research and theory, it has fallen short of its primary objective of developing measurement procedures that could inform teaching and learning. Lohman (1994) has suggested that there are at least two reasons for the apparent failure: (a) The assumption that psychometric tests would make good cognitive tasks, or vice versa, was generally false, and (b) cognitive performance does not decompose into components as neatly as was hypothesized. Lohman suggests that process analyses of individual performance are most useful when there are instructionally relevant differences in how subjects perform, and these kinds of differences are not likely to be found in psychometric test items designed and selected to produce homogeneous patterns of response. The regression slope, which is often the parameter of interest in componential models, is a product of the Person × Item interaction, and being a component of measurement error it is generally minimized in standardized ability tests. Furthermore, when heterogeneous performance does occur, it is likely to create problems for the regression model, and this is particularly true if the heterogeneity is within individuals.

Another fundamental problem with this entire approach is that many of the tasks and performances that have served as the focus of intensive study are very distal to the classroom learning environments and learning activities that one would ultimately hope to affect. A beginning identification of this problem can be discerned in the comments of Lauren Resnick (1979), who raised questions about the ultimate benefits of this approach to the study of individual differences at the time it was a burgeoning enterprise:

> The assumption appears to be that these processes will be, in the main, the same ones called upon in performing important school tasks. I believe this assumption may be largely incorrect. Let me explain why. It seems probable that it is not *performance* on IQ tests that involves the same processes as learning in school. Rather it is *learning how* to perform well on IQ tests that involves the same processes. . . . I think it is highly unlikely that current IQ tests, even with the reinterpretation that is now proceeding, will be able to do this. Instead we will need tests that are more direct measures of the processes involved in *learning* to perform test and school tasks. I think these tests are likely to be actual samples of learning on tasks chosen to display the relevant processes as directly as possible. (Resnick, 1979, pp. 212–213)

Resnick's comments, although critical of the logic of the general approach, maintain a focus on the processes of cognition, an emphasis that was the dominant theoretical motif in cognitive psychology during the 1970s. The emphasis on process continued during the 1980s, although the nature of the tasks that were studied changed substantially as the field of cognitive psychology matured. In particular, researchers began to move beyond the study of processing in artificial or ''toy'' tasks and took on the serious and difficult task of modeling performance in complex domains of achievement in which knowledge as well as process was a significant hallmark of accomplishment and expertise.

Early examples of cognitive task analysis in school subjects (e.g., Klahr, 1977; Lesgold, Pellegrino, Fokkema, & Glaser, 1977) used information-processing concepts to describe various domains of achievement. Detailed models were developed and applied to the analysis of procedural knowledge components of the mathematics curriculum, including addition, subtraction, multiplication, and operations with fractions (Brown & Burton, 1978). Performance in a domain such as geometry required a more complicated set of models encompassing procedures and knowledge structures represented within a production system architecture, including cognitive mechanisms for setting goals and searching a problem space. In other complex domains of learning and performance, the modeling focused on conceptual knowledge, which was represented as semantic and propositional networks. Attention also turned to the analysis of forms of problem representation, the transfer of learning and problem solving to new situations, and ultimately to the characteristics of knowledge and performance that distinguish novice from expert performance.

The pioneering work of de Groot (1946/1978) on chess masters introduced the study of expertise, and a quarter century passed before the follow-up studies by Chase and Simon (1973) began to influence cognitive research and theory. Despite the seeming disconnection between chess and education, the paradigm of contrasting experts and novices both informed and stimulated the subsequent study of expertise in fields as diverse as physics (Chi, Feltovich, & Glaser, 1981), volleyball (Allard & Starkes, 1980), medicine (Lesgold et al., 1988; Patel & Groen, 1986), and writing (Bereiter & Scardamalia, 1987). A psychology of subject-matter expertise and complex learning slowly developed, and the study of individual differences became firmly embedded in domains of endeavor that were acquired over long periods of time.

Ultimately, the study of expertise and competence emerged as an alternative approach to the study of individual differences, one that was no longer linked to psychometrically defined constructs of aptitude or intelligence and their associated artifacts such as test item types. Rather, the study of individual differences focused on attained knowledge and related cognitive processes that are the object of deliberate instruction, practice, and learning. Verbal protocols and expert-novice comparisons provided the building blocks for a theory of expertise that described the acquisition and structure of declarative and procedural knowledge in various domains of human performance (Chi, Glaser, & Farr, 1988). In contrast to process-based aptitude performance, which was the focus of much of the research conducted in the 1970s, expertise is both knowledge based and process based; the primary process, and perhaps the characterizing feature of expertise (see, e.g., Bereiter & Scardamalia, 1993), is the continuous acquisition and restructuring of domain-based knowledge.

By the beginning of the 1990s, cognitive psychology and the study of individual differences had matured to the point where the groundwork was laid for fulfilling three of the four conditions necessary to achieve a psychology of instruction (Glaser, 1976). Analyses of competence and expertise were plentiful. Numerous descriptions existed of novice performance and misconceptions in multiple domains of achievement. There were even the beginnings of work focused on the design of conditions that foster the acquisition of competence taking into account theories of learning and expertise, including an extensive body of work on metacognitive monitoring (Brown & Palinscar, 1989), self-explanation (Chi, Bassock, Lewis, Reimann, & Glaser, 1989), and the deliberate employment of various strategies for text comprehension (Cote, Goldman, & Saul, 1998; Goldman, 1997). All of the work was firmly rooted in a rich psychology of cognition and cognitive development.

While it may have been overly optimistic, the argument was nonetheless made that the stage had been set for substantial changes in how issues of assessment could be approached.

> Essential characteristics of proficient performance have been described in various domains and provide useful indices for assessment. We know that, at specific stages of learning, there exist different integrations of knowledge, different forms of skill, differences in access to knowledge, and differences in the efficiency of performance. These stages can define criteria for test design. We can now propose a set of candidate dimensions along which subject-matter competence can be assessed. As competence in a subject-matter grows, evidence of a knowledge base that is increasingly coherent, principled, useful and goal-oriented is displayed, and test items can be designed to capture such evidence. (Glaser, 1991, p. 26)

Such optimism about what was now known and about the possibilities for genuine change in assessment design and practice was reinforced more recently by Greeno, Pearson, and Schoenfeld (1997) in the course of evaluating many of the cumulative results of cognitive science research on expertise and understanding in the domains of mathematics and literacy. Greeno et al. (1997) frame their discussion of what is now known squarely in the context of the opportunities, issues, and conundrums that such knowledge poses for an effective integration with assessment and instructional practice. We consider these issues further in the subsequent two sections after first discussing some aspects of the policy landscape that have profoundly shaped current assessment practice.

INTEGRATING COGNITIVE THEORY AND RESEARCH WITH ASSESSMENT AND INSTRUCTIONAL PRACTICES

While cognitive theory and research made substantial progress toward understanding and explicating individual differences in a wide array of cognitive performances, the worlds of assessment and instruction also underwent a number of important changes. Critical to these changes was a fundamental policy shift regarding the purpose and objectives of education, with increased demands for tests to simultaneously monitor and promote instructional changes. These changes, which set the stage and created an impetus for connecting cognitive and curricular theory with assessment practices, are briefly reviewed here. Subsequently, we recount efforts of the last two decades to align assessment with calls for curricular changes and conclude with the implications of this work for achievement testing.

An Impetus for Change

At the time of Cronbach's 1957 address, certain seminal changes in the world of assessment were already under way. For example, standardized, norm-referenced achievement test batteries had become an established part of educational practice, and their use was as prevalent as that of aptitude test batteries, if not more so. Questions about the condition of our schools and about levels of students' academic achievement received particular attention as the Soviets' launching of Sputnik fueled fears of a decrease in America's international dominance and a threat to future national security. Stemming from these and other events was a wave of educational reform with increased emphasis on mathematics and science preparation for the academically talented.

The post-Sputnik era of the 1960s also witnessed increased access to educational opportunity, particularly with respect to students with special educational needs. Federal legislation obliged the states to assume responsibility for the provision of equal educational opportunity. In 1965, the Elementary and Secondary Education Act (ESEA), Title I, called for financial assistance and special services for low-income students and districts and required performance data on students receiving assistance and evaluation data on outcomes of funded programs. As interest and enthusiasm grew for indicators that would measure progress toward the goal of providing all students with a good education, questions were raised about the quality and content of American education.

The federal government responded with two initiatives. The first was the Equality of Educational Opportunity Survey (EEOS), which provided information on the achievement of more than 600,000 children in elementary and secondary schools. The analysis of these data, the Coleman Report (Coleman et al., 1966), documented the enormous variation in achievement of 12th graders and showed that graduation rates revealed little about what graduates learned in school. More revealing forms of assessment seemed necessary.

The second federal initiative of the 1960s was the creation of the National Assessment of Educational Progress (NAEP). As the first of the successive waves of NAEP scores appeared, clear differences in the scores of various populations of students called into question the impact of reforms and brought to the forefront issues of educational equity and opportunity. Concerns about the achievements of American students have been amplified over time by numerous documents and reports using NAEP and other indicator data, including the 1983 publication of *A Nation at Risk* (National Commission for Educational Excellence, 1983). Issuance of the latter constituted a watershed in promoting public awareness of issues regarding American education.

Searching questions about the quality of education persist, and increasingly they have taken shape through debate focused on issues of accountability. Citizens, educators, and policymakers, at multiple levels from local school districts to the federal bureaucracy, want to know whether the investments that have been made in education in the ensuing decades are reaping rewards. Thus, efforts to devise useful assessment and accountability measures have proliferated along with separate state assessment programs. The federal initiative expanded so that the NAEP could provide state data and facilitate cross-state comparisons in a number of subject areas at specific grade levels. The first trial state assessments were administered in 1992, and through the NAEP program states can now compare the performance of their students with that of students in other states and the nation at large.

The 1990s also ushered in the standards-based reform movement (e.g., Bush, 1991; National Education Goals Panel, 1991), a broad national policy agenda involving content standards (i.e., what students should know), delivery standards (i.e., how schools will ensure that all students have a fair chance of achieving the standards), and performance standards (i.e., the level at which students should know the important content). Representative of such a trend is the May 1990 action of the National Assessment Governing Board in approving a document endorsing the establishment of three national levels of subject-matter achievement: basic, proficient, and advanced. These performance standards were then slated to be incorporated into reporting of the national assessment results. The generation, reporting, and validation of outcomes relative to such standards has become an issue of considerable study and debate within the NAEP assessment program (National Academy of Education, 1993, 1997; National Research Council, 1999). Various validity concerns have been raised, including the small numbers of students supposedly performing at the proficient level. In separate developments, recent results from the Third International Mathematics and Science Study

(TIMMS) have drawn attention to variations across Grades 4, 8, and 12 in the performance of American students relative to their counterparts in other countries. The TIMMS data have also highlighted differential curricular emphases and teaching practices across countries that may account for some of the observed performance differences.

Changing the Nature of Assessment Practice

The stage was thus set for changes in cognitive, curricular, and instructional theory to have an impact on assessment practice during the decade of the 1980s. In addition to the general influences noted earlier, two specific issues are worth mentioning as they propelled research to focus the assessment-instruction debate in ways that highlighted one over the other. The first was rooted in arguments concerning the pervasive negative influence of standardized tests on instructional practices. It was claimed that the content and format of typical achievement tests unduly influenced both the "what" and the "how" of teaching and learning. Standardized tests were seen as perpetuating a focus on memorization of isolated bits of factual knowledge and procedures that could be easily retrieved on tests composed largely of multiple-choice items (e.g., Fredericksen, 1984). Such an approach to assessment and instruction was largely a product of earlier behaviorist assumptions about the nature of knowledge and learning (see also Greeno et al., 1997). In contrast, important aspects of cognition and learning such as conceptual understanding, reasoning, and complex problem solving were often ignored, in part because they were more difficult to implement in the context of standardized assessments of achievement. As a consequence, these aspects of cognition and achievement were often neglected in the classroom learning milieu. The logic of the argument regarding negative consequences was that if important performances were not required on the tests, then they would not be the focus of work in the classroom.

In somewhat separate developments, the new curricular standards also fueled a desire to expand the scope of instructional practice to encompass broader goals for student learning (e.g., National Council of Teachers of Mathematics [NCTM], 1989; National Research Council, 1996). Work began on improving the content and process of classroom learning and largely deemphasized dominant assessment approaches. The latter were perceived as generally irrelevant to the types of teaching and learning that were desired. Much of the effort in this area was closely tied to emergent cognitive science understandings of the nature of reasoning and problem solving in knowledge-rich domains of achievement and the emergent efforts at curricular reform in mathematics and science.

In what follows, we first describe initial efforts to change the nature of assessment practice and the impact of these exploratory efforts on the current state of testing. We then provide examples of classroom assessment practices specifically designed to guide teaching and learning. Each approach highlights important considerations for the future of linking cognition and assessment.

Assessment-Driven Instruction

The perceived power of assessments to reform curriculum and instruction—and, as a consequence, improve teaching and learning—stimulated numerous development efforts in state and national testing programs (National Council on Education Standards and Testing, 1992). These large-scale testing efforts were influenced simultaneously by a cognitive perspective on thinking, reasoning, and problem solving in subject matters and traditional psychometric concerns for reliability and validity. With a goal toward influencing instructional practice (i.e., assessment-driven instruction), a number of significant changes were introduced to the design of assessments. First, the development process was informed by multiple perspectives, including cognitive psychologists, teachers, and subject-matter and measurement specialists. Second, the tasks generally consisted of open-ended prompts or exercises requiring students to write explanations, carry out a set of procedures, design investigations, or otherwise reason with targeted subject matter. Third, innovative multilevel scoring criteria or rubrics that gave consideration to procedures, strategies, and quality of response were favored over right/wrong scoring. Not surprisingly, given the state of the art, psychometric properties, particularly reliability, were of primary concern, as were practical considerations such as the time and cost of administration and scoring.

Connecticut and California provided models for other state testing programs, as they were among the first to develop and pilot new forms of assessment as part of statewide testing programs. Connecticut experimented with a variety of item or task forms, including extended performances involving groups of students, oral presentations, self-assessment, and model-based reasoning. California emphasized written explanations in mathematics and science and coordinated performance-based assessments tied to real-world situations such as recycling. Scoring rubrics were introduced as a way to judge the qualitative differences in student performance, and exemplars of various levels of understanding were provided for this purpose. Experienced teachers were recruited to develop various task forms in ways that were aligned with curricular frameworks for content and cognitive outcomes (e.g., application, inference). Furthermore, large numbers of teachers participated in the scoring of statewide samples of student responses to these new forms of assessment.

The application and use of rubrics to describe performance was viewed as a means to provide large-scale professional development in ways that would potentially affect classroom practice. Teachers, it was thought, would take from their scoring experiences innovative ideas for changing their own instructional practices, particularly an appreciation for observable qualitative differences in performance that signal relative understanding in a subject matter. The assumption that teachers will teach to the test rests on the notion that good tests will lead to good instruction. However, as we describe later, empirical support for the quality of the measures was not forthcoming for some time.

At about the time various states were experimenting with new forms of assessment, researchers in a number of disciplines began efforts to develop and articulate procedures for the design and evaluation of performance-based assessments that were predicated on constructive notions of teaching and learning (e.g., Baker, Freeman, & Clayton, 1991). Essentially, these exploratory efforts were subject specific, guided by a diverse range of participants (teachers, subject-matter and measurement specialists), and pursued under the broad umbrellas of cognitive psychology, curriculum reform, and traditional psychometric concerns for reliability and validity. Two examples follow, one in mathematics and one in science.

QUASAR (Quantitative Understanding: Amplifying Student Achievement and Reasoning), a national project, was designed to improve mathematics instruction for middle school students in economically disadvantaged communities (Silver & Stein, 1996). As part of this project, a set of open-ended assessment tasks was developed to monitor and evaluate the impact of the mathematics instructional program in participating schools. It is important to note that the tasks composing the QUASAR Cognitive Assessment Instrument (QCAI) were linked to important cross-program instructional goals, not within-school curriculum. Specifications for the QCAI were based on the construct domain of mathematics, as described in the *Curriculum and Evaluation Standards for School Mathematics* (NCTM, 1989), and included three components: cognitive processes (understanding mathematical concepts and procedures, problem solving, reasoning, creating mathematical arguments), mathematical content (number and operation, estimation, patterns, algebra, geometry and measurement, data analysis, probability and statistics), and modes of representation (pictorial, graphic, tabular, arithmetic and algebraic symbolic representations). A focused-holistic scoring rubric was developed for each task, and student performance was scored on a scale of 0 to 4, with each score level indicating a qualitatively different combination of mathematical knowledge, strategic knowledge, and communication skills (Lane, 1993).

In the area of science, Shavelson, Baxter, and Pine used a sampling theory approach to select tasks from a domain and then focus analysis on the extent to which students' performance on the sample of tasks was generalizable to the larger domain. Students were given a problem (e.g., determine the content of six boxes) and relevant equipment (batteries, bulbs, wires) and asked to document their answer and how they arrived at that answer (i.e., evidence). This task-based approach to assessment development was deliberately sensitive to the differential effectiveness of procedures or solution strategies as an essential criterion for evaluating performance (Baxter, Shavelson, Goldman, & Pine, 1992). For each of the developed assessments, student performance was scored in terms of the correctness of the answer and the quality of the procedures, strategies, or evidence used to support one's answer. Generalizability studies indicated that student performance was low and inconsistent across tasks and that assessment methods (e.g., hands on, computer, short answer) tapped different aspects of performance (Shavelson, Baxter, & Gao, 1993). Correlational evidence suggests that the performance-based measures are only moderately associated with standardized measures of aptitude and achievement. Furthermore, the performance of students with differing instructional histories was distinguishable on knowledge-based tasks and not on the tasks that required science-process skills with minimal demands for science content knowledge.

Initial efforts, such as those described, to develop alternative forms of assessment were of great interest to various audiences, and the outcomes of these efforts laid important groundwork for ensuing developments. Technical concerns, particularly the generalizability of performance across tasks, and practical concerns for cost of developing, administering, and scoring these new forms of assessment led to more complex assessment designs in which multiple methods were mixed. Content coverage could be accomplished through multiple-choice items, and cognitive concerns for thinking and reasoning might be more appropriately measured through a select number of constructed response or performance-based measures.

Their impact on large-scale testing (e.g., the NAEP program) notwithstanding, these efforts had minimal impact on instruction, in part because they were somewhat divorced from specific opportunities to learn. Rather, samples of tasks from curricular frameworks were administered and scored regardless of whether students had studied the topic or not. Typically, content coverage matters overrode cognitive considerations in task development, in part because there was no experience, guidelines, or models for how to do otherwise. More important, a number of studies suggested that the relatively poor performance of students was indicative of a lack of domain-specific knowledge, which imposed constraints on the ability of students to think and reason with their knowledge in differing contexts (Baxter, Elder, & Glaser, 1996; Baxter & Glaser, 1998; Hamilton, Nussbaum, & Snow, 1997; Magone, Cai, Silver, & Wang, 1994). The message emerging from this entire line of activity is that embedding assessments in the curriculum is essential if the goal is to use assessment innovations to influence and shape teaching and learning in ways consistent with instructional goals envisioned by reformers.

Validity and Effectiveness Concerns

As noted earlier, much of the assessment development effort in the 1980s through the early 1990s was guided by curricular frameworks that articulated in various forms the content and cognitive constructs to be assessed. Goals for assessing higher order thinking, problem solving, reasoning, and the like were stated prior to development of test items and tasks, but procedures for evaluating the extent to which these goals were achieved in the final assessment batteries were seldom undertaken. Nor were procedures or structures available to guide assessment development to ensure that goals were achieved. Linn, Baker, and Dunbar (1991), among others, cautioned that surface-level changes in assessments (e.g., constructed response, multiple scoring levels) were not sufficient to ensure that relevant cognitive processes were being tapped or that the tasks were any more complex than traditional forced-choice items. They suggested a set of "special" validity criteria to emphasize the unique characteristics of these "new" forms of assessment. These include consequences, fairness, transfer and generalizability, cognitive complexity, content quality, content coverage, meaningfulness, cost, and efficiency.

Techniques and methods for gathering and evaluating relevant evidence of these criteria have only begun to appear in the literature. Hamilton, Nussbaum, and Snow (1997) report on the use of interview procedures to supplement traditional psychometric analysis. Fifth-grade students were observed and interviewed while carrying out a science performance task. Six categories of cognitive demands were identified from the protocols: use of working memory, use of language and communication, metacognitive skills, application of prior knowledge and expectations, acquisition of new knowledge, and use of scientific processes. Analysis of a number of performance tasks led the authors to conclude that "interviews help in a new way to identify the actual cognitive demands and affordances of the test and the function in test performance of different kinds of content knowledge and reasoning for different students" (Hamilton, Nussbaum, & Snow, 1997, p. 199).

In a related effort, Baxter and Glaser (1998) subjected a variety of assessment tasks and scoring criteria to a cognitive analysis, using as a basis for their approach cognitive theories of expertise and competence. For this purpose, they developed an analytic framework that considers the nature of cognitive activity likely to be elicited by tasks with particular requirements for science content and science process skills. The cognitive activities, derived from studies on expertise, included problem representation, monitoring, strategy use, and explanation. As the subject-matter demands for content increase from lean to rich and the science process demands increase from constrained to open, consequent changes in opportunities for students to display each of these cognitive dimensions are afforded. For example, tasks that are content rich but process constrained permit opportunities for explanation but little else. In contrast, tasks that are content rich and process open provide opportunities for students to display knowledge-based differences in problem representation, monitoring, strategy use, and explanation.

Using this cognitive-content framework, Baxter and Glaser analyzed the extent to which alternative assessments are measuring cognitive capabilities that distinguish levels of competence. Examples from analyses of a diverse range of science assessments developed in state and district testing programs illustrate (a) matches and mismatches between the intentions of test developers (i.e., measure higher order thinking) and the nature and extent of cognitive activity elicited in an assessment situation and (b) the correspondence between the quality of observed cognitive activity and performance scores. In particular, a number of the tasks did not provide opportunities for students to display differential levels of understanding. In other situations, scoring systems were designed in ways that attended to the superficial aspects of performance—those that are easy to count, for example—rather than the quality of the thinking and reasoning underlying performance.

As this work demonstrates, changes in task format and scoring criteria did not always result in fundamental changes in what was being assessed. The problem, in part, stemmed from a lack of sufficient theory and technique to guide a more informed approach. In part, however, the problem stemmed from the context in which the assessments were designed and used (i.e., large-scale testing programs). As we describe in the following sections, the link between the instructional experiences of students and measures of learning is easier to forge within a classroom than across schools and districts in a state.

Integrating Assessments With Instructional Practice

During the 1990s, a body of work emerged that pursues a different route to improving learning outcomes by focusing directly on integrating assessment into classroom practice rather than attempting to change classroom practice indirectly through large-scale, externally administered assessments. An underlying premise behind this work is that the development and implementation of classroom-based assessment is fundamental to creating enhanced learning outcomes and opportunities for all students, especially when dealing with challenging subject matter and high expectations regarding conceptual understanding, reasoning, and transfer. As such, cognitive theories of knowledge and performance served as a basis for the design of learning environments and associated assessment practices to be used in diagnosing student understandings, monitoring the effects of instruction on learning, and promoting knowledge construction in these situations (e.g., Cognition and Technology Group at Vanderbilt [CTGV], 1997, 1998; Duschl & Gitomer, 1997; Hunt & Minstrell, 1994; Minstrell, 1992, 1999; White & Fredericksen, 1998).

Classroom-based formative assessment strategies have also been recognized as fundamental to implementing the new standards for instruction recommended by groups such as the National Research Council (1996) and the National Council of Teachers of Mathematics (1989). Such standards emphasize the importance of teaching in ways that promote deep understanding by students. To accomplish such ends for all students, teachers need to be more aware of the preconceptions that their students bring to new learning situations, to teach in ways that make students' thinking "visible" to themselves and other students, and to help students reflect on and reconcile their conceptions with those of others. Formative assessment thus becomes an essential part of the repertoire of effective teaching behaviors.

A recent review of classroom-based formative assessment by Black and Wiliam (1998) shows that such practices can have significant effects on overall student learning outcomes, with typical effect sizes of between .4 and .7. However, the same review reveals relatively sparse evidence of widespread deployment of such formative assessment practices. The Black and Wiliam findings are understandable, since the integration of formative assessment into classroom instructional practice can be a time- and information-intensive process. It must be managed, monitored, and used in environments with high levels of information flow and potential information overload. Thus, it is often difficult for teachers to instantiate and sustain serious formative assessment in environments that lack sophisticated data storage, manipulation, and feedback systems. It is not surprising, then, that various forms of technology have often proven useful in implementing successful formative assessment strategies at the classroom level. This is not to imply, however, that formative assessment integral to instructional practice is predicated on a technology infrastructure. What matters most is the careful probing and analysis of student understandings, which leads to sensitive student-specific adjustments in the overall learning environment, thereby influencing individual learning trajectories.

Two different sets of examples serve to illustrate work that has been undertaken on classroom-based formative assessment strategies. Common among these efforts is a focus on knowledge-rich domains, an emphasis on the development of conceptual understanding in a domain, and a view that assessment development should stem from an analysis of the various ways in which students conceptualize or explain situations or events. That is, an understanding of knowledge development in the domain serves as the guide for developing instruction and assessment. The assessment, in turn, provides feedback on instruction by calling attention to levels of student understanding at various times and in various contexts. Explicit attention to the iterative nature of teaching, learning, and assessment is the hallmark of the approaches described next. The differences lie primarily in the breadth of the subject matter that is the focus of this instruction-assessment cycle.

Mental Models and Misconceptions

Much of the cognitive research on mental models has been done in the domain of physical science, and research on classroom-embedded assessment draws heavily upon this work to support a process of cognitive diagnosis. The latter permits subsequent instructional contexts to be selected so as to challenge a student's conceptions of scientific phenomena and events and guide the student in the direction of increasingly sophisticated and complex conceptual understandings (e.g., Hunt & Minstrell, 1994; Minstrell, 1992; Minstrell & Stimpson, 1996; White & Fredericksen, 1998). Thus, the problems presented to students may not look substantially different from those found in a normal physical science curriculum. Rather, at thc heart of the difference is how such problems are posed, how student responses are evaluated, and the role of each within an overall instructional system.

Minstrell and his colleagues (Hunt & Minstrell, 1994; Levidow, Hunt, & McKee, 1991) developed Diagnoser, a relatively simple computer program designed to evaluate the consistency of students' reasoning in particular situations. The system was designed from an analysis of students' understanding of physics problems and the categories or pieces of knowledge (termed facets) that students apply when solving these problems. "A facet is a convenient unit of thought, an understanding or reasoning, a piece of content knowledge or strategy seemingly used by the student in making sense of a particular situation" (Minstrell, 1992, p. 2). Facet clusters, such as the one shown in Table 1, are sets of related elements grouped around a physical situation (e.g., forces on interacting objects) or around

TABLE 1
Separating Fluid/Medium Effects From Gravitational Effects: Facets of Student Understanding

310—pushes from above and below by a surrounding fluid medium lend a slight support

311—a mathematical formulaic approach (e.g., rho × g × h1 − rho × g × h2 = net buoyant pressure)

314—surrounding fluids don't exert any forces or pushes on objects

315—surrounding fluids exert equal pushes all around an object

316—whichever surface has greater amount of fluid above or below the object has the greater push by the fluid on the surface

317—fluid mediums exert an upward push only

318—surrounding fluid mediums exert a net downward push

319—weight of an object is directly proportional to medium pressure on it

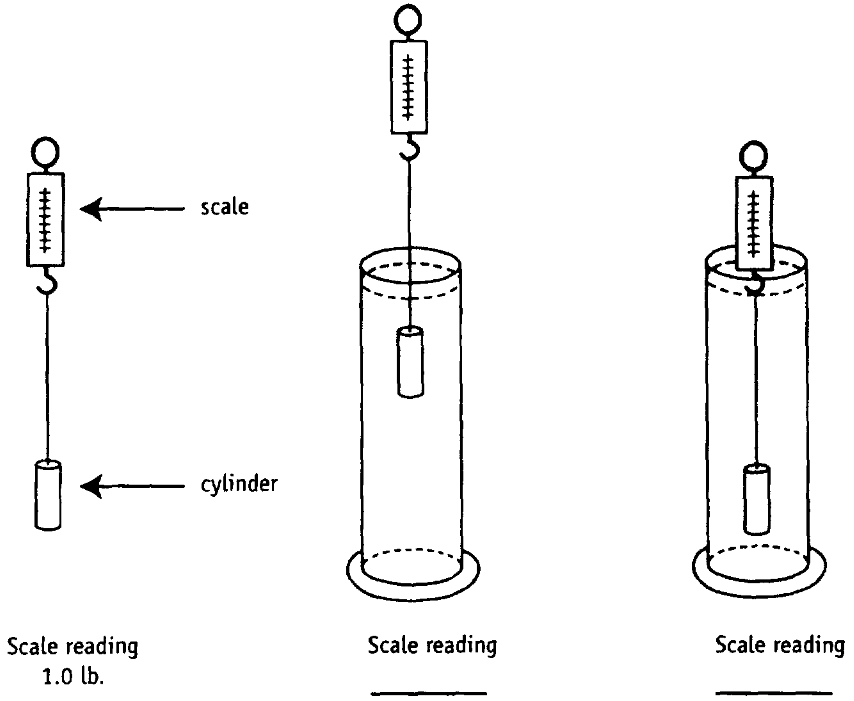
some conceptual idea (e.g., meaning of average velocity). The individual facets of students' thinking refer to individual pieces, or constructions of a few pieces of knowledge and/or strategies of reasoning. Within a cluster, facets can be sequenced in an approximate order of development that ranges from appropriate, acceptable understandings for introductory physics to those that represent limited understandings or, in some cases, serious misunderstandings. Facets in the middle of this range frequently arise from formal instruction but may represent overgeneralizations or undergeneralizations in a student's knowledge structure.

Systematic knowledge of the levels at which students understand and represent physical concepts, principles, and/or situations is the starting point for developing highly informative assessment tasks. Figure 1 is an example of a constructed response item designed to probe levels of understanding from the facet cluster in Table 1. As discussed by Minstrell (1999), student responses to this item can be mapped to the facets in the cluster shown in Table 1 in a relatively straightforward manner. Students may be thinking that weight is due to the downward push by air (319), or they may believe that fluids (air or water) only push downward (318) or only push upward (317); that fluids push equally from above, below, and all around (315); that fluids do not push at all on objects in them (314); or that there is a differential in the push depending on how much fluid is above or below the object (316). If they do understand that there is a greater push from below than from above due to the greater pressure at greater depth, they may express it in a formulaic way (311) or with a rich conceptual description (310).

Single items such as that shown in Figure 1, even when coupled with qualitative evaluation frameworks such as the facet cluster in Table 1, seldom provide sufficient information to ascertain the specificity versus generality and appropriateness of a student's understanding. However, sets of items, or item families, can be constructed to assess the context specificity of a student's understanding. By considering the response patterns across pairs or sets of items such as those shown in Figures 1 and 2, an evaluation can be provided of how much a student's

FIGURE 1
Example Constructed-Response Item: Separating Fluid/Medium Effects From Gravitational Effects

A solid cylinder is hung by a long string from a spring scale. The reading on the scale shows that the cylinder weighs 1.0 lb.



About how much will the scale read if the cylinder which weighs 1.0 lbs. is submerged just below the surface of the water? What will it read when the cylinder is much deeper in the water?

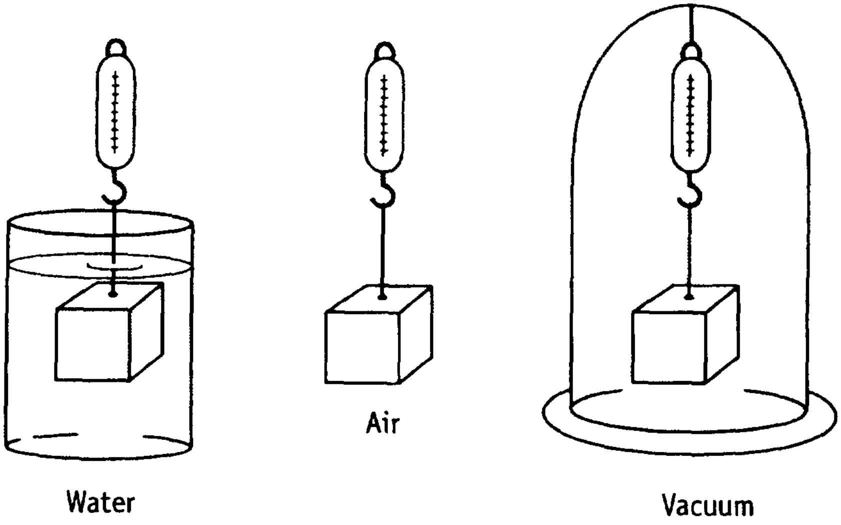
Briefly explain how you decided.

Figure 1 is reprinted from *Grading the Nation's Report Card,* National Research Council, 1999; reprinted with permission of the National Academy Press, Washington, DC.

understanding is tied to the specific surface situation described in a given problem. For the conceptual domain illustrated, it is not uncommon for student understanding of the effects of a medium to achieve a more sophisticated level for the water than air context. Minstrell (1999) has indicated that interpretable patterns of responding across items can be obtained for various physical concepts and situations. It is also worth noting that many of the tasks and associated scoring rubrics developed by Minstrell and others for use in a classroom setting can be used in large-scale survey assessments as well (National Research Council, 1999).

FIGURE 2
Example Multiple-Choice Item: Separating Fluid/Medium Effects From Gravitational Effects

These pictures show three identical blocks attached to the spring scale. In one case the block is in the water, in another it is in air, and in the third the block is in a vacuum. In the air, the scale represents 10 lbs. to the nearest 0.1 lb.



The scale readings would be

A. about the same in all three environments.
B. noticeably less in water but about the same in air and in a vacuum.
C. noticeably less in air and in water.
D. noticeably more in water and noticeably less in a vacuum.

Figure 2 is reprinted from *Grading the Nation's Report Card*, National Research Council, 1999; reprinted with permission of the National Academy Press, Washington, DC.

The Diagnoser computer program presents sets of problems such as the ones illustrated and records student responses and justifications as a means of identifying a student's understanding; then it provides instructional prescriptions when needed. The classroom instructor is also provided information about the range of student understanding for the class and thus can adjust lessons accordingly to assist students in developing a proper conceptual representation of the given problem domain. Minstrell (1999) has described how an integrated formative assessment and instruction system can produce significant changes in student achievement levels and conceptual understanding in various areas of the high school physics curriculum.

In similar work, White and Fredericksen (1998) direct their attention to learning and assessment environments that help young students acquire appropriate mental models for basic physical laws and their application across situations. Computer-based representations challenge students' existing conceptions and stimulate cross-student debates and experimentation to resolve discrepancies between what students think and what evidence from various inquiries seems to demonstrate. A cyclical sequence of hypothesize, test, generalize is promoted and supported by the software and the overall instructional design.

Support for Problem-Based Learning

The second set of examples differs from the work just described by focusing on larger units of scientific inquiry and instruction that have the character of more extended problem-based and project-based learning situations. Frequently, such situations engage students in individual and collaborative problem-solving activities organized around a complex real-world scenario. Such inquiry-based activities have been recommended in various standards for mathematics and science learning and teaching (e.g., NCTM, 1989; NRC, 1996).

An example of embedding assessment strategies within such extended inquiry activities can be found in work pursued by the Cognition and Technology Group at Vanderbilt on the development of a conceptual model for integrating curriculum, instruction, and assessment in science and mathematics (Barron et al., 1995, 1998; CTGV, 1994, 1997). The resultant SMART Model (Scientific and Mathematical Arenas for Refining Thinking) involves frequent opportunities for formative assessment by both students and teachers and includes an emphasis on self-assessment to help students develop the ability to monitor their own understanding and find resources to deepen it when necessary (Brown, Bransford, Ferrara, & Campione, 1983; Stiggins, 1994). The SMART Model involves the explicit design of multiple cycles of problem solving, self-assessment, and revision in an overall problem-based to project-based learning environment.

Activity in the problem-based learning portion of SMART typically begins with a video-based problem scenario such as the "Stones River Mystery" (SRM; Sherwood et al., 1995). SRM tells the story of a group of high school students who, in collaboration with a biologist and hydrologist, are monitoring the water in Stones River. The video shows the team visiting the river and conducting various water quality tests. Students in the classroom are asked to assess the water quality at a second site on the river. They are challenged to select tools that they can use to sample macroinvertebrates and test dissolved oxygen, to conduct these tests, and to interpret the data relative to previous data from the same site. Ultimately, they find that the river is polluted owing to illegal dumping of restaurant grease. Students then must decide how to clean up the pollution.

The problem-based learning activity includes three sequential modules: macroinvertebrate sampling, dissolved oxygen testing, and pollution cleanup. Each follows the same cycle of activities: initial selection of a method for testing or cleanup, feedback on the initial choice, revision of the choice, and a culminating task. The modules are preliminary to the project-based activity in which students conduct actual water quality testing at a local river. In executing the latter, they are provided with a set of criteria by which an external agency will evaluate written reports and accompanying videotaped presentations.

Within each activity module, selection, feedback, and revision make use of the SMART Web site, which organizes the overall process and supports three high-level functions. First, it provides individualized feedback to students and serves as a formative evaluation tool. The feedback suggests aspects of students' work that are in need of revision and classroom resources that students can use to help them revise. The feedback does not tell students the ''right answer.'' Instead, it sets a course for independent inquiry by the student. The Web feedback is generated from data that individual students enter.

As an example, when students begin working on macroinvertebrates, they are given a catalog of sampling tools and instruments. Many of these are ''bogus'' and collect the wrong kind of sample; others are ''legitimate'' and will gather a representative sample of macroinvertebrates. The catalog items are specially designed to include contrasting cases that help students discover the need to know certain kinds of information. Students are asked to choose and justify their choice of tool. To help them make their choices, they are provided with resources, some of which are on-line, that they can use to find out about river ecosystems, macroinvertebrates, and water quality monitoring. Once students have made an initial set of choices, they use the SMART Web site. They enter their catalog choices and select justifications for their choices. Once students have submitted their catalog order on-line, the SMART Web site sends them individualized feedback. The catalog items and foils are designed to expose particular misconceptions. The feedback that students receive from the SMART Web site highlights why the selected tool is problematic and suggests helpful resources (sections of on-line and off-line resources, hands-on experiments, and peers). This form of feedback has been used in similar work on mathematics problem solving, and results suggest that it can be an effective stimulus for guided inquiry and revision by students.

The second function of the SMART Web site is to collect, organize, and display the data collected from multiple distributed classrooms (''SMART Lab''). Data displays are automatically updated as new data are submitted to the database by students. The data in SMART Lab consist of students' answers to problems and explanations for their answers. Each class's data can be displayed separately from the distributed classroom's data. This feature enables the teacher and her or his class to discuss different solution strategies and, in the process, address important concepts and misconceptions. These discussions provide a rich source of information for the teacher on how her or his students are thinking about a problem and are designed to stimulate further student reflection.

The third section of the SMART Web site consists of explanations by student-actors (''Kids Online''). The explanations are text based with audio narration, and they are errorful by design. Students are asked to critically evaluate the explanations and provide feedback to the student-actor. The errors seed thinking and discussion on concepts that are frequently misconceived by students. At the same time, students learn important critical evaluation skills.

The ability of students and teachers to make progress through the various cycles of work and revision and achieve an effective solution to the larger problem depends on a variety of resource materials carefully designed to assist the learning and assessment process. Students who use these resources and tools learn significantly more than students who go through the same instructional sequence for the same amount of time, but without the benefit of the tools and the embedded formative assessment activities, and their performance in a related project-based learning activity is significantly enhanced (Barron et al., 1995, 1998).

Incomplete Solutions and Unresolved Conceptual Problems

In the preceding discussion, we highlighted two general ways in which assessment practice has been connected with concepts arising from theories of cognition and learning. In the first set of examples, large-scale and largely summative assessment practices have been expanded with the goal of sampling a wider range of cognitive performances and affecting instructional practice at the classroom level. In the second set of examples, formative assessment tactics have been incorporated directly into teaching strategies for complex content with the goal of enhancing student learning outcomes. Both lines of work are important because they take the enhancement of classroom-based learning and instructional processes as the nexus for the connection between cognitive theory and assessment practice.

While much has been learned through these efforts, each approach has specific limitations. For example, we have described how significant attention to innovations in task format and concern for instructionally informative scoring criteria frequently conflicted with traditional concerns for reliability and validity (e.g., Dunbar, Koretz, & Hoover, 1991; Shavelson, Baxter, & Gao, 1993; Shavelson, Baxter, & Pine, 1991). Despite changes in task format and score emphasis, efforts to develop alternative assessments proceeded without a theory of assessment aligned with contemporary research on cognition and learning. Little attention was given to the underlying assumptions that "new" forms of assessment are direct measures of complex performances and that they can and do change classroom teaching and learning in positive ways. Indeed, procedures for evaluating the cognitive aspects of assessment situations, once developed, pointed to the difficulty in developing tasks and scoring systems consistent with goals for measuring thinking, reasoning, and problem solving (Baxter & Glaser, 1998; Hamilton et al., 1997).

It also appears to be the case that as assessment development and use moves away from the classroom teaching and learning situation, validity issues seem to take a back seat to issues of reliability and generalizability. In contrast, when assessments are integral parts of instructional practice, validity appears to be the primary issue and very often the major technical criterion by which assessments are judged. It is unfortunate that a better balance has yet to be achieved within and across approaches. By selectively focusing on a specific assessment purpose (summative vs. formative) as applied to a specific assessment context (large scale and high stakes vs. classroom based and low stakes), one or more critical issues of inference are largely ignored.

Available evidence also suggests that traditional methods for test design supplemented by cognitive analysis of content and process can usefully inform test development so that the stated goals may be more closely approximated than has generally been the case. However, layering new techniques on top of the old is but a temporary solution to the development of appropriate measures of subject-matter achievement. What is needed is a fundamentally different approach to assessment design, an approach that is grounded in theories of developing competence in specific subject matters and that is supported by relevant psychometric analyses. To inform educational practice, it is essential to define critical differences between successful and unsuccessful student performance and model these differences in a way that makes relevant cognitive activity apparent to teachers and students. Psychometric technology that has emerged in the context of selection and aptitude testing is not particularly appropriate for these purposes of achievement assessment. Clearly, no approach we have considered thus far attempts to fully confront the conceptual conflict between theories of cognition and learning and contemporary models and methods of psychometric technique. Before evaluating the effectiveness of attempts at more synthetic solutions, it is therefore important to examine the scope of the conceptual divide separating contemporary cognitive and psychometric theories.

As we have indicated, cognitive theory and research emphasize the knowledge structures and processes underlying understanding in various substantive performance domains. In considering such matters, the cognitive perspective also focuses on the ways in which such understandings are constructed by individuals. Such a perspective on the nature of knowledge and skill thus raises serious questions about what should be assessed and the manner of assessment. With regard to the latter, it has been argued that the assessment technologies currently in use to develop, select, and score test items and tasks, and thus to determine summary scores, treat content domains and cognition as consisting of separate pieces of information (e.g., facts, procedures, and definitions). This fragmentation of knowledge into discrete exercises and activities is the hallmark of "the associative learning and behavioral objectives traditions," which dominated American psychology for most of this century (Greeno et al., 1997). This "knowledge in pieces" view has dominated learning theory and instructional practice in America, as well as assessment and testing technology (Mislevy, 1993). Much of current testing technology is based on an underlying theory that allows tasks to be treated as independent, discrete entities that can be accumulated and aggregated in various ways to produce overall scores. Furthermore, test forms are compiled according to a simple substitution of one item for another or one exercise for another based on parameters of item difficulty. Mislevy (1996) captured part of the conceptual clash between a cognitively based approach to assessment and current psychometric methods.

> To some extent in any assessment comprising multiple tasks, what is relatively hard for some students is relatively easy for others, depending on the degree to which the tasks relate to the knowledge

> structures that students have, each in their own way, constructed. From the trait-behavioral perspective, this is noise, or measurement error, that leads to: low reliability under classical test theory (CTT); low generalizability under generalizability theory; and low item discrimination parameters under item response theory (IRT). It obscures what one is interested in from that perspective, namely, locating people along a single dimension as to a general behavioral tendency as defined in terms of this particular domain of tasks. For inferences concerning overall proficiency in this sense, tasks that do not line people up in the same way are less informative than ones that do. (p. 392)

Standard test theory, that is, both classical true-score theory and item response theory, appears to be largely incompatible with the implications and findings of contemporary psychological theory and research. The former arose from the pressures to make selective decisions that required the ranking of students from high to low aptitude or achievement. Classical test theory (CTT) originated at the turn of the century in the work of Spearman (1904a, 1904b). Over time, the statistical tools available for analyses within the realm of CTT grew in both number and sophistication (e.g., Lord & Novick, 1968). However, the confounding of items (making up a test) with persons (taking the test) is a major shortcoming of CTT. There is no direct mechanism within CTT to compare test scores that are derived from different sets of items or obtained from groups of different ability levels.

Originating in the work of Lord (1952) and Rasch (1960), item response theory (IRT) was developed in response to these shortcomings. IRT proposes that an individual's performance is a product of his or her proficiency, and the probability of responding correctly to a given item is a function of his or her overall proficiency parameter and one or more item parameters (e.g., difficulty). As such, the item response rather than the overall test score becomes the object of measurement. When an IRT model can be supported, the scores on any two subsets of items can be compared on the same scale of measurement. Likewise, the scores of groups of different developmental or ability levels can be compared at one or multiple points in time.

Because of these features, IRT has facilitated contemporary practice in test design and analysis. For example, adaptive testing relies on estimates of which item would provide the most information given an individual's proficiency and responses to previous items. IRT has also had a major influence on large-scale educational assessment programs such as NAEP by making it possible to equate sets of items administered at different points in time and under different subject-matter frameworks. For example, multiple matrix sampling and IRT were used in developing a reading scale for the 1984 NAEP (Educational Testing Service, 1985) that allowed comparisons of reading scores across assessments conducted from 1971 to 1984 and across age cohorts of 9-, 13-, and 17-year-old students. Despite the technical and practical advantages of IRT methodology, characterizing test performance as the product of a unidimensional proficiency variable is at odds with current conceptions of achievement. Furthermore, there has been very little change in how understandings of the nature of developing subject-matter competence should influence the content, item types, or interpretation of tests.

Contemporary cognitive theorists would argue that inferences about the nature of a student's level of knowledge and achievement in a given domain should not focus on individual, disaggregated bits and pieces of information arrayed along a unidimensional item difficulty scale. More important than the questions students answer correctly is the overall pattern of responses that students generate across a set of items or tasks. The pattern of responses reflects the connectedness of the knowledge structure that underlies conceptual understanding and skill in a domain of academic competence. Thus, it is the pattern of performance over a set of items or tasks explicitly constructed to discriminate between alternative profiles of knowledge that should be the focus of assessment. The latter can be used to determine the level of a given student's understanding and competence within a given subject-matter domain. Such information is interpretive and diagnostic, highly informative, and potentially prescriptive. The examples provided earlier in this section for science learning (e.g., CTGV, 1997; Minstrell, 1992) illustrate just such an approach to assessment (see also National Research Council, 1999).

Two other features of subject-matter competence merit attention in designing assessments that are responsive to the changing educational environment. First, competence is defined, in part, by the extent to which knowledge and skills are transferable and applicable in a variety of tasks and circumstances. To know something is not simply to reproduce it but to be able to apply or transfer that knowledge in situations that are more or less similar to the originally acquired competence. Second, people's knowledge, understanding, and skill are reflected in their capacity to carry out significant, sustained performances both independently and in collaboration with others in a group. Especially important are group situations that emphasize distributed expertise and the sharing of knowledge across individuals to enable successful performance of a major task.

PURSUIT OF THEORY AND CONSTRUCT-DRIVEN ASSESSMENT

Efforts to develop alternative assessments aligned with highly sought after instructional changes, and advances in statistical method and technique for addressing important measurement problems, have for the most part proceeded with minimal attention to the developments in cognitive psychology. It is also clear that the task- or item-based approach to assessment design that relies on postdevelopment statistical criteria such as item difficulty to make decisions about the final test form has been influenced little by new developments. As noted by Mislevy (1993), ''It is only a slight exaggeration to describe the test theory that dominates educational measurement today as the application of 20th century statistics to 19th century psychology'' (p. 19). Thus, the problem we continue to face is how theory and research on cognition and instruction can allow us to go beyond current testing and assessment practices to achieve an understanding of an individual's underlying cognitive competence that can also be of benefit to instructional practice.

These problems are not new. The integration of assessment practices with advances in knowledge of cognition and human learning has been advocated for some time (Anastasi, 1967; Cronbach, 1957; Glaser, 1981; Nickerson, 1989; Resnick & Resnick, 1992). Nevertheless, efforts to articulate the rationale for such an approach through discussions of relevant cognitive research and its implications for assessment design were slow to develop (e.g., Glaser, 1986; Mislevy, 1993; Snow & Lohman, 1989, 1993; Sternberg, 1984; Wittrock & Baker, 1991). Furthermore, the discussion generally lacked the specificity necessary to effectively guide assessment design. The essence of the dilemma was well captured by Snow and Lohman (1989):

> It seems clear to us, at least today, that cognitive psychology has no ready answers to the educational measurement problems of yesterday, today, or tomorrow. But it also seems clear to us that cognitive psychology has opened a spectrum of questions about what educational measurements do and do not represent and an equally large spectrum of methods for investigating such questions. (p. 320)

Theory and Construct-Driven Aptitude Assessment

After considering much of the work that has ensued on the cognitive analysis of aptitude, Snow and Lohman (1989, 1993) argued that developing a new generation of tests that seek to more accurately describe cognitive performance will require a new test theory, one that fully embodies a componential account of human information processing. Such an effort must be able to account for the various components of cognitive task performance such as stimulus encoding, feature comparison, rule induction, rule application, and response. In addition to explaining various cognitive components or processes, componential measurement theory will have to deal with the use of different strategies across individuals as well as shifting strategy use within individuals (e.g., Kyllonen, Lohman, & Woltz, 1984). Issues of learning and changing contributions of automatic and controlled processing must also be considered (e.g., Ackerman, 1987). Modeling and interpreting each of these dimensions, including the potential changes in these dimensions, may profit from a testing system that is adaptive (i.e., the succession of items for an individual will be based on substantive observations obtained during testing). Finally, a componential measurement theory will also have to deal with assessing the combined metrics of speed and accuracy so that their effects and interactions within and between individuals are accurately represented. Given such a complex, multifaceted model, the measurement of human performance can no longer be interpreted as representing a single dimension or trait.

Rather than develop an entirely new test theory, some have attempted to develop compound models that build on standard test theory and that include the cognitive variables believed to affect test performance. Faceted tests, proposed by Guttman (1970), are an early example of this kind of measurement application. More recent developments include hybrid approaches such as "tectonic plate" (Wilson, 1989), latent class (Haertel, 1984), and componential models (Embretson, 1984; Whitely, 1980).

A well-developed example of a hybrid approach is the incorporation of assumptions regarding cognitive complexity into Rasch models for certain types of aptitude test performance. Incorporating cognitive complexity into a Rasch model requires substituting a mathematical model of cognitive complexity for the item difficulty. An early example of such a model is Fisher's (1973) linear logistic latent trait model (LLTM), in which items are scored on factors affecting complexity. Embretson (1993) has used LLTM to model performance in two spatial ability tasks, the Space Relations Test from the Differential Aptitude Test (DAT) and the Spatial Learning Ability Test (SLAT; Embretson & Waxman, 1989). In doing so, she drew upon some of the earlier-mentioned cognitive components research on spatial ability and then demonstrated how LLTM can be used to (a) evaluate different processing models of test performance by comparing model fits, (b) evaluate the construct representation of a test using parameter estimates to determine the effect of cognitive variables that are represented by factors, and (c) select items for subsequent testing based on their cognitive representation, complexity, and difficulty.

One potential shortcoming of the LLTM method is that it requires a strong a priori model of task complexity and its relationship with difficulty. This was possible to some extent for the spatial aptitude items given considerable empirical analyses of cognitive processing factors in task performance (e.g., Kyllonen, Lohman, & Woltz, 1984; Mumaw & Pellegrino, 1984; Pellegrino & Kail, 1982; Shepard & Cooper, 1983). An alternative is the multicomponent latent trait model (MLTM; Embretson, 1983), which also requires an initial componential model but does not require the model to specify the relationships of stimulus features and complexity. To work without an a priori theory of difficulty, MLTM uses data from responses to standard items and responses to subtask items representing components of the standard item. Another latent trait model described by Embretson, the general component latent trait model (GLTM; Embretson, 1984), combines the LLTM and MLTM to create a more general model.

Embretson also has presented the multidimensional Rasch model for learning and change (MRMLC; Embretson, 1991), which treats change as a latent variable within a multidimensional item response model, thereby attempting to resolve a number of the problems and paradoxes associated with measuring change. The MRMLC, however, depends on a number of strong assumptions and requires adherence to a set design and complex administration conditions. More specifically, items are administered under successive conditions, with the first condition being the standard condition and the remaining conditions being the conditions of change, which can be either positive or negative (e.g., practice, instruction, or stress induction). Embretson (1993, p. 143) writes that MRMLC is "a dynamic, rather than a static, concept of ability" in which "performance is changing in both dimensionality and level, due to individual differences in modifiability."

Such merging of cognitive theories of performance with psychometric models and methods in the context of aptitude assessment represents a substantial advance in bridging the conceptual divide (see Fredericksen, Mislevy, & Bejar, 1993).

Work of this type clearly has significance for enhancing various aspects of selection testing and entrance exams such as the SAT and the Graduate Record Examination. Unfortunately, improvements in aptitude assessment design and development, despite their importance, provide little or no information to promote educational attainment in areas of academic achievement and instructional relevance. Furthermore, it is highly unlikely that the same merger of constructs that appear to work for certain domains of aptitude, albeit with complicated test administration and validation designs, can be productively extended to the assessment of achievement. Rather, the situation of achievement assessment is more complex, as described next.

Beyond Aptitude Assessment: Theory-Based Assessment of Achievement

While the gap between cognitive theory and assessment practice has narrowed somewhat in the world of aptitude testing, the gap has remained quite substantial in the world of achievement testing. As described earlier, achievement test developers, when faced with the challenge of developing alternative forms of assessment, generally adopted a task-centered approach as opposed to a construct-centered approach to generate assessments symmetric with instructional practice. In the intervening years, performance-based assessments and other alternatives to multiple-choice testing have come to play an increasingly central role in various state and national testing programs because of their perceived power to reform curriculum and instruction and, as a consequence, improve teaching and learning (National Council on Education Standards and Testing, 1992). In addition to their role as a political change agent, performance-based assessments are expected to measure student outcomes and provide standards for future performance, support comparisons across educational settings, and produce indices of change within and across these settings.

The perceived power of assessments to effect teaching and learning in optimal ways, although generally unsubstantiated, may be realized through the integration of assessment and learning as an interacting system. Heretofore, standardized assessment and the conditions of instruction and schooling have coexisted largely as decoupled systems. The move to standards-based assessment and reporting of performance in terms of achievement levels brings to the forefront the extent of the disconnection between the vision of reformers for equitable, sustained opportunities for acquiring knowledge and the experiences of students in various instructional settings. Given this disconnection between assessment and instruction, researchers and educators alike have called for changes that would result in test formats being more aligned with instructional tasks and test results being more useful for instructional decision making (Glaser, 1986; Linn, 1986; Nitko, 1989). Accomplishing this goal will require the integration of theories of knowledge and instruction with new psychometric models that describe acquired competence in subject-matter learning. Within this context, the theory-based constructs to be measured must be emphasized prior to test development and then used to generate item or task characteristics that are intended to influence the performance of more or less proficient students (National Research Council, 1999). In this way, assessments can be designed with predictable cognitive demands for specific groups of test takers (Nichols, 1994; Nichols, Chipman, & Brennan, 1995; Nichols & Sugrue, 1997). Messick summarized the approach as follows:

> A construct-centered approach would begin by asking what complex of knowledge, skills, or other attribute should be assessed, presumably because they are tied to explicit or implicit objectives of instruction or are otherwise valued by society. Next, what behaviors or performances should reveal those constructs, and what tasks or situations should elicit those behaviors? Thus, the nature of the *construct guides the selection or construction of relevant tasks as well as the rational development* of construct-based scoring criteria and rubrics. (1994, p. 17)

Mislevy and his colleagues have elaborated and extended Messick's ''guiding questions'' and formulated an assessment design theory centered around notions of evidence and inference. Broadly speaking, theories of learning and knowledge acquisition based in cognitive psychology provide the conceptual structure for formulating what constitutes evidence, and the statistical power of mathematical probability provides an empirical structure for drawing inferences from the attained evidence. The strength of this approach lies in its efforts to shift the focus of test design from items to substantive theory, that is, what is known about the display of differential competence in a domain, joined with the use of inferential reasoning for judging the evidence derived from the test situation. Examples of this approach to structure inferences about proportional reasoning, mixed-number subtraction, foreign-language learning, and accomplishment in a studio art program, among others, attest to the possibility for broad application in the future.

As described by Mislevy (1995, 1996), an iterative three-stage, assessment design process of necessity begins with a theory of performance in a domain, followed by generation of a set of tasks the responses to which can be analyzed/interpreted within that theory and then statistical comparisons of expected to observed performance to draw conclusions about level of student understanding within the targeted domain. Discrepancies between theoretical models of performance and patterns of observed performance in particular situations provide information for improving or changing the substantive theory underlying the initial design.

In designing assessments, one begins by characterizing differential competence or levels of knowing in a domain. For this purpose, a set of variables denotes relevant aspects of performance that signal differences in knowledge and skill (i.e., student models). These variables can take on any of a number of forms (e.g., quantitative, qualitative, or some combination) depending on the domain theory from which they are derived. The goal is to characterize performance in ways that capture the essential or critical distinctions between those with differing levels of knowledge and experience. In this regard, consideration of the key concepts in a domain, ways of understanding or misunderstanding them, and the common developmental trajectories through which learning progresses serves as an appropriate guide.

With the substantive theory of performance sufficiently delineated, tasks or situations are designed so as to provide an opportunity for students to display differential levels of understanding. Certainty of conclusions depends on the relevance of the obtained evidence to the domain of study and theory of performance. From theory and data, one posits probabilities for the ways that students with different configurations of knowledge, skills, or other distinguishing characteristics of performance (e.g., problem representation, strategy use) will solve problems, answer questions, and so on. Given a student's particular configuration of knowledge and skills, what is the probability that he or she will respond to a given task or situation in a particular way (e.g., right/wrong, quality of explanation)? These are initial performance expectations (conditional probabilities) based on expert opinion, theory, or pilot studies. These probabilities can be revised after performance is observed and inferences drawn about the appropriate or most likely level of competence or understanding as defined by the student model.

> The essential thing is to define a space of ''student models''—simplified characterizations of students' knowledge, skill, and/or strategies, indexed by variables that signify their key aspects (1995, p. 43). A properly structured statistical model embodies the salient qualitative patterns in the application at hand and spells out, within that framework, the relationships between conjectures and evidence. It overlays a substantive model for the situation with a model for our knowledge of the situation, so that we may characterize and communicate what we come to believe—as to both content and conviction—and why we believe it—as to our assumptions, our conjectures, our evidence, and the structure of reasoning. (Mislevy, 1996, p. 13)

The design just mentioned entails a myriad of complexities of practical application both theoretically and statistically. As efforts proceed in this direction, detailed examples of the design and decision-making process for creating assessments for school subjects will be necessary. Method and theory will also evolve, and techniques will need to be elaborated and refined as definitions of competence are extended. While much remains to be done at a practical level, an important cognitive-psychometric discourse has been established around achievement assessment, and a context has been established for bridging the conceptual divide in this critical domain of assessment and instructional practice.

Combining Aptitude and Achievement Assessment With Instruction: ATIs Reconsidered

The theoretical landscape now looks considerably different than it did in the 1960s and 1970s, when much of the ATI work reviewed by Cronbach and Snow (1977) was originally conducted. What can we now say about the logic and potential of such research in light of current understandings of cognition and assessment, especially with regard to the goal of using such knowledge to improve instructional and learning outcomes? This question can be approached by first examining two examples of second-generation ATI work and then considering whether the potential exists for building on such work and developing an expanded research agenda exploring AATIs: Aptitude × Achievement × Treatment interactions.

Shute's (1992, 1993) research using macroadaptive intelligent tutoring systems (ITS) can be characterized as second-generation ATI work because her hypotheses and assessments are derived from contemporary information-processing theories, particularly the ACT-R (Anderson, 1983, 1993) and four-sources (Kyllonen & Shute, 1989) theories. The ACT-R theory describes a three-stage process of cognitive skill acquisition involving structures and processes in working, declarative, and procedural memory, while the four-sources theory is a learning skills taxonomy that defines a four-dimensional space of subject matter, learning environment, desired learning outcomes, and learner attributes.

To evaluate the interaction of students' associative learning skills (AL) and different learning environments, Shute (1992) used an ITS designed to teach the basic principles of electricity, which could be operated in either of two modes, a rule-application environment or a rule-induction environment. The rule-application environment was a high-structure treatment in which each of the variables and their relationships in a problem were fully specified and explained. In the rule-induction treatment, which was a low-structure environment, the tutor identified the relevant variables in a problem, but the student was left to induce their relationships. In the criterion tasks, two types of knowledge were evaluated, with declarative assessments requiring students to answer factual questions and procedural assessments requiring students to apply Ohm's and Kirchhoff's laws in solving problems.

For declarative knowledge, Shute (1992) found that the rule-induction environment was optimal for high-AL students, whereas low-AL students acquired declarative knowledge better in the rule-application environment. For procedural knowledge, high-AL students performed better in the rule application environment and low-AL students did not perform well in either environment. Thus, there was a three-way interaction of AL, learning environment, and knowledge type. These results are best understood in terms of the overall match of the learning environment to the student's aptitude and the type of knowledge to be learned, consistent with the four-sources theory. In another ITS study that taught flight engineering, Shute (1993) found a three-way interaction of working memory capacity, general knowledge, and two types of problem sets. High-capacity/low-knowledge students performed better when assigned to an extended problem-set treatment, whereas low-capacity/high-knowledge students performed better in a constrained problem-set treatment.

Swanson's (1990) ATI work can also be characterized as second generation. His research shows the potentially dynamic nature of ATIs and how treatments may be adapted in dynamic situations to benefit the learner. Swanson employed two human tutors to teach optics to college students and trained the tutors to use three different tutoring approaches: high structure, low structure, and contingent. The latter method was based on principles such as Bruner's (1978) concept of scaffolding and Vygotsky's (1978) zone of proximal development. In contingent tutoring, learning is fostered by creating connections between what is to be learned and things that are already known, and students are given more or less support depending on the difficulty they are experiencing. Thus, the tutor switches between different levels of high- and low-structure treatment depending on how the student is performing the task. Across students varying in general aptitude, the contingent treatment was best for low-aptitude students and the low-structure treatment was best for high-aptitude students.

Swanson's study shows that students change as they develop knowledge and cognitive skills, as do the relationships of aptitudes, performance, and optimal treatments (see also Ackerman, 1988, 1989, 1996, 1997). Thus, ATIs are contextualized in the person-situation interaction, and these contexts change as learning progresses. Any research program that ignores this set of relationships not only will fall short of its potential to effect positive change but will probably find itself swamped in contradictory findings, the product of normal changes accompanying learning. Given the results and complexities of earlier ATI research, careful attention must be given to these issues; otherwise, there is the distinct possibility of history repeating itself!

The research of Shute (1992, 1993) and Swanson (1990) is informative and represents the beginning of a second generation of ATI research. For such research to have practical benefits, much more work remains to be done. The goal of such research should be to create a knowledge base that supports a system of instruction whereby individuals of mixed aptitudes can reach higher criterion levels of performance in various achievement domains. For this to come about, several points bear consideration. First and foremost, the assessments of aptitude and achievement, as well as the instructional treatments, should be based on detailed cognitive theories appropriate to each set of constructs. The ACT-R and four-sources theories provide a good starting point in terms of a substantive theory of cognitive aptitudes. Assessments based on ACT-R are likely to focus on the structures and processes of the working, declarative, and procedural memory systems. However, missing from the work of Shute and Swanson are detailed and integrated assessments of students' knowledge structures in the content domain to be learned (e.g., Mislevy, 1995, 1996; Tatsuoka, 1983).

As mentioned in the preceding discussion of achievement assessment, efforts to assess knowledge structures have begun to integrate traditional psychometric methods with new techniques such as Bayesian inference networks to create measurement models capable of assessing students' achievement relative to optimal and suboptimal domain performance models. Such models can be used to trace a student's progress through a given knowledge space and to adjust instruction accordingly. Since learning is a process of change, considerable focus should be on the changing relations of aptitudes, knowledge structures, and performance. Evaluating the changing relations of basic information-processing constructs such as working memory capacity and detailed knowledge-structure models can provide a more accurate depiction of the learning process than has previously been available. Given that knowledge structures represent levels of achievement, it may be heuristic to think of this type of investigation as an aptitude-achievement-treatment interaction (AATI) design.

One thing AATI research must emphasize relative to traditional ATI research is learning as a process of change. Historically, ATI designs involved an aptitude pretest, an educational treatment, and an achievement posttest. The posttest was then regressed on the pretest, and the interaction was assessed. If the interaction was significant, it could potentially serve as a guide to a macroadaptive treatment. In practice, however, this was rarely done, because unaccounted-for contextual effects such as change limited the generalizations that could be made from ATI results. One of the reasons for this is that change has been difficult to assess (see, e.g., Bereiter, 1963; Cronbach & Furby, 1970; Willet, 1988). However, developments in statistical techniques such as hierarchical linear modeling (Bryk & Raudenbush, 1992) and latent growth-curve modeling (Willet & Sayer, 1994) provide powerful new methods for describing change and incorporating this into the AATI design.

The optimal application of AATI findings will probably involve combined microadaptive and macroadaptive treatments that also take into account the situated nature of learning and encompass aspects of aptitude such as affect and conation (see, e.g., Snow, Corno, & Jackson, 1996). Given the complexities of microadaptation and human interaction, one possible environment for exploring and applying AATI findings is likely to be computer-based instructional and formative assessment systems that serve as an adjunct to classroom instruction. Ultimately, the goal of AATI research is to understand the development of knowledge and cognitive skill and provide optimal conditions for students with different aptitude profiles as they progress along a given learning trajectory. When this is finally done, AATI research may begin to play a truly constructive role in the educational process, in ways originally envisioned by Cronbach.

ACCOMPLISHMENTS AND ASPIRATIONS

In this chapter, we have tried to show how the relationships between the "two disciplines" of scientific psychology have changed over time since Cronbach's original plea for their unification. In so doing, we have looked at the initial atheoretical disconnection in the pursuit of Aptitude × Treatment interactions, which was followed by the cognitive analysis of aptitudes and the related cognitive analysis of performance in the study of expertise. With knowledge of the cognitive components of complex performance, the field then attempted more sophisticated and model-based understanding of the relationships between attained achievement and the conditions of learning. Assessment of performance and conditions of learning are now being studied as dynamically related events in experimental instructional situations. And efforts have been made to shift the focus of assessment design from items or tasks to important constructs derived from an understanding of the nature of competent performance in a domain and how it might be accessed, displayed, and scored in particular situations.

The historical path we have described shows continued effort to unite cognitive science with psychometrics in ways that benefited both psychology and education. The evidence is clear that we have accomplished much, even if the merger has been tentative and somewhat strained at times. Nevertheless, we remain quite removed from a theory and technology of assessment design that effectively meets the needs of various user communities and that is consistent with our current knowledge regarding the nature of expertise and achievement. Traditional disciplinary boundaries need to be redrawn and new disciplines defined in ways that focus a coordinated educational research and development agenda around the relationship between instruction and assessment. In this context, something more than the machinery of CTT and IRT is needed, and the sophistication of such an approach may well vary as a function of the aspects of cognition that are of primary concern as well as the types of inferences we wish to make and the purposes for which we make them.

In the future, test theory will have to account for differences among and within individuals in terms of knowledge structures such as schemas, mental models, and semantic and procedural networks. Relevant here are measures of current knowledge and skills as well as a means of characterizing change in the compilation and configuration of these structures as learning progresses. Given a student's particular constellation of cognitive processing efficiency, mental models, and position in a learning trajectory, what is the next step that will provide the student with optimal learning opportunities within a domain? This is not a new idea, particularly as regards individual variation among learners vis-à-vis the goals of testing and assessment in the schools.

> Teachers and schools need information on individuals that is oriented toward instructional decision rather than prediction. Tests in a helping society are not mere indexes which predict that the individual child will adjust to the school or which relieve the school from assisting the student to achieve as much as possible. The test and the instructional decision should be an integral event. (Glaser, 1981, p. 924)

Creating conditions that support learning and foster subject-matter competence requires the adoption of new, more enlightened and productive attitudes toward assessment. Critical here is the notion that assessment can and should be instrumental in providing information to facilitate contemporary goals for instruction and inclusion rather than antiquated goals for ranking and selection. Competence will no longer be defined in terms of number of correct responses. Rather, emphasis will shift to characterizing the consistency, nature, and quality of performance under varying conditions. Given the complexity of knowledge and skills that are the focus of standards-based reforms, techniques and methods must be developed and sampled in ways that promote those aspects of subject-matter achievement that are most valued in particular contexts (e.g., school, workplace). Of necessity, attention will focus on the inferences that can be drawn regarding student learning and competence from meaningful combinations of evidence.

Certain conditions of assessment that are currently being advocated will certainly be pervasive in the future. One is socially situated assessment. Assessment will require performance in group efforts where students contribute to community tasks and assist others. Shared performance encourages and promotes a strong sense of community and an effective workforce, as learning becomes attuned to the constraints and resources of the environment. In this context, students develop and question their definitions of competence, observe how others reason, and receive feedback on their own problem-solving efforts. An important aspect of this social setting is that students develop facility in accepting help and stimulation from others.

A second issue is the display of competence. Advanced information technology will be used to openly display standards and criteria for competent performance to parents, teachers, and students. The performance criteria, by which the successful education of students is to be judged, must be as recognizable as possible so that they can motivate and direct learning and community expectations. This display particularly illustrates the relevance and utility of knowledge and skill that is being acquired for use and transfer to different life circumstances.

A third key issue is cognitive significance. Assessment will provide content coverage and not neglect significant processes of performance such as raising questions, representing and planning a problem prior to solution, and offering conceptual explanations. Constructing instructionally relevant assessment situations necessitates analyzing the cognitive requirements or demands of situations and designing related scoring procedures that attend to the differential complexity of the performances of more or less experienced learners.

With these criteria as a guide, the assessment of learning and achievement can be designed to provide useful information about content and skill that should be studied or taught in order to improve performance. For this purpose, testing must be an integral part of instruction so that assessment helps guide teachers and students toward the attainment of educational goals.

> If assessments are occasional externally controlled events used primarily for aggregated measures of student achievement levels, they are unlikely to be constructed in ways that provide rich information about the processes of student learning and their individual, idiosyncratic approaches to different kinds of tasks and opportunities. Consequently, teachers will have little opportunity to use the results to understand the complex nuances of student learning in ways that support more successful instruction, and little information on which to act in trying to rethink their daily practices. (Darling-Hammond, 1994, p. 20)

Equally important is the utility of assessment information for increasing proficiency in students' ability to learn. The sheer amount of available and changing information will force curriculum and assessment to emphasize the utility of current learning and the organization of a student's knowledge to support future learning. Students will need to develop an attitude of attaining the knowledge and skill that is necessary for the intellectual activity of handling a large volume of information; they learn to take multiple perspectives, to generate key organizational concepts, and to make analogies to new situations. In essence, a part of their education focuses on generative abilities to update their knowledge-based competence and develop learning skills and strategies for efficiently accessing the resources for reasoned problem solving.

Thus, in the future, the assessment of achievement will encompass cognitive abilities, disciplinary performance objectives, measurement procedures, and instructional practices. "The consequent increase in complexity sometimes seems daunting, particularly because of the interdisciplinary nature of much of the discussion. Nonetheless, the effort is worthwhile because the ultimate goal is a body of theory and methods that should be immensely more valuable to the world of education" (Braun, 1993, p. 385). The effort begins by acknowledging assessment and instruction as integral systems that foster access to effective education and to the attainment of subject-matter competence and learning proficiencies. Research on one issue cannot proceed without cognizance of and integration of the other. Nor can either be effective without attention to the contributions of cognitive science to our understanding of the development and use of knowledge. Assessments of the specific cognitive processes and structures involved in learning and achievement can and should be subjected to empirical scrutiny that will challenge or support extant theories of learning and achievement. The success of this iterative endeavor is dependent on a shared agenda focused on the equitable improvement of educational opportunities and attainment. As recognized long ago by Cronbach, to keep such theoretical and empirical efforts separate is to ensure their respective inadequacy.

REFERENCES

Ackerman, P. L. (1986). Individual differences in information processing: An investigation of intellectual abilities and task performance during practice. *Intelligence, 10,* 101–139.

Ackerman, P. L. (1987). Individual differences in skill learning: An integration of psychometric and information processing perspectives. *Psychological Bulletin, 102,* 3–27.

Ackerman, P. L. (1988). Determinants of individual differences during skill acquisition: Cognitive abilities and information processing. *Journal of Experimental Psychology: General, 117,* 288–318.

Ackerman, P. L. (1989). Individual differences and skill acquisition. In P. L. Ackerman, R. J. Sternberg, & R. Glaser (Eds.), *Learning and individual differences* (pp. 164–217). New York: Freeman.

Allard, F., & Starkes, J. L. (1980). Perception in sport: Volleyball. *Journal of Sport Psychology, 2,* 22–33.

Anastasi, A. (1967). Psychology, psychologists, and psychological testing. *American Psychologist, 22,* 297–306.

Anderson, J. R. (1983). *The architecture of cognition.* Cambridge, MA: Harvard University Press.

Anderson, J. R. (1993). *Rules of the mind.* Hillsdale, NJ: Erlbaum.

Baker, E. L., Freeman, M., & Clayton, S. (1991). Cognitive assessment of history for large-scale testing. In M. C. Wittrock & E. L. Baker (Eds.), *Testing and cognition* (pp. 131–153). Englewood Cliffs, NJ: Prentice Hall.

Barron, B., Schwartz, D. L., Vye, N., Moore, A., Petrosino, A., Zech, L., Bransford, J. D., & the Cognition and Technology Group at Vanderbilt. (1998). Doing with understanding:

Lessons from research on problem and project-based learning. *Journal of Learning Sciences, 7,* 271–311.

Barron, B., Vye, N. J., Zech, L., Schwartz, D., Bransford, J. D., Goldman, S. R., Pellegrino, J., Morris, J., Garrison, S., & Kantor, R. (1995). Creating contexts for community-based problem solving: The Jasper Challenge Series. In C. Hedley, P. Antonacci, & M. Rabinowitz (Eds.), *Thinking and literacy: The mind at work* (pp. 47–71). Hillsdale, NJ: Erlbaum.

Baxter, G. P., Elder, A. D., & Glaser, R. (1996). Knowledge-based cognition and performance assessment in the science classroom. *Educational Psychologist, 31,* 133–140.

Baxter, G. P., & Glaser, R. (1998). The cognitive complexity of science performance assessments. *Educational Measurement: Issues and Practice, 17*(3), 37–45.

Baxter, G. P., Shavelson, R. J., Goldman, S. R., & Pine, J. (1992). Evaluation of procedure-based scoring for hands-on science assessment. *Journal of Educational Measurement, 29,* 1–17.

Bereiter, C. (1963). Some persisting dilemmas in the measurement of change. In C. W. Harris (Ed.), *Problems in measuring change* (pp. 3–20). Madison: University of Wisconsin Press.

Bereiter, C., & Scardamalia, M. (1987). *The psychology of written composition.* Hillsdale, NJ: Erlbaum.

Bereiter, C., & Scardamalia, M. (1993). *Surpassing ourselves: An inquiry into the nature and implications of expertise.* Chicago: Open Court.

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

Bond, L., & Glaser, R. (1979). ATI, mostly A and T and not much of I [Review of Aptitudes and instructional methods by L. J. Cronbach & R. E. Snow]. *Applied Psychological Measurement, 3,* 137–140.

Bransford, J. D., Brown, A., & Cocking, R. (1999). *How people learn: Brain. mind, experience and school.* Washington, DC: National Academy Press.

Braun, H. (1993). Comments on Chapters 11–14. In N. Fredericksen, R. J. Mislevy, & I. I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. 385–390). Hillsdale, NJ: Erlbaum.

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

Brown, A., & Palinscar, A. M. (1989). Guided, cooperative learning and individual knowledge acquisition. In L. B. Resnick (Ed.), *Knowing, learning, and instruction: Essays in honor of Robert Glaser* (pp. 393–451). Hillsdale, NJ: Erlbaum.

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

Bruner, J. S. (1978). The role of dialogue in language acquisition. In A. Sinclair, R. J. Jarvell, & W. J. M. Levelt (Eds.), *The child's conception of language* (pp. 241–256). New York: Springer.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis methods.* Newbury Park, CA: Sage.

Bush, G. W. (1991). *America 2000: An educational strategy.* Washington, DC: U.S. Department of Education.

Carroll, J. B. (1976). Psychometric tests as cognitive tasks: A new "structure of intellect." In L. B. Resnick (Ed.), *The nature of intelligence* (pp. 27–56). Hillsdale, NJ: Erlbaum.

Carroll, J. B. (1978). How shall we study individual differences in cognitive abilities? Methodological and theoretical perspectives. *Intelligence, 2,* 87–115.

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology, 4,* 55–81.

Chi, M. T. H., Bassock, M., Lewis, M. W., Reimann, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science, 13,* 145–182.

Chi, M. T. H., Feltovich, P., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science, 5,* 121–152.

Chi, M. T. H., Glaser, R., & Farr, M. (Eds.). (1988). *The nature of expertise.* Hillsdale, NJ: Erlbaum.

Cognition and Technology Group at Vanderbilt. (1994). From visual word problems to learning communitites: Changing conceptions of cognitive research. In K. McGilly (Ed.), *Classroom lessons: Integrating cognitive theory and classroom practice* (pp. 157–200). Cambridge, MA: MIT Press/Bradford Books.

Cognition and Technology Group at Vanderbilt. (1997). *The Jasper Project: Lessons in curriculum, instruction, assessment, and professional development.* Mahwah, NJ: Erlbaum.

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). Greenwich, CT: Ablex.

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D., & York, R. L. (1966). *Equality of educational opportunity.* Washington, DC: U.S. Government Printing Office.

Cote, N., Goldman, S. R., & Saul, E. (1998). Students making sense of informational text: Relations between processing and representation. *Discourse Processes, 25,* 1–53.

Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American Psychologist, 12,* 671–684.

Cronbach, L. J. (1975). Beyond the two disciplines of scientific psychology. *American Psychologist, 30,* 116–127.

Cronbach, L. J., & Furby, L. (1970). How should we measure change—Or should we? *Psychological Bulletin, 74,* 68–80.

Cronbach, L. J., & Gleser, G. C. (1957). *Psychological tests and personnel decisions.* Urbana: University of Illinois Press.

Cronbach, L. J., & Snow, R. E. (1977). *Aptitudes and instructional methods: A handbook for research on interactions.* New York: Irvington.

Darling-Hammond, L. (1994). Performance-based assessment and educational equity. *Harvard Educational Review, 64,* 5–30.

de Groot, A. (1978). *Thought and choice in chess.* The Hague: Mouton. (Original work published 1946)

Dunbar, S. B., Koretz, D. M., & Hoover, H. D. (1991). Quality control in the development and use of performance assessments. *Applied Measurement in Education, 4,* 289–303.

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

Educational Testing Service. (1985). *The reading report card: Progress toward excellence in our schools. Trends in reading over four national assessments.* Princeton, NJ.

Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93,* 179–197.

Embretson, S. E. (1984). A general latent trait model for response processes. *Psychometrica, 49,* 175–186.

Embretson, S. E. (Ed.). (1985). *Test design: Developments in psychology and psychometrics.* New York: Academic Press.

Embretson, S. E. (1991). A multidimensional item response model for learning processes. *Psychometrica, 56,* 495–515.

Embretson, S. E. (1993). Psychometric models for learning and cognitive processes. In N. Fredericksen, R. J. Mislevy, & I. I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. 125–150). Hillsdale, NJ: Erlbaum.

Embretson, S. E., & Waxman, M. (1989). *Models for processing and individual differences in spatial folding*. Unpublished manuscript.

Fisher, G. (1973). Linear logistic test model as an instrument in educational research. *Psychologica, 37,* 359–374.

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

Fredericksen, N., Mislevy, R. J., & Bejar, I. I. (Eds.). (1993). *Test theory for a new generation of tests*. Hillsdale, NJ: Erlbaum.

Glaser, R. (1976). Components of a psychology of instruction: Toward a science of design. *Review of Educational Research, 46,* 1–24.

Glaser, R. (1981). The future of testing: A research agenda for cognitive psychology and psychometrics. *American Psychologist, 36,* 923–936.

Glaser, R. (1986). A cognitive science perspective on selection and classification and on technical training. *Advances in Reading/Language Research, 4,* 253–268.

Glaser, R. (1991). Expertise and assessment. In M. C. Wittrock & E. L. Baker (Eds.), *Testing and cognition* (pp. 17–30). Englewood Cliffs, NJ: Prentice Hall.

Goldman, S. R. (1997). Learning from text: Reflections on the past and suggestions for the future. *Discourse Processes, 23,* 357–398.

Goldman, S. R., & Pellegrino, J. W. (1984). Deductions about induction: Analysis of developmental and individual differences. In R. J. Sternberg (Ed.), *Advances in the psychology of human intelligence* (Vol. 2, pp. 149–197). Hillsdale, NJ: Erlbaum.

Greeno, J. G., Pearson, P. D., & Schoenfeld, A. H. (1997). Implications for the National Assessment of Educational Progress of research on learning and cognition. In R. Glaser, R. Linn, & G. Bohrnstedt (Eds.), *Assessment in transition: Monitoring the nation's educational progress, background studies*. Stanford, CA: National Academy of Education.

Guttman, L. (1970). Integration of test design and analysis. In *Proceedings of the 1969 Invitational Conference on Testing Problems* (pp. 53–65). Princeton, NJ: Educational Testing Service.

Haertel, E. H. (1984). An application of latent class models to assessment data. *Applied Psychological Measurement, 8,* 333–346.

Hamilton, L. S., Nussbaum, E. M., & Snow, R. E. (1997). Interview procedures for validating science assessments. *Applied Measurement in Education, 10,* 181–200.

Hunt, E. (1978). Mechanics of verbal ability. *Psychological Review, 85,* 109–130.

Hunt, E., Frost, N., & Lunnenborg, C. (1973). Individual differences in cognition: A new approach to intelligence. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 7, pp. 87–122). New York: Academic Press.

Hunt, E., & Lansman, M. (1975). Cognitive theory applied to individual differences. In W. K. Estes (Ed.), *Handbook of learning and cognitive processes: Introduction to concepts and issues* (Vol. 1). Hillsdale, NJ: Erlbaum.

Hunt, E., Lunnenborg, C., & Lewis, J. (1975). What does it mean to be high verbal? *Cognitive Psychology, 7,* 194–227.

Hunt, E., & Minstrell, J. (1994). A cognitive approach to teaching physics. In K. McGilly (Ed.), *Classroom lessons: Integrating cognitive theory and classroom practice* (pp. 51–74). Cambridge, MA: MIT Press/Bradford Books.

Klahr, D. (1976). *Cognition and instruction*. Hillsdale, NJ: Erlbaum.

Kotovsky, K., & Simon, H. A. (1973). Empirical tests of a theory of human acquisition of concepts for sequential events. *Cognitive Psychology, 4,* 399–424.

Kyllonen, P. C. (1993). Aptitude testing inspired by information processing: A test of the four-sources model. *Journal of General Psychology, 120,* 375–405.

Kyllonen, P. C., Lohman, D. F., & Woltz, D. J. (1984). Componential modeling of alternative strategies for performing spatial tasks. *Journal of Educational Psychology, 76,* 1325–1345.

Kyllonen, P. C., & Shute, V. J. (1989). A taxonomy of learning skills. In P. L. Ackerman, R. J. Sternberg, & R. Glaser (Eds), *Learning and individual differences* (pp. 117–163). New York: Freeman.

Lane, S. (1993). The conceptual framework for the development of a mathematics performance assessment instrument. *Educational Measurement: Issues and Practice, 12,* 16–23.

Larry P. v. Riles, 495 F. Supp. 926 (N.D. Cal. 1979).

Law, D. J., Morrin, K. A., & Pellegrino, J. W. (1995). Training effects and working memory contributions to skill acquisition in a complex coordination task. *Learning and Individual Differences, 7,* 207–234.

Law, D. J., Pellegrino, J. W., & Hunt, E. (1993). Comparing the tortoise and the hare: Gender differences in dynamic spatial reasoning tasks. *Psychological Science, 4,* 35–40.

Lesgold, A. M., Pellegrino, J. W., Fokkema, S. D., & Glaser, R. (Eds.). (1977). *Cognitive psychology and instruction* (NATO Conference Series III, Human Factors, Vol. 5). New York: Plenum.

Lesgold, A. M., Rubinson, H., Feltovich, P., Glaser, R., Klopfer, D., & Wang, Y. (1988). Expertise in a complex skill: Diagnosing X-ray pictures. In M. T. H. Chi, R. Glaser, & M. J. Farr (Eds.), *The nature of expertise* (pp. 311–342). Hillsdale, NJ: Erlbaum.

Levidow, B. B., Hunt, E., & McKee, C. (1991). The DIAGNOSER: A HyperCard tool for building theoretically based tutorials. *Behavior Research Methods, Instruments and Computers, 23,* 249–252.

Linn, R. L. (1986). Educational testing and assessment: Research needs and policy issues. *American Psychologist, 41,* 1153–1160.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20,* 5–21.

Lohman, D. F. (1994). Component scores as residual variation (or why the intercept correlates best). *Intelligence, 19,* 1–11.

Lord, F. M. (1952). A theory of test scores. *Psychometrica Monographs, 7*(4, Pt. 2).

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Magone, M. E., Cai, J., Silver, E. A., & Wang, N. (1994). Validating the cognitive complexity and content quality of a mathematics performance assessment. *International Journal of Educational Research, 21,* 317–340.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13–23.

Minstrell, J. (1992, April). *Facets of students' knowledge: A practical view from the classroom.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

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.

Minstrell, J., & Stimpson, V. (1996). A classroom environment for learning: Guiding students' reconstruction of understanding and reasoning. In L. Schauble & R. Glaser (Eds.), *Innovations in learning: New environments for education* (pp. 175–202). Mahwah, NJ: Erlbaum.

Mislevy, R. J. (1993). Foundations of a new test theory. In N. Frederiksen, R. J. Mislevy, & I. I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. 19–39). Hillsdale, NJ: Erlbaum.

Mislevy, R. J. (1995). Probability-based inference in cognitive diagnosis. In P. D. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 43–71). Hillsdale, NJ: Erlbaum.

Mislevy, R. J. (1996). *Evidence and inference in educational assessment* (CSE Technical Report 414). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing, Graduate School of Education and Information Studies, University of California, Los Angeles.

Morrin, K. A., Law, D. J., & Pellegrino, J. W. (1994). Structural modeling of information coordination abilities: An evaluation and extension of the Yee, Hunt, and Pellegrino model. *Intelligence, 19,* 117–144.

Mumaw, R. J., & Pellegrino, J. W. (1984). Individual differences in complex spatial processing. *Journal of Educational Psychology, 76,* 920–939.

National Academy of Education. (1993). *The trial state assessment: Prospects and realities.* Stanford, CA.

National Academy of Education. (1997). *Assessment in transition: Monitoring the nation's educational progress.* Stanford, CA.

National Commission for Educational Excellence. (1983). *A nation at risk.* Washington, DC: U.S. Government Printing Office.

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

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

National Educational Goals Panel. (1991). *Measuring progress toward the national educational goals: Potential indicators and measurement strategies.* Washington, DC: U.S. Government Printing Office.

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

National Research Council. (1999). *Grading the nation's report card: Evaluating NAEP and transforming the assessment of educational progress.* Washington, DC: National Academy Press.

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

Nichols, P. D., Chipman, S. F., & Brennan, R. L. (Eds.). (1995). *Cognitively diagnostic assessment.* Hillsdale, NJ: Erlbaum.

Nichols, P. D., & Sugrue, B. (1997). *Construct-centered test development for NAEP's short forms.* Washington, DC: National Center for Education Statistics.

Nickerson, R. S. (Ed.). (1989). Special issue on educational assessment. *Educational Researcher, 18,* 3–33.

Nitko, A. J. (1989). Designing tests that are integrated with instruction. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 447–474). New York: MacMillan.

Patel, V. L., & Groen, G. L. (1986). Knowledge based solution strategies in medical reasoning. *Cognitive Science, 10,* 91–116.

Pellegrino, J. W., & Glaser, R. (1979). Cognitive correlates and components in the analysis of individual differences. *Intelligence, 3,* 187–214.

Pellegrino, J. W., & Glaser, R. (1982). Analyzing aptitudes for learning: Inductive reasoning. In R. Glaser (Ed.), *Advances in instructional psychology* (Vol. 2, pp. 269–345). Hillsdale, NJ: Erlbaum.

Pellegrino, J. W., & Hunt, E. (1989). Computer controlled assessment of static and dynamic spatial reasoning. In R. F. Dillon & J. W. Pellegrino (Eds.), *Testing: Theoretical and applied perspectives* (pp. 174–198). New York: Praeger.

Pellegrino, J. W., & Kail, R. V. (1982). Process analyses of spatial aptitude. In R. J. Sternberg (Ed.), *Advances in the psychology of human intelligence* (Vol. 1, pp. 311–365). Hillsdale, NJ: Erlbaum.

Perfetti, C. A., & Goldman, S. R. (1976). Discourse memory and reading comprehension skill. *Journal of Verbal Learning and Verbal Behavior, 15,* 33–42.

Perfetti, C. A., & Hoagaboam, T. (1975). Relationship between simple word decoding and reading comprehension skill. *Journal of Educational Psychology, 67,* 461–469.

Perfetti, C. A., & Lesgold, A. M. (1982). *Reading ability.* Hillsdale, NJ: Erlbaum.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danish Institute for Educational Research.

Resnick, L. B. (1979). The future of IQ testing. In R. J. Sternberg & D. K. Detterman (Eds.), *Human intelligence: Perspectives on its theory and measurement* (pp. 203–215). Norwood, NJ: Ablex.

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.

Shavelson, R. J., Baxter, G. P., & Gao, X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement, 30,* 215–232.

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

Shepard, R. N., & Cooper, L. A. (1983). *Mental images and their transformations.* Cambridge, MA: MIT Press.

Sherwood, R. D., Petrosino, A. J., Lin, X., Lamon, M., & the Cognition and Technology Group at Vanderbilt. (1995). Problem-based macro contexts in science instruction: Theoretical basis, design issues, and the development of applications. In D. Lavoie (Ed.), *Towards a cognitive-science perspective for scientific problem solving* (pp. 191–214). Manhattan, KS: National Association for Research in Science Teaching.

Shute, V. J. (1992). Aptitude-treatment interactions and cognitive skill diagnosis. In W. J. Regian & V. J. Shute (Eds.), *Cognitive approaches to automated instruction* (pp. 15–43). Hillsdale, NJ: Erlbaum.

Shute, V. J. (1993). A macroadaptive approach to tutoring. *Journal of Artificial Intelligence in Education, 4,* 61–93.

Silver, E. A., & Stein, M. K. (1996). The QUASAR project: The "revolution of the possible" in mathematics instructional reform in urban middle schools. *Urban Education, 30,* 476–521.

Snow, R. E. (1980). Aptitude and achievement. *New Directions in Testing and Measurement, 5,* 39–59.

Snow, R. E. (1989a). Aptitude interactions as a framework for individual differences in learning. In P. L. Ackerman, R. J. Sternberg, & R. Glaser (Eds.), *Learning and individual differences: Advances in theory and research* (pp. 13–59). New York: Freeman.

Snow, R. E. (1989b). Cognitive-conative aptitude interactions in learning. In R. Kanfer, P. L. Ackerman, & R. Cudek (Eds.), *Abilities, motivation, and methodology: The Minnesota Symposium on Learning and Individual Differences* (pp. 435–474). Hillsdale, NJ: Erlbaum.

Snow, R. E. (1994). Abilities in academic tasks. In R. J. Sternberg & R. K. Wagner (Eds.), *Mind in context: Interactionist perspectives on human intelligence* (pp. 3–37). New York: Cambridge University Press.

Snow, R. E., Corno, L., & Jackson, D. (1996). Individual differences in affective and conative functions. In D. Berliner & R. Calfee (Eds.), *Handbook of research in educational psychology* (pp. 186–242). New York: Macmillan.

Snow, R. E., & Lohman, D. F. (1989). Implications of cognitive psychology for educational measurement. In R. L. Linn (Ed.), *Educational measurement* (pp. 263–331). New York: Macmillan.

Snow, R. E., & Lohman, D. F. (1993). Cognitive psychology, new test design, and new test theory: An introduction. In N. Fredericksen, R. J. Mislevy, & I. I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. 1–17). Hillsdale, NJ: Erlbaum.

Spearman, C. (1904a). The proof and measurement of the correlation between two things. *American Journal of Psychology, 15*, 201–292.

Spearman, C. (1904b). "General intelligence" objectively determined and measured. *American Journal of Psychology, 18*, 161–169.

Sternberg, R. J. (1977). *Intelligence, information processing, and analogical reasoning: The componential analysis of human abilities*. Hillsdale, NJ: Erlbaum.

Sternberg, R. J. (1984). What cognitive psychology can (and cannot) do for test development. In B. S. Plake (Ed.), *Social and technical issues in testing: Implications for test construction and usage* (pp. 39–60). Hillsdale, NJ: Erlbaum.

Stiggins, R. (1994). *Student-centered classroom assessment*. New York: Merrill.

Swanson, J. H. (1990). One-to-one instruction: An experimental evaluation of effective tutoring strategies (Doctoral dissertation, Stanford University, 1990). *Dissertation Abstracts International, 50A*, 8.

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20*, 345–354.

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

White, B. Y., & Fredericksen, J. R. (1998). Inquiry, modeling, and metacognition: Making science accessible to all students. *Cognition and Instruction, 16*, 3–118.

Whitely, S. E. (1980). Multicomponent latent trait models for ability tests. *Psychometrica, 45*, 479–494.

Willet, J. B. (1988). Questions and answers in the measurement of change. In *Review of research in education* (Vol. 15, pp. 345–422). Washington, DC: American Educational Research Association.

Willet, J. B., & Sayer, A. G. (1994). Using covariance structure analysis to detect correlates and predictors of individual change over time. *Psychological Bulletin, 116*, 363–381.

Wilson, M. R. (1989). Saltus: A psychometric model of discontinuity in cognitive development. *Psychological Bulletin, 105*, 276–289.

Wittrock, M., & Baker, E. (1991). *Testing and cognition*. Englewood Cliffs, NJ: Prentice Hall.

Woltz, D. J. (1988). An investigation of the role of working memory in procedural skill acquisition. *Journal of Experimental Psychology: General, 117*, 319–331.

Woltz, D. J., & Shute, V. J. (1993). Individual differences in repetition priming and its relationship to declarative knowledge acquisition. *Intelligence, 17*, 333–359.

Manuscript received January 4, 1999
Accepted April 1, 1999